The evolution of the mitochondrial proteome in animals

Viraj Muthye

Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Bioinformatics Commons

Recommended Citation
https://lib.dr.iastate.edu/etd/17752

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
The evolution of the mitochondrial proteome in animals

by

Viraj Rajendra Muthye

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Dennis Lavrov, Co-major Professor
Carolyn Lawrence-Dill, Co-major Professor
Karin Dorman
Robert Jernigan
Iddo Friedberg

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University
Ames, Iowa
2019

Copyright © Viraj Rajendra Muthye, 2019. All rights reserved.
DEDICATION

To my wife, younger brother, parents and friends for their unconditional support, commitment and encouragement throughout my life
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>v</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>vii</td>
</tr>
<tr>
<td>CHAPTER 1. GENERAL INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Mitochondria: An Introduction to the Organelle</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Overview of Mitochondrial Functions</td>
<td>3</td>
</tr>
<tr>
<td>1.3 Overview of Mitochondrial Protein Import</td>
<td>4</td>
</tr>
<tr>
<td>1.4 Characterization of the Mitochondrial Proteome</td>
<td>6</td>
</tr>
<tr>
<td>1.4.1 Experimental approaches</td>
<td>7</td>
</tr>
<tr>
<td>1.4.2 Computational approaches</td>
<td>7</td>
</tr>
<tr>
<td>1.4.3 Integrative approaches</td>
<td>10</td>
</tr>
<tr>
<td>1.5 Databases for Mitochondrial Proteomes</td>
<td>10</td>
</tr>
<tr>
<td>1.6 An Introduction to Metazoan Phylogeny</td>
<td>11</td>
</tr>
<tr>
<td>1.7 Dissertation Organization</td>
<td>13</td>
</tr>
<tr>
<td>1.8 References</td>
<td>14</td>
</tr>
<tr>
<td>1.9 Tables and Figures</td>
<td>20</td>
</tr>
<tr>
<td>CHAPTER 2. CHARACTERIZATION OF MITOCHONDRIAL PROTEOMES OF NON-BILATERIAN ANIMALS</td>
<td>22</td>
</tr>
<tr>
<td>2.1 Abstract</td>
<td>22</td>
</tr>
<tr>
<td>2.2 Introduction</td>
<td>23</td>
</tr>
<tr>
<td>2.3 Materials and Methods</td>
<td>27</td>
</tr>
<tr>
<td>2.3.1 Predicting mitochondrial proteomes</td>
<td>27</td>
</tr>
<tr>
<td>2.3.2 Analyses of inferred mt-proteomes</td>
<td>29</td>
</tr>
<tr>
<td>2.4 Results and Discussion</td>
<td>31</td>
</tr>
<tr>
<td>2.4.1 Mitochondrial proteomes in nonbilaterian animals</td>
<td>31</td>
</tr>
<tr>
<td>2.4.2 Identification of common animal mt-proteins (CAMPs) and bilaterian-specific mt-proteins</td>
<td>33</td>
</tr>
<tr>
<td>2.4.3 Identification of predicted nonbilaterian mt-proteins with no ortholog in the reference mt-proteomes</td>
<td>34</td>
</tr>
<tr>
<td>2.4.4 Conservation of proteins involved in core mitochondrial functions</td>
<td>36</td>
</tr>
<tr>
<td>2.4.5 Analysis of mitochondrial targeting signals of mt-proteins</td>
<td>36</td>
</tr>
<tr>
<td>2.4.6 Analysis of mitochondrial protein domains</td>
<td>38</td>
</tr>
<tr>
<td>2.5 Conclusion</td>
<td>40</td>
</tr>
<tr>
<td>2.6 Acknowledgement</td>
<td>41</td>
</tr>
<tr>
<td>2.7 References</td>
<td>41</td>
</tr>
<tr>
<td>2.8 Tables and Figures</td>
<td>51</td>
</tr>
</tbody>
</table>
CHAPTER 3. CAUSES AND CONSEQUENCES OF MITOCHONDRIAL PROTEOME SIZE-VARIATION IN ANIMALS ................................................................. 59

3.1 Abstract ................................................................. 59
3.2 Introduction ............................................................. 59
3.3 Materials and Methods ................................................. 62
  3.3.1 Assembling animal mt-proteomes .................................. 62
  3.3.2 Identification of orthologous groups .................................. 63
  3.3.3 Identification of Mitochondrial Targeting Signals (MTS) .................. 63
  3.3.4 Functional analysis of mt-proteins ...................................... 64
  3.3.5 Data availability .................................................... 64
3.4 Results ................................................................. 65
  3.4.1 Evolution of animal mt-proteomes ..................................... 65
  3.4.2 Functional analysis of mt-proteins ...................................... 68
  3.4.3 Role of MTS in mt-proteome evolution .................................. 70
3.5 Discussion ............................................................. 70
3.6 Conclusion .............................................................. 73
3.7 Acknowledgement ....................................................... 74
3.8 References .............................................................. 74
3.9 Tables and Figures ....................................................... 79

CHAPTER 4. MMPDB AND MITOPREDICTOR: TOOLS FOR FACILITATING COMPARATIVE ANALYSIS OF ANIMAL MITOCHONDRIAL PROTEOMES .......... 84

4.1 Abstract ................................................................. 84
4.2 Introduction .............................................................. 85
4.3 Materials and Methods .................................................. 87
  4.3.1 The Metazoan Mitochondrial Proteome Database ...................... 87
  4.3.2 MitoPredictor ....................................................... 89
4.4 Results and Discussion .................................................. 95
  4.4.1 The Metazoan Mitochondrial Proteome Database ...................... 95
  4.4.2 MitoPredictor ....................................................... 98
4.5 Acknowledgement ....................................................... 100
4.6 References .............................................................. 101
4.7 Tables and Figures ....................................................... 104
4.8 Supplementary Materials ............................................... 109
  4.8.1 Selection of machine-learning algorithm ............................... 109
  4.8.2 Evaluation of the Random Forest model ............................... 110
  4.8.3 Evaluation of prediction performance of MitoPredictor and SubCons features 113

CHAPTER 5. GENERAL CONCLUSION ........................................... 115

5.1 References .............................................................. 120
ACKNOWLEDGMENTS

It is difficult to acknowledge all the people who have directly and indirectly helped me through this wonderful and challenging journey at Iowa State University. I feel that the page limitation does not do justice to the reasoning behind including this in the dissertation. Neither the brevity of my acknowledgements nor the order reflects their significance.

First and foremost, the reason why I do what I do, I wish to acknowledge my best friend and my wife, Bhakti Bansode. Nothing would have been possible without her support, love and friendship. I still am amazed how her companionship makes me feel at home in a house so far away from home. It is as much her journey as is mine. I want to thank my brother, for being the amazing guy that he has always been. It is tough to describe the support and love my parents, and my wife’s parents, have given me throughout this journey, which made the more difficult parts seem a lot less difficult. I want to acknowledge my other brother, my partner-in-crime, Raj Rege, for his friendship and patience through the years. And I want to acknowledge my Ames family- Gaurav Kandoi, Pulkit Kanodia, Akshay Yadav, Surya and Saranya for always being there unconditionally, tolerating my quirks, celebrating my joys and helping me through the tough times.

It was quite a transition from biotechnology to bioinformatics, and the BCB community here played an important role in making sure that transition went smoothly. Dr. Dennis Lavrov has been an amazing mentor throughout these fun and challenging years. I feel extremely lucky to have been a part of his lab, and learned so much from him. He has inspired, guided and challenged me throughout my PhD years, for which I am extremely grateful. The friendly, supportive and inspiring environment he created in lab is the reason why I am now obsessed with mitochondria. When I started my PhD, I had very limited knowledge in programming. This made the first year in the program crucial. An amazing first rotation in Dr. Carolyn Lawrence-Dill’s lab really helped me get started in this program. I learned a lot from Dr. Lawrence-Dill, Dr. Andorf and Dr. Cannon, who
made sure that I was well-set in this program and really made me confident that I could do this. Everyone from my committee - Dr. Karin Dorman, Dr. Robert Jernigan and Dr. Iddo Friedberg—have been extremely supportive and their suggestions have helped me shape my dissertation and research. And last, but in no way the least, I wish to acknowledge Trish Stauble for everything she has done, for her constant motivation and her love for the program and everyone in BCB.

Finally, I also wish to acknowledge my teaching community. During my stay here, I got the wonderful opportunity to teach every semester. I want to thank Linda Westgate, Dr. Jim Colbert and Chris Myers for their superb assistance and support through the years, making the six hours of teaching each week a stress-buster experience. I also want to thank every single student I had for teaching me much more than how much I taught them.

I may have missed quite a few people in this section, and for that I apologize. I wish to end this section with two words which changed my life forever: Hello World!
Mitochondria are subcellular organelles in eukaryotes which possess their own genome. While they are most well-known for their role in energy metabolism via oxidative phosphorylation, research has shown that mitochondria are involved in diverse critical cellular functions like Fe/S cluster biosynthesis, apoptosis, signaling, etc. In mammals, over 1,500 proteins carry out these functions in the mitochondria. A small portion of these proteins (∼1%) is contributed by the mitochondrial genome, whereas a vast majority (∼99%) are encoded in the nuclear genome and transported into the organelle. This set of nuclear-encoded mitochondrial proteins is defined as the “mitochondrial proteome”. The primary objective of my research is to analyze the evolution of the mitochondrial proteome in animals, and to develop tools for facilitating the comparative analysis of animal mitochondrial proteomes.

For obtaining a broad picture of animal mitochondrial proteome evolution, it is necessary to examine the mitochondrial proteomes of both bilaterian and non-bilaterian animals. All experimentally-characterized mitochondrial proteomes in animals are from Bilateria. This is unfortunate, since the comparative analysis of animal mitochondrial genomes has shown that most of the mitochondrial genomic diversity in animals can be found in the four phyla of non-bilaterian animals (Porifera, Cnidaria, Ctenophora, and Placozoa). In this dissertation, we carry out the first comparative analysis of mitochondrial proteomes from non-bilaterian animals. We use bioinformatic techniques to predict the mitochondrial proteomes from non-bilaterian animals. We detect a large variation in the size and content of the inferred mitochondrial proteomes of non-bilaterian animals. The size of the inferred mitochondrial proteomes ranges from 454 proteins in Kudoa iwatai to 2,119 proteins in Leucosolenia complicata. We find that much of the variation in the size of the mitochondrial proteomes in non-bilaterian animals is due to the number of proteins with a mitochondrial targeting signal, but no ortholog to any human or yeast protein. Additionally, we also identify
several instances of mitochondrial neolocalization in the non-bilaterian mitochondrial proteomes. Conversely, ∼ 2.5% of the human mitochondrial proteome has no ortholog in any non-bilaterian species, representing potential bilaterian mitochondrial innovations. Next, through a comparative analysis of the experimentally-characterized mitochondrial proteomes of bilaterian animals, we investigate the causes and functional implications of the variation in size and content of the animal mitochondrial proteomes. We find that the animal mitochondrial proteome is a dynamic entity, with a small core of mitochondrial proteins that are conserved in all four animals, and a large number of lineage-specific gains and losses. Of the several factors responsible for the size-variation in the four animal mitochondrial proteomes, we find that the gain of novel mitochondrial proteins in mammals and loss of conserved mitochondrial proteins in the two ecdysozoans are the main contributors. Interestingly, while nearly one-fifth of each animal mitochondrial proteome consists of proteins that underwent mitochondrial neolocalization in animals, the majority of these neolocalized proteins lack a canonical mitochondrial targeting signal.

While carrying out comparative analysis of mitochondrial proteomes in animals, researchers encounter two main challenges: 1) data on experimentally-characterized animal mitochondrial proteomes is scattered across several databases, and 2) most animal phyla lack a species with an experimentally-characterized mitochondrial proteome. To address these challenges, we develop two tools to facilitate the comparative analysis of mitochondrial proteomes in animals- 1) the Metazoan Mitochondrial Proteome database, which consolidates data on animal mitochondrial proteomes from various sources, and 2) MitoPredictor, a novel machine-learning tool to predict mitochondrial proteins in animals, using three sources of information: orthology, mitochondrial-targeting signal prediction and protein-domain information.
CHAPTER 1. GENERAL INTRODUCTION

1.1 Mitochondria: An Introduction to the Organelle

Mitochondria are double-membrane containing organelles in most eukaryotic cells. While they are most well-known for their role in ATP production via oxidative phosphorylation (Hatefi, 1985), research on mitochondrial biology and function has now expanded the organelar functional repertoire to include processes like Fe/S cluster biosynthesis (Stehling and Lill, 2013), amino-acid metabolism (King, 2007), apoptosis (Wang and Youle, 2009) and cellular signaling (Tait and Green, 2012). In animals, besides the nucleus, the mitochondrion is the only sub-cellular organelle which possesses its own genome. It is well-established that all mitochondria originated from a free-living \( \alpha \)-proteobacterium (Gray et al., 1999) within the order Rickettsiales (Wang and Wu, 2015). The origin of mitochondria can be traced back to an ancient endosymbiosis event, around 2 billion years ago, between a free-living \( \alpha \)-proteobacterium and an archaean host cell, most closely related to the Asgard archaea (Spang et al., 2015; Zaremba-Niedzwiedzka et al., 2017). Since the endosymbiosis event, most of the genes encoded in the mitochondrial genome were either transferred to the nuclear genome or lost, resulting in a vast reduction of the size of the mitochondrial genome (Adams and Palmer, 2003).

The human mitochondrial genome was the first animal mitochondrial genome to be sequenced (Anderson et al., 1981). It is a small, compact circular molecule (~16.6 kb), with only 13 protein-coding genes (subunits of the complexes of the Oxidative Phosphorylation system), two rRNA genes, and 22 tRNA genes, one large non-coding region, a little intergenic region, and a modified genetic code. With over 6,000 sequenced animal mitochondrial genomes (Lavrov and Pett, 2016), the human mitochondrial genome proves to be fairly representative of the majority of the bilaterian animals (but see exceptions in (Lavrov and Pett, 2016)). However, bilaterian animals represent just one of the five major branches of the animal phylogeny. Characterization of mitochondrial genomes
from the other four animal lineages - Phyla Porifera, Cnidaria, Ctenophora, Placozoa - referred to as “non-bilaterian animals” revealed a much larger diversity of mitochondrial structure (Lavrov and Pett, 2016) than present in bilaterian animals alone (Section 1.6).

Analysis of mitochondrial genomes is just one aspect of understanding mitochondrial evolution, since only a minuscule portion of the total set of proteins functioning in the organelle are contributed by the mitochondrial genome (≈ 1%). The majority of the proteins involved in the diverse set of mitochondrial functions are encoded in the nuclear genome, synthesized on cytosolic ribosomes and imported into the organelle. These proteins are defined as the “mitochondrial proteome”. Once inside the organelle, some of these nuclear-encoded proteins interact with mitochondria-encoded proteins and function in energy metabolism. Others are involved in various critical cellular processes (discussed below). In animals, mitochondrial proteomes have been experimentally-characterized in human, mouse, Caenorhabditis elegans and Drosophila melanogaster. Surprisingly, while mitochondrial genomes of these animals are near-identical, the size of animal mitochondrial proteome varies, from 838 proteins in D. melanogaster to 1,700 proteins in Homo sapiens (Calvo et al., 2015; Hu et al., 2015; Smith and Robinson, 2018; Li et al., 2009a).

Thus, to obtain a complete picture of mitochondrial evolution in animals, it is necessary to 1) carry out comparative analysis of both, the animal mitochondrial genomes and the animal mitochondrial proteomes, and 2) include both bilaterian and non-bilaterian animals in such analysis. In the following chapters, we use bioinformatic techniques to analyze the evolution of the mitochondrial proteome in animals and provide tools to facilitate further analysis of animal mitochondrial proteomes. Here, we give a brief introduction to mitochondrial functions, mitochondrial protein-import pathways, strategies for characterization of animal mitochondrial proteomes and existing databases for animal mitochondrial proteins.
1.2 Overview of Mitochondrial Functions

Oxidative phosphorylation is by far the most well-known function of mitochondria. In this process, electrons are transferred through multiple mitochondrial inner-membrane protein complexes called the electron transport chain (ETC)\(^1\). The ETC complexes (complex I-IV) transport electrons to the final acceptor, oxygen, to form water. Some of these complexes pump protons across the inner mitochondrial membrane. The resulting electrochemical gradient is used by a protein complex (ATP-synthase) to generate ATP (Saraste, 1999; van der Bliek et al., 2017). The mitochondrial genome contributes subunits to all respiratory complexes, except complex II which is composed solely of nuclear-encoded proteins. Complex II links oxidative phosphorylation to the Tricarboxylic Acid Cycle (TCA)\(^2\), as it functions in both processes. The TCA cycle links metabolism of carbohydrates, proteins, and fats to energy metabolism. Catabolism of carbohydrates, proteins and fats results in acetyl-CoA, which gets oxidized in the TCA cycle thereby producing NADH and FADH\(_2\). NADH and FADH\(_2\) transfer electrons to the inner-membrane respiratory complexes, thus beginning oxidative phosphorylation (Raimundo et al., 2011; van der Bliek et al., 2017).

Mitochondria are also involved in the production of various co-factors like Iron-sulfur (Fe/S) clusters. In eukaryotes, Fe/S proteins are involved in diverse and indispensable functions in the cell (Lill et al., 2012; Stehling and Lill, 2013). The importance of Fe/S cluster biosynthesis is evident from analysis of mitosomes (double-membraned organelles in protists derived from mitochondria), which have lost critical mitochondrial pathways like oxidative phosphorylation, TCA cycle and fatty-acid oxidation. However, they contain all the required machinery for Fe/S cluster biosynthesis (Tovar et al., 2003). Recent studies have also implicated mitochondrial involvement in several other functions, including cellular signaling (Chandel, 2015), apoptosis (Wang and Youle, 2009) and innate immune responses (West et al., 2011).

---

\(^1\)ETC: electron transport chain
\(^2\)TCA: Tricarboxylic Acid Cycle
The centrality of nuclear-encoded mitochondrial proteins in diverse cellular processes necessitates efficient protein import into the organelle. Defects in mitochondrial protein-import have been linked to several disorders (Table S1 in (Nicolas et al., 2019).) In the next section, we briefly discuss how nuclear-encoded proteins are imported into the organelle.

1.3 Overview of Mitochondrial Protein Import

The majority of mitochondrial proteins are encoded by genes in the nuclear genome, synthesized on cytosolic ribosomes and then imported into the organelle (Dudek et al., 2013). These proteins contain distinct mitochondrial targeting signals which direct them to the organelle. These nuclear-encoded mitochondrial proteins not only have to be targeted and imported into the mitochondria, but also correctly sorted into the appropriate sub-organelar compartment - mitochondrial outer-membrane (OM), mitochondrial inner-membrane (IM), mitochondrial intermembrane-space (IMS) and mitochondrial matrix. Five major protein import pathways are known to be involved in mitochondrial protein import (Wiedemann and Pfanner, 2017).

Around 70% of mitochondrial proteins are imported via the classical presequence / mitochondrial targeting signal pathway (Fukasawa et al., 2015). Proteins imported via this pathway are synthesized with an N-terminus mitochondrial targeting presequence/signal (MTS). These MTS are found typically in the first 90 amino-acid residues. They are characterized by a high abundance of arginine and a depletion of negatively charged residues (Schneider et al., 1998). They form an amphipathic α-helix, which contains hydrophobic residues on one face and positively charged residues on the opposite face (von Heijne, 1986). The TOM complex (translocase of mitochondrial outer-membrane) and TIM complex (translocase of the mitochondrial inner-membrane) transport these proteins across the outer and inner-membranes respectively. The presequence translocase-

---

3OM: Outer Mitochondrial Membrane
4IM: Inner Mitochondrial Membrane
5IMS: Inter-Membrane Space
6MTS: Mitochondrial Targeting Signal
7TOM: Translocase of Mitochondrial outer-membrane
8TIM: Translocase of the Mitochondrial inner-membrane
associated motor (PAM)\(^9\) then imports these proteins into the mitochondrial matrix. Some proteins, however, contain a hydrophobic signal following the MTS. These proteins are inserted into the inner-mitochondrial membrane, instead of the matrix. After protein import, the MTS are cleaved by the Mitochondrial Processing Peptidase (MPP)\(^10\). Following this cleavage event, some proteins are cleaved by additional proteases, resulting in the functional mitochondrial protein (Wiedemann and Pfanner, 2017).

Other pathways are required for import and sorting of all outer-membrane and most inner-membrane mitochondrial proteins, which lack a cleavable MTS, but possess distinct internal mitochondrial targeting signals. Inner-mitochondrial proteins without cleavable MTS are imported by two different pathways - 1] The carrier pathway is responsible for the import of the large family of inner-membrane metabolite carrier proteins. They are transported across the outer mitochondrial membrane by the TOM complex like nearly all mitochondrial proteins. After that, small TIM chaperones in the intermembrane space and the TIM22 complex transport them to the inner-membrane. 2] The mitochondrial intermembrane space import and assembly (MIA)\(^11\) machinery is responsible for the transport of cysteine-rich inner-mitochondrial membrane proteins. Meanwhile, the outer-membrane mitochondrial proteins are imported and sorted by two different pathways - 1] \(\beta\)-barrel proteins of the mitochondrial outer-membrane follow the \(\beta\)-barrel pathway, in which they are transported across the outer-membrane by the TOM complex and inserted into the outer-membrane by the sorting and assembly machinery SAM\(^12\) complex. 2] Outer-membrane proteins with \(\alpha\)-helical transmembrane segments are imported by the mitochondrial import machinery (MIM)\(^13\) complex (Wiedemann and Pfanner, 2017).

Some mitochondrial proteins possess non-canonical targeting signals. \textit{e.g.} the C-terminus targeting signal in DNA helicase (Hmi1p) in yeast. Some of these non-canonical targeting signals can be found in LocSigDB, a database of manually-curated protein localization signals (Negi et al., 2015).

\(^9\)PAM: Presequence Translocase-Associated Motor  
\(^{10}\)MPP: Mitochondrial Processing Peptidase  
\(^{11}\)MIA: Mitochondrial Inter-membrane Space Import and Assembly  
\(^{12}\)SAM: Sorting and Assembly Machinery  
\(^{13}\)MIM: Mitochondrial Import Machinery
For multiple mitochondrial proteins, neither the import pathway, not the targeting signal are known. In addition to unidentified targeting signals and pathways, and additional complexity in mitochondrial protein import is introduced due to dual-localization of proteins. Dual-localized proteins are proteins which function in more than one subcellular compartment (Yogev and Pines, 2011). By one estimate, nearly a third of yeast mitochondrial proteins are dual-localized (Ben-Menachem et al., 2011). In addition to cataloging which proteins are imported into the mitochondria, it is also important to find out how they are imported. In the next section, we briefly summarize the different techniques used in the characterization of mitochondrial proteomes.

1.4 Characterization of the Mitochondrial Proteome

A complete inventory of all proteins functioning in mitochondria is critical for understanding mitochondrial biology, evolution, and pathogenesis. Currently, mitochondrial proteomes have been experimentally-characterized in all major groups of eukaryotes: animals (Pagliarini et al., 2008; Calvo et al., 2015; White et al., 2011; Li et al., 2009a; Hu et al., 2015), plants (Salvato et al., 2014; Lee et al., 2013; Rao et al., 2017), fungi (Morgenstern et al., 2017) and protists (Zhang et al., 2010; Gawryluk et al., 2014a; Seidi et al., 2018) (Table 1.1). Strategies for characterization of mitochondrial proteomes can be divided into two categories: experimental and computational.
1.4.1 Experimental approaches

The most widely-used experimental technique to identify mitochondrial proteins is to purify mitochondria, separate mitochondrial proteins, and use mass-spectrometry (MS) (Calvo and Mootha, 2010). MS-based approaches have significantly improved since the earliest mitochondrial proteome characterization analysis performed in 1998, in which a total of 46 mitochondrial proteins were identified from human placental mitochondria (Rabilloud et al., 1998). For instance, Pagliarini et al. identified 3881 mitochondrial proteins from 14 mouse tissues (Pagliarini et al., 2008). An extensive review of mammalian mitochondrial proteomic analyses is provided in (Chen et al., 2010). Other experimental techniques use microscopy for visual inspection of mitochondrial localization, like tagging proteins with GFP (Green Fluorescent Protein). Such experimental approaches are particularly useful in identifying mitochondrial proteins which lack identifiable/known targeting signals and proteins which lack homology to known mitochondrial proteins.

1.4.2 Computational approaches

The number of experimentally-characterized animal mitochondrial proteomes is small and their phylogenetic distribution is limited; all experimentally-characterized animal mitochondrial proteomes are from Bilateria (Chordata, Arthropoda and Nematoda). The lack of experimentally-characterized mitochondrial proteomes from non-bilaterian animals, as well as from additional phyla of bilaterian animals, prevents us from obtaining a broad view of animal mitochondrial proteome evolution. For instance, even well-studied model organisms like *Danio rerio* (zebrafish) lack complete mitochondrial proteomes. Computational techniques for identifying mitochondrial proteins, like mitochondrial targeting signal (MTS) prediction, orthology search, phylogenetic profiling, and co-expression analysis, play an important role in addressing this lack of information.
1.4.2.1 Prediction of mitochondrial targeting signals

The most common computational technique to identify mitochondrial proteins is to predict the N-terminus mitochondrial targeting signals (MTS). As discussed in Section 1.3, it is the predominant method of protein import into the mitochondria. While MTS are not conserved at the level of primary sequence, their properties (discussed in Section 1.3) are well conserved in eukaryotes. These properties are used by several software to identify mitochondrial proteins (TargetP (Emanuelsson et al., 2007), MitoFates (Fukasawa et al., 2015), Predotar (Small et al., 2004) etc. reviewed in (Rao et al., 2017)). There are several advantages of utilizing this approach to identify mitochondrial proteins. MTS-prediction can identify proteins that may be missed by MS-based techniques, like low-abundance proteins or proteins which exhibit tissue-specific expression. Thus, an MTS-based approach would be very useful to identify potential novel mitochondrial proteins in an organism. However, if a protein lacks an MTS or possesses an MTS with different properties than known MTS, current MTS detection software will miss it. As we see in our research (Chapter 3), MTS are well-conserved in ancestral conserved mitochondrial proteins, but several novel animal-specific mitochondrial proteins lack MTS. Additionally, MTS prediction tools also generate a large number of false-positives. One way to reduce this error is to use multiple predictors for generating a consensus result.

1.4.2.2 Orthology analysis

Orthologs are genes derived from a common ancestor by a speciation event (Lafond et al., 2016), as opposed to paralogs, where the homology is due to a gene-duplication event. Several tools have been developed for detection of orthologs. A common method for detection of orthologs is Reciprocal Best Blast Hit (RBBH). RBBH identifies two genes as orthologs if both genes are the best-scoring Blast hits of each other. Some tools, like In-Paranoid (Sonnhammer and Östlund, 2014), OrthoMCL (Li et al., 2003) and Proteinortho (Lechner et al., 2011) are extensions of this RBBH technique, which identify co-orthologs among multiple species. Co-orthologs are results of a gene-duplication event following speciation, due to which two or more genes in one species are
Orthologs of one or more genes in another species (Koonin, 2005). An in-depth comparison of recent orthology tools is provided in (Nichio et al., 2017).

Orthology analysis is useful in the identification of well-conserved mitochondrial proteins. However, this technique would miss potential novel mitochondrial proteins (which lack orthology to known mitochondrial proteins). Additionally, there are several known instances of neolocalization, i.e. a non-mitochondrial protein being recruited to mitochondria. In our research, we found that a large portion of neolocalized proteins in mammalian mitochondria do not possess a detectable MTS (Chapter 3). In such cases, both orthology-inference and MTS-prediction would lead to misannotation of the subcellular localization of proteins.

1.4.2.3 Other computational approaches

Another approach to identify mitochondrial proteins, which uses sequence homology, is using protein domain content to identify mitochondrial proteins. Such approaches search for mitochondria-specific protein domains to predict mitochondrial proteins. One such method, SubMitoPred (Kumar et al., 2018), which uses protein domain content as one of the predictive features, can also identify the sub-mitochondrial localization of a protein. Other techniques include co-expression analysis (analysis of expression profiles of mRNA or protein in multiple tissues), induction (identifying transcripts specifically up-regulated during mitochondrial biogenesis or proliferation) and phylogenetic profiling (prediction of gene-function based on gene-presence/absence across species) (Calvo et al., 2015).
1.4.3 Integrative approaches

The most comprehensive and accurate lists of mitochondrial proteins have been generated using integrative approaches, i.e. combining both experimental and computational techniques (Calvo et al., 2015; Smith and Robinson, 2018). Mitochondrial proteomes of multiple eukaryotes have been characterized using such combined approaches - mammals (human, mouse) (Calvo et al., 2015), the ecdysozoan D. melanogaster (Hu et al., 2015) and protists (Acanthamoeba castellanii (Gawryluk et al., 2014a)). In MitoCarta, multiple approaches (Mass-spectrometry, homology to yeast, MTS-prediction, protein-domain analysis, co-expression analysis, homology to Rickettsia, induction) were used to identify mitochondrial proteins in human and mouse. Similarly in IMPI (Integrated Mitochondrial Protein Index), mitochondrial proteomes of human, mouse and rat were identified using various techniques like Mass-spectrometry, GFP-tagging assays, MTS prediction by multiple targeting software, antibody analysis from Human Protein Atlas, and literature survey. Interestingly, IMPI identified a large number of nuclear genes encoding mitochondrial proteins which were missed by MitoCarta 2.0 (569). Conversely, 80 mitochondrial genes were identified by MitoCarta 2.0 and not by IMPI.

1.5 Databases for Mitochondrial Proteomes

Experimentally-verified animal mitochondrial proteomes have been deposited in various databases, most of which are dedicated solely to mammals. MitoMiner v4.0 (Smith and Robinson, 2018) is the most comprehensive mammalian mitochondrial proteome database. It includes reference mitochondrial gene-lists for human and mouse, based on results of MitoCarta 2.0 and IMPI. MitoMiner consolidates localization information from multiple sources, such as MS-based assays, GFP-tagging and microscopy, Gene-Ontology analysis, prediction of MTS, antibody staining results, KEGG pathway analysis, and disease information from OMIM (Online Mendelian Inheritance in Man) (Hamosh et al., 2005). Information regarding partial mitochondrial proteomes of rat and zebrafish are also provided in MitoMiner. Within mammals, most databases are dedicated to human mitochondria (MitoProteome (Cotter et al., 2004), HMPdb (https://bioinfo.nist.gov/)). Outside of mammals,
the most comprehensive mitochondrial proteome database has been developed for the arthropod model-organism *D. melanogaster*. The GLAD database (Gene List Annotation for Drosophila) integrates data from multiple sources - Gene-Ontology, Uniprot annotations, other mitochondrial databases and literature (Hu et al., 2015). Some animal mitochondrial proteomes have not been deposited in any database - *C. elegans* and rabbit (White et al., 2011). Thus, valuable information regarding animal mitochondrial proteomes is scattered across multiple databases, and no one database houses the mitochondrial proteomes of vertebrate and invertebrate model organisms. In Chapter 4, we address this need, and create the Metazoan Mitochondrial Proteome Database (MMPdb), which consolidates data on mitochondrial proteins from human, mouse, *C. elegans* and *D. melanogaster* to facilitate comparative analysis of animal mitochondrial proteomes. In the next section, we take a brief look at animal phylogeny, with the sole purpose of introducing the different animal datasets used in the dissertation.

### 1.6 An Introduction to Metazoan Phylogeny

In Chapters 2, 3 and 4, we use data from both bilaterian and non-bilaterian animal species to analyze animal mitochondrial proteome evolution. The goal of this section is to give a brief introduction to animal phylogeny. All animals belong to one of the five main branches of the animal phylogeny: Bilateria (represented by multiple phyla) and the four phyla of non-bilaterian animals: Porifera (sponges), Cnidaria (jellyfish, coral, anemones), Ctenophora (comb jellies) and Placozoa (placozoans) (Ax, 2012). Bilateria comprises of over a million described species, and majority of the metazoan phyla. Bilaterian animals can be further divided into two large groups: Protostomia and Deuterostomia (and possibly, Xenacoelomorpha). Dueterostomes are comprised of chordates (Cephalochordata, Urochordata and Vertebrata) and ambulacrarians (Echinodermata and Hemichordata). In protostomes, there are three major lineages- Chaetognatha, Ecdysozoa, and Spiralia (Dunn et al., 2014). The two model organisms: *C. elegans* and *D. melanogaster* belong to phylum Nematoda and phylum Arthropoda, respectively, within Ecdysozoa.
The non-bilaterian animals are comprised of four phyla- Porifera, Cnidaria, Ctenophora and Placozoa. Sponges are further divided into four classes, based on cellular organization, chemical composition of protective structures called spicules and molecular phylogenetic analysis: Demospongiae, Homoscleromorpha, Hexactinellida and Calcarea (Renard et al., 2018; Wörheide et al., 2012). Cnidarians are subdivided into two major groups - Anthozoa (sea anemones, gorgonians, soft and stony corals) and Medusozoa (hydroids, jellyfish, syphonophores) (Kayal et al., 2013). These also include a group of obligate intracellular parasites - Myxozoa and Polypodiozoa (Chang et al., 2015). Placozoans are represented by just two known species (Srivastava et al., 2008; Eitel et al., 2018). Ctenophora is comprised of two classes, distinguished by the presence/absence of their feeding tentacles: Tentaculata (present) and Nuda (absent). The phylogenetic relationship between the four non-bilaterian phyla and bilaterian animals is still under debate (Dunn et al., 2015; Dohrmann and Wörheide, 2013; Pisani et al., 2015; Whelan et al., 2015), but does not affect results of any of our analyses.

Unlike bilaterian animals, there is substantial variation in the mitochondrial genomes of non-bilaterian animals, with respect to gene-content, rate of evolution, alternate genetic codes and genome architecture (Lavrov and Pett, 2016). For instance, linear mtDNA, instead of the circular molecule in mammals, is present in Medusozoa (Cnidaria) and Calcarea (Porifera). Mitochondrial genomes of all calcareous sponges sequenced to date are linear and multipartite. e.g., all genes in Leucosolenia complicata are present on individual linear chromosomes. There is considerable variation in the size of mitochondrial genomes in non-bilaterian animals; the smallest animal mitochondrial genome is found in the ctenophore Mnemiopsis leidyi (10kb), while one of the largest animal mitochondrial genomes is from the calcareous sponge Clathrina clathrus (51kb). Compared to bilaterian animals, the protein-coding gene content of non-bilaterian animals is more variable, including additional mitochondrial genes (atp9 in all sponges) and loss of mitochondrial genes (atp6 and atp8 in ctenophores). For a detailed review on non-bilaterian mitochondrial genomic diversity, see (Lavrov and Pett, 2016). The non-bilaterian phyla encompass most of the mitochondrial genomic diversity witnessed in animals. We hypothesize that these changes would be reflected in
their mitochondrial proteomes. Thus, to understand the evolution of animal mitochondrial proteome, it is necessary to include species from both non-bilaterian and bilaterian phyla. However, all of the experimentally verified mitochondrial proteomes in animals are from bilaterian animals. Hence, in Chapter 2, we used bioinformatic methods to predict and characterize mitochondrial proteomes from the four non-bilaterian phyla to obtain a broad picture of conservation and evolution of mitochondrial proteomes in animals.

1.7 Dissertation Organization

This dissertation consists of a general introduction (Chapter 1), three journal chapters (Chapters 2-4) and general conclusions (Chapter 5).

Chapter 1 provides a general introduction and details the background of this dissertation.

Chapter 2 includes a published manuscript, in which we use computational methods to predict and characterize mitochondrial proteomes in the four phyla of non-bilaterian animals (Porifera, Cnidaria, Ctenophora, and Placozoa). The manuscript is published under the title “Characterization of mitochondrial proteomes of nonbilaterian animals.” IUBMB life 70.12 (2018): 1289-1301. Data and code for this analysis can be found on the project page on Open Science Framework: Muthye, Viraj, and Dennis Lavrov. 2018. Analysis of Mitochondrial Proteomes in Non-Bilaterian Animals (August 2018). OSF. September 4. doi:10.17605/OSF.IO/WJHDS.

Chapter 3 includes a manuscript under review in the journal Mitochondrion. In this manuscript, we identify the causes and functional implications of variation in the size of mitochondrial proteomes in bilaterian animals, by performing a comparative analysis of the publicly available mitochondrial proteomes of human, mouse, Caenorhabditis elegans and Drosophila melanogaster. Data and code for this analysis can be found on the project page at the Open Science Framework: Muthye, Viraj, and Dennis Lavrov. 2019. Data for Causes and Consequences of Animal Mt-Proteome Size Variation. OSF. June 25. osf.io/a49yw.
Chapter 4 includes a manuscript under review in the journal Mitochondrion. In this, we develop two tools for facilitating comparative analysis of mitochondrial proteomes in animals—1) the Metazoan Mitochondrial Proteome Database (MMPdb), which consolidates data on the experimentally-characterized mitochondrial proteomes of vertebrate and invertebrate animals, and 2) MitoPredictor, a novel machine-learning tool for predicting mitochondrial proteins in animals. Data and code for MMPdb can be found on the project page at the Open Science Framework: Muthye, Viraj, Dennis Lavrov, and Gaurav Kandoi. 2019. Data for Metazoan Mitochondrial Proteome Database (MMPdb). OSF. July 2. osf.io/gfyq9. Data and code for the pipeline “mitopredictor” can be found on the project page on Open Science Framework : Muthye, Viraj. 2019. Data for Mitopredictor. OSF. July 1. osf.io/sprzy].

Chapter 5 outlines the general conclusions to the dissertation and provides future directions for research and improvements which can be made to the dissertation.

1.8 References


## 1.9 Tables and Figures

Table 1.1: List of experimentally-characterized mitochondrial proteomes in eukaryotes.

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of mitochondrial proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Animals</strong></td>
<td></td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>~1700</td>
</tr>
<tr>
<td><em>Mus musculus</em></td>
<td>1519</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td>1090</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>838</td>
</tr>
<tr>
<td><em>Oryctolagus cuniculus</em></td>
<td>995</td>
</tr>
<tr>
<td><strong>Plants</strong></td>
<td></td>
</tr>
<tr>
<td><em>Solanum tuberosum</em></td>
<td>1060</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>1005</td>
</tr>
<tr>
<td><strong>Fungi</strong></td>
<td></td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>~1000</td>
</tr>
<tr>
<td><strong>Protists</strong></td>
<td></td>
</tr>
<tr>
<td><em>Toxoplasma gondii</em></td>
<td>400</td>
</tr>
<tr>
<td><em>Acanthamoeba castellanii</em></td>
<td>1033</td>
</tr>
<tr>
<td><em>Trypanosoma brucei</em></td>
<td>1008</td>
</tr>
</tbody>
</table>
Figure 1.1: The taxonomic groups used in Chapters 2, 3 and 4 of the dissertation. In Chapter 2, the mitochondrial proteomes of non-bilaterian animals were inferred and the experimentally-characterized mitochondrial proteomes of human (Chordata) and yeast (Fungi) were used. In Chapters 3 and 4, the experimentally-characterized mitochondrial proteomes of human (Chordata), mouse (Chordata), *Caenorhabditis elegans* (Nematoda), and *Drosophila melanogaster* (Arthropoda) were used. The experimentally-characterized mitochondrial proteomes of yeast (Fungi) and the protist *Acanthamoeba castellanii* were used as outgroups.
CHAPTER 2. CHARACTERIZATION OF MITOCHONDRIAL PROTEOMES OF NONBILATERIAN ANIMALS

Modified from a manuscript published in IUBMB Life

Viraj Muthye\textsuperscript{1,2} and Dennis Lavrov\textsuperscript{1,2}

\textsuperscript{1} Bioinformatics and Computational Biology Program, Iowa State University, 2437 Pammel Drive, Ames, Iowa 50011, USA
\textsuperscript{2} Department of Ecology, Evolution and Organismal Biology, Iowa State University, 241 Bessey Hall, Ames, Iowa 50011, USA

2.1 Abstract

Mitochondria require \( \sim 1,500 \) proteins for their maintenance and proper functionality, which constitute the mitochondrial proteome (mt-proteome)\textsuperscript{1}. Although a few of these proteins, mostly subunits of the electron transport chain complexes, are encoded in mitochondrial DNA (mtDNA)\textsuperscript{2}, the vast majority are encoded in the nuclear genome and imported to the organelle. Previous studies have shown a continuous and complex evolution of mt-proteome among eukaryotes. However, there was less attention paid to mt-proteome evolution within Metazoa, presumably because animal mtDNA and, by extension, animal mitochondria are often considered to be uniform. In this analysis, two bioinformatic approaches (ortholog-detection and Mitochondrial Targeting Sequence prediction) were used to identify mt-proteins in 23 species from four nonbilateral phyla: Cnidaria, Ctenophora, Placozoa, and Porifera, as well as two choanoflagellates, the closest animal relatives. Our results revealed a large variation in mt-proteome in nonbilateral animals in

---

\textsuperscript{1} mt-proteome: mitochondrial proteome
\textsuperscript{2} mtDNA: mitochondrial DNA
size and composition. Myxozoans, highly reduced cnidarian parasites, possessed the smallest inferred mt-proteomes, while calcareous sponges possessed the largest. About 513 mitochondrial orthologous groups were present in all nonbilaterian phyla and human. Interestingly, 42 human mitochondrial proteins (mt-proteins) were not identified in any nonbilaterian species studied and represent putative innovations along the bilaterian branch. Several of these proteins were involved in apoptosis and innate immunity, two processes known to evolve within Metazoa. Conversely, several proteins identified as mitochondrial in nonbilaterian phyla and animal outgroups were absent in human, representing cases of possible loss. Finally, a few human cytosolic proteins, such as histones and cytosolic ribosomal proteins, were predicted to be targeted to mitochondria in nonbilaterian animals. Overall, our analysis provides the first step in characterization of mt-proteomes in nonbilaterian animals and understanding evolution of animal mt-proteome.

2.2 Introduction

Mitochondria are membrane-bounded organelles present in most eukaryotic cells and involved in crucial cellular functions, including Fe/S (ironsulfur) cluster biosynthesis, metabolism of amino and fatty acids, and generation of energy via oxidative phosphorylation (Nunnari and Suomalainen, 2012). According to the prevailing view, the origin of mitochondria lies in an ancient endosymbiotic event between an α-proteobacterium, possibly from the order Rickettsiales (Andersson et al., 1998; Fitzpatrick et al., 2005; Gray et al., 1999) but see (Martijn et al., 2018), and an archaean host, most closely related to a recently discovered group called the Asgards (Martin et al., 2016; Spang et al., 2015; Zaremba-Niedzwiedzka et al., 2017). Although all energy-producing mitochondria retained their own genome (mtDNA), its gene content is highly reduced compared to that of modern α-proteobacteria (Petersen et al., 2014; Burger et al., 2013). In particular, mt-genomes in animals contain between 14 and 41 genes and encode 12 to 15 known proteins, nearly all of them being subunits of oxidative phosphorylation (OXPHOS) complexes I, III, IV, and, usually

---

3 mt-protein: mitochondrial protein
4 mtDNA: mitochondrial DNA/genome
5 OXPHOS: oxidative phosphorylation
but not always, V (Lavrov and Pett, 2016). Such limited coding capacity of mtDNA would not allow autonomous existence and/or function of mitochondria and necessitates a massive import of nuclear encoded proteins (Meisinger et al., 2008).

Thanks to the protein import, the number of distinct proteins found in mitochondria is estimated to be between 1,000 and 2,000 or about two orders of magnitude greater than the number of proteins encoded in mt-genomes (Meisinger et al., 2008; Calvo et al., 2015; Gawryluk et al., 2014a; Salvato et al., 2014; Millar et al., 2001). Among animals, human and mouse mitochondrial proteomes (mt-proteomes) have been well characterized, with several curated databases publicly available (e.g., Mitocarta, Integrated Mitochondrial Protein Index (IMPI) (Calvo et al., 2015; Smith and Robinson, 2015, 2009)). In fungi, large-scale proteomic studies have been performed in yeast Saccharomyces cerevisiae (Reinders et al., 2006; Renvoisé et al., 2014), with the Saccharomyces Genome Database (SGD) (http://www.yeastgenome.org/) providing a well-curated list of mitochondrial proteins (mt-proteins) (Cherry et al., 2012). In plants, mt-proteomes have been analyzed in several species (reviewed in (Rao et al., 2017)), including Solanum tuberosum (potato) and Arabidopsis thaliana (the thale cress) (Salvato et al., 2014; Millar et al., 2001; Heazlewood et al., 2004). In addition, mt-proteomes have been characterized for a few unicellular eukaryotes (Gawryluk et al., 2014a; Zhang et al., 2010).

Comparison of mt-proteomes across eukaryotic lineages revealed a mosaic of protein conservation, acquisition and loss. One driving force behind mt-proteome evolution appears to be interactions between nuclear-encoded proteins and mt-encoded molecules (Hashimoto et al., 2001; Fuku et al., 2015). Examples of such cyto-nuclear co-evolution include changes in nuclear-encoded mitochondrial ribosomal proteins, possibly in response to reductive evolution in mt-rRNA structures (Sharma et al., 2003) and loss of nuclear-encoded proteins following the loss of mitochondrial tRNA genes (Pett and Lavrov, 2015). Another driving force in mt-proteome evolution is acquisition of novel functions by mitochondria, such as intrinsic apoptosis (Elmore, 2007; Oberst et al., 2008) and innate immunity (Mills et al., 2017) in animals. Interestingly, changes have been observed even

---

6 mt-proteome: mitochondrial proteome
7 IMPI: Integrated Mitochondrial Protein Index
within the otherwise well-conserved oxidative phosphorylation machinery, including the loss of the whole complex 1 (NADH-ubiquinone oxidoreductase) in fungi (Marcet-Houben et al., 2009; Gabaldón et al., 2005a; Hirst, 2013), loss of complexes I and III in Chromera velia, a phototrophic relative of parasitic apicomplexans (Flegontov et al., 2015), and addition of tissue-specific subunits in mammals (Pagliarini et al., 2008).

Although we know that the mt-proteome evolved extensively across eukaryotic lineages, we know far less about its evolution within animals. This is because most mt-proteomic studies have been focused on mammals (but see (Lotz et al., 2013; Alonso et al., 2005)), which represent one phylum (Chordata) in one of the five major lineages of animals (Bilateria). The other four major branches of animal phylogenetic tree are formed by the so-called nonbilaterian animals: phyla Cnidaria, Ctenophora, Placozoa, and Porifera (Fig. 2.1). The origin of these major animal lineages can be traced back to the Precambrian, the time at which environmental conditions had been very different from those in recent seas. In particular, the oxygen concentration had drastically increased during the Precambrian-Cambrian transition, potentially triggering the Cambrian explosion of macroscopic animals (Chen et al., 2015). Thus, mitochondria in different lineages of animals had to adapt independently to higher environmental oxygen concentration, as well as larger size and greater morphological complexity of modern animals that resulted from this change. Our previous studies have shown that nonbilaterian phyla encompass most of the mitochondrial genomic diversity observed in animals, as manifested by their various genomic architecture, gene content, genetic code, nucleotide composition, and rates of sequence evolution (Lavrov and Pett, 2016). Mt-proteomes in these lineages had to accommodate these changes.

As an initial step in investigating evolution of animal mt-proteome, we searched for mt-proteins in 23 animal species representing all four nonbilaterian phyla, using publicly available transcriptomic resources (Table 2.1). We also identified mt-proteins in two choanoflagellate species, which, along with the well-characterized yeast mt-proteome, serve as outgroups for our analysis.

There are two main approaches applicable for computational characterization of mt-proteomes: by sequence similarity to known mitochondrial counterparts and through identification of mito-
chondrial targeting sequences (MTS) \(^8\). The first approach aims to identify orthologs of known mt-proteins and has been implemented in a variety of programs reviewed in References (Altenhoff et al., 2016; Tekaia, 2016), including Proteinortho (Lechner et al., 2011). Proteinortho detects co-orthologs in protein datasets and has been shown to perform as well as OrthoMCL (Lechner et al., 2011; Li et al., 2003), one of the most commonly used tools in the field, but is faster and hence better suited for analyses of larger datasets. The second approach predicts mt-proteins by the presence of MTS. Although several import pathways exist in mitochondria (Harbauer et al., 2014; Fox, 2012; Chacinska et al., 2009), the majority of the mitochondrial matrix proteins, along with many inner-membrane and inter-membrane space proteins, are imported via the MTS pathway (Vögtle et al., 2009) that depends on the presence of an MTS- a short (15-50 aa) sequence at the N-terminus of a peptide, usually enriched in hydrophobic and positively charged amino acids. Upon import, the MTS is cleaved, resulting in a mature protein. Several software exist to predict the MTS (see Supporting Information table S3 in (Rao et al., 2017)). Among them, TargetP (Emanuelsson et al., 2000, 2007), a neural network-based tool and MitoFates (Fukasawa et al., 2015), a support vector machine (SVM) classifier-based approach, are commonly used. MTS prediction is particularly attractive for de-novo characterization of mt-proteome because it does not require prior information on homologous molecules.

Here, we show that the inferred mt-proteomes in nonbilaterian animals are diverse in size and in protein content. We identified both proteins shared across all animals and proteins present only in nonbilaterian animals. In addition, we used a set of conserved mt-proteins (from the tricarboxylic acid cycle [TCA]\(^9\)) to analyze MTS in nonbilaterian animals. Finally, we analyzed the diversity of protein domains—evolutionarily conserved functional modules—-in inferred mt-proteins, as an additional approach to investigate the functional diversity of the inferred mt-proteomes.

---

\(^8\)MTS: mitochondrial targeting signal

\(^9\)TCA: tricarboxylic acid cycle
2.3 Materials and Methods

2.3.1 Predicting mitochondrial proteomes

The overall approach of the study is shown in Figure 2.2. Assembled transcriptomes of 23 nonbilaterian species and two choanoflagellate species (Table 2.1) were downloaded from public databases (Supporting Information S1). Encoded proteins were predicted with Transdecoder v.5.3.0 ((Haas et al., 2013), http://transdecoder.github.io) with the default settings. Mt-proteomes were inferred using two computational approaches: (i) ortholog-detection, utilizing Proteinortho v5.16b (Lechner et al., 2011) against reference mt-proteomes of human and yeast and (ii) detection of MTS using TargetP v1.1 (Emanuelsson et al., 2000, 2007) and MitoFates (Fukasawa et al., 2015). All datasets and scripts used in this analysis are available on the Open Science Framework project page (https://osf.io/wjhds/) (Muthye, Viraj, and Dennis Lavrov. 2018. Analysis of mt-proteomes in Non-Bilaterian Animals (August 2018). Open Science Framework. September 3. osf.io/wjhds).

In some of the transcriptomes used, we identified a large percentage of identical transcripts, which would have been problematic for our analysis. Thus, complete predicted proteomes of nonbilaterian animals and choanoflagellates were clustered with CD-HIT (Fu et al., 2012; Li et al., 2001) with the sequence identity cutoff of 98%. CD-HIT clusters proteins based on a sequence similarity threshold and returns one representative sequence for each group.

2.3.1.1 Approach 1: Ortholog-detection by Proteinortho

We compiled a reference human mt-proteome, using the lists of mt-proteins from two sources: Integrated mt-protein Index (IMPI) vQ2 and MitoCarta v2.0. The yeast mt-proteome was downloaded from the Saccharomyces Genome Database (Martin et al., 2016).
Proteinortho v5.16b was run on the total proteome of human and yeast (downloaded from Uniprot (TheUniProtConsortium, 2016)) and the predicted nonbilaterian and choanoflagellate proteomes using default parameters: an e-value of 1e-05, 25% identity and 50% coverage cut-offs for BlastP searches. From the resulting orthologous groups (OG)\textsuperscript{10}, those containing the reference human and yeast mt-proteins were extracted.

### 2.3.1.2 Approach 2: Mitochondrial targeting signal (MTS)-prediction

TargetP v1.1 (Emanuelsson et al., 2000, 2007) and MitoFates (Fukasawa et al., 2015) were used to identify proteins targeted to the mitochondria by detecting MTS. MitoFates was run using the Metazoa option for animals and choanoflagellates and the Fungi option for yeast. TargetP was run using the Non-plant version. Mitochondrial processing peptidase (MPP)\textsuperscript{11} cleavage sites identified by MitoFates were used to extract the predicted MTS in the mt-proteins for further analysis. The MTS was defined as the part of the sequence of the protein from the first residue at the N-terminus end till the MPP cleavage site. We categorized a protein as mitochondrial if it was predicted to be targeted to mitochondria by both MitoFates and TargetP (all reliability classes [RCs]\textsuperscript{12}). TargetP outputs a RC for each prediction, which is an indication of the strength of that prediction. RCs range from 1 to 5, where 1 denotes the strongest prediction and 5 denotes the weakest. The rationale for this requirement, which a protein will be considered as mitochondrial only when both predictors detect a MTS, was to decrease the number of false positives and thus to increase confidence in prediction of MTS. The results can be demonstrated on the characterized human mt-proteome from MitoCarta v.2, which consists of 1,158 mitochondrial genes. TargetP alone identified 759 of these mt-proteins when all RC were used and also had 2,277 false positives (non-mt-proteins identified as mitochondrial). Most of these false positives were in lower RC, but 884 had RC1-3. MitoFates identified 576 mt-proteins as mitochondrial, but also had 365 false positives. When the two searches were combined, the number of true positive decreased < 5%

\textsuperscript{10}OG: orthologous group \\
\textsuperscript{11}MPP: mitochondrial processing peptidase \\
\textsuperscript{12}RC: reliability class
relative to the MitoFates search (to 548), but the number of false positive declined 26% (to 270). Assuming that these results can be extended to other animals, we would expect ∼67% of proteins identified based on the MTS prediction in our study to be mitochondrial. Although we also expect to miss ∼50% of mt-proteins by this approach, some of these proteins should be identified by the orthology-search.

Presence of a complete N-terminus end is critical for MTS prediction. In case of a fragmented protein, lacking some or all of the N-terminus, the prediction software results will be based on the N-terminus end of the fragment and incorrect. Therefore, for the MTS analysis, only proteins, with a methionine at position one, were used. The resulting lists of proteins from approach 1 and approach 2 were then combined to generate the inferred mt-proteome for each analyzed species.

2.3.1.3 Removal of possible non-metazoan contamination

Next, for each of these inferred mt-proteomes, a BlastP search was carried out against the NR database. In all the nonbilaterian species, only proteins that possessed a metazoan top blast hit were retained for further analysis. In the two choanoflagellate species, only proteins that possessed a eukaryotic top blast hit were retained.

2.3.2 Analyses of inferred mt-proteomes

2.3.2.1 General characterization

Inferred mt-proteins were divided into three categories: proteins predicted to be mitochondrial by approach 1 alone (proteins orthologous to reference mt-proteins but without recognized MTS), by approach 2 alone (proteins with a detectable MTS but no orthologs in reference mt-proteins), and by both approaches. BoxPlotR (http://shiny.chemgrid.org/boxplotr/) was used to plot lengths of inferred mt-proteomes, complete predicted proteomes and transcriptomes. We identified a set of common animal mt-proteins (CAMPs) from the Proteinortho results. In this study, we define a CAMP as an orthology group that contains a human mt-protein, as well as proteins from at least 13

13CAMP: common animal mt-proteins
six sponges, at least three cnidarians, at least one ctenophore and the placozoan. PANTHER v.13.1
Overrepresentation Test (Released 20171205) (Mi et al., 2016) was carried out to identify which
gene ontology (GO) biological process terms were enriched in CAMPs.

2.3.2.2 Identification of proteins involved in selected mitochondrial pathways

Lists of human proteins involved in the TCA cycle (TCA), oxidative phosphorylation (OX-
PHOS) pathway, Fe/S cluster assembly (FES)\textsuperscript{14}, mitochondrial biogenesis (MtBio)\textsuperscript{15}, mitochondrial
contact site and cristae organizing system (MICOS)\textsuperscript{16} complex and Intrinsic pathway of apoptosis
were obtained from the KEGG database and literature (Elmore, 2007; Kanehisa and Goto, 2000;
Kanehisa et al., 2016, 2015; Stehling et al., 2014; Huynen et al., 2016). Lists of mitochondrial
ribosomal proteins were downloaded from the Ribosomal Protein Gene Database (Nakao et al.,
2004). Corresponding amino-acid sequences were downloaded from Uniprot. OGs containing these
proteins were extracted from the Proteinortho results and inspected for the presence of orthologs
in analyzed species.

2.3.2.3 Protein domain analysis using PFAM

The Perl script pfam\_scan.pl (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/Tools/) with PFAM
31.0 HMM libraries was used to identify protein domains (Finn et al., 2015); Venny v2.1 (Oliveros,
2007) was used to visualize the results. GO analysis of the protein domains was performed using
dcGO (Fang and Gough, 2012). We wrote a bash script to identify the most abundant protein
domains in the mt-proteomes and to identify a set of protein domains shared by all nonbilaterian
animals.

2.3.2.4 Mitochondrial targeting signal (MTS) Analysis

The MPP cleavage site prediction by MitoFates was used to extract the MTS from the mt-
proteins. Sequence length and amino-acid composition of the MTS were calculated. Principle

\textsuperscript{14}FES: Fe/S cluster assembly
\textsuperscript{15}MtBio: mitochondrial biogenesis
\textsuperscript{16}MICOS: mitochondrial contact site and cristae organizing system
component analysis (PCA)\textsuperscript{17} of amino-acid composition of the MTS was performed and visualized using ClustVis (Metsalu and Vilo, 2015). BoxPlotR (http://shiny.chemgrid.org/boxplotr/) was used to plot lengths of the inferred MTS. The net charge was calculated using the Peptides package V2.3 in R, using the Henderson-Hasselbalch equation and the EMBOSS pKa scale.

**Supplementary Materials**


### 2.4 Results and Discussion

#### 2.4.1 Mitochondrial proteomes in nonbilaterian animals

We used two bioinformatic approaches- ortholog-detection by Proteinortho and identification of MTS by TargetP and MitoFates- to infer mt-proteomes for 23 nonbilaterian species: 13 sponges representing all four Poriferan classes: Calcarea, Demospongiae, Hexactinellida, and Homoscleromorpha; seven cnidarians with species from the two major cnidarian lineages, Anthozoa and Medusozoa, as well as Myxozoa and Polypodiozoa; two ctenophores; and one placozoan. In addition, mt-proteomes were inferred for choanoflagellates *Monosiga brevicollis* and *Salpingoeca rosetta*, the closest known relatives of animals.

Among nonbilaterian species, the size of the inferred mitochondrial proteome was on average 1,054 proteins and ranged from 454 proteins in the myxozoan *Kudoa iwatai* to 2,119 proteins in the calcarean sponge *Leucosolenia complicata* (Figure 2.3, Supporting Information S2). Substantial variation in size was also observed within each nonbilaterian phylum. For instance, within Porifera, the glass sponge *Hyalonema populiferum* possessed the smallest inferred mt-proteome (493

\textsuperscript{17}PCA: Principle component analysis
proteins), while the calcarean sponges *L. complicata* possessed the largest (2,119 proteins). The majority of inferred mt-proteins in each species were identified by approach 1 (ortholog-detection by Proteinortho, mean = 832) (Supporting Information S2). The number of such proteins varied between 271 in *Pleurobrachia bachei* to 1,433 in *L. complicata*. Fewer proteins were identified by approach 2 (MTS prediction, mean = 348), ranging from 91 in *Kudoa iwatai* to 1,018 in *Sycon ciliatum*. Interestingly, the number of proteins identified by both approaches was small (between 27 in *K. iwatai* and 272 in *S. ciliatum* and *L. complicata*; mean = 127).

The large number of mt-proteins identified in *L. complicata* by approach 1 was, in part, because of the detection of multiple co-orthologs (several genes in one lineage which are orthologous to one or more genes in another lineage due to lineage-specific duplication). Thus, 145 mitochondrial proteins present as a single-copy protein in human were represented by multiple co-orthologs in *L. complicata*. These proteins included subunits of the TCA cycle, oxidative phosphorylation, and mitochondrial translation (Supporting Information File S1). In addition, *L. complicata* had one of the largest number of proteins identified by presequence prediction alone (686). In particular, we identified 87 cases of possible neolocalization following duplication in *L. complicata*, that is, OGs where at least one ortholog from *L. complicata* was predicted to possess a presequence and at least one that was not. Some examples included ribosomal protein L27 (RL27), Aflatoxin B1 aldehyde reductase member 2 (ARK72), succinate-CoA ligase subunit beta (SUCB2) and Ironsulfur cluster assembly enzyme (ISCU) (Supporting Information File S6).

At the other extreme, the low number of mt-proteins detected in *K. iwatai* was due to both the low number of orthologs detected (390) as well few proteins bearing MTS (91). Human mt-proteins lacking orthologs in *K. iwatai* included several critical mitochondrial proteins. This is consistent with previous reports that myxozoans, being endoparasites, have lost several well-conserved genes found in most multicellular animals (Chang et al., 2015). However, further analysis revealed that some of these results could be explained by the strict nature of the utilized ortholog-detection pipeline combined with high rates of sequence evolution in Myxozoa. To that end, we looked at TCA, FES, and the MICOS complex proteins in the complete predicted proteome of *K. iwatai*, and
used HMMER v.3.1b2 (Eddy, 2011) to identify the proteins missed by ortholog detection. HMM profiles for individual proteins were downloaded from EggNOG v.4.5.1 (Huerta-Cepas et al., 2015). In TCA and FES cycle, we identified homologs of all proteins missed by our ortholog detection. Conversely, we only identified a fragment of one possible homolog of MICOS complex subunit MIC60.

The mean length of inferred proteins varied substantially among the species from 291 amino acids in the cnidarian *Aurelia aurita* to 453 amino acids in the sponge *Aphrocallistes vastus* (Supporting Information File 8). Notably, in most analyzed species, the mean length of inferred mt-proteins was smaller than that in reference mt-proteomes human (450 amino acids) and yeast (508 amino acids). This result is likely a reflection of incomplete mRNA sequences in assembled transcriptomes and/or addition of some small peptides predicted to be targeted to mitochondria.

### 2.4.2 Identification of common animal mt-proteins (CAMPs) and bilaterian-specific mt-proteins

To identify the core of mt-proteome in animals, we compared inferred mt-proteomes in nonbilaterian animals among themselves and with human mt-proteins. We identified 513 proteins shared across all animal groups used in this study given the criteria explained in the section METHODS AND MATERIALS and called them Common Animal mt-proteins (CAMPs). The average number of CAMPs detected in studied nonbilaterian species was 381, with both the smallest (178) and largest (492) number coming from the phylum Ctenophora (*P. bachei* and *M. leidyi*, respectively; (Supplementary materials S3). As expected, central mitochondrial processes were over-represented in the CAMPs dataset when compared to the total human proteome, including mitochondrial transport, ferredoxin metabolic process, fatty acid beta-oxidation, tricarboxylic acid cycle, glycolysis, mitochondrial translation, oxidative phosphorylation and tRNA aminoacylation for protein translation (Supporting Information S3).

Next, we identified human mt-proteins that did not have orthologs in any species in this study and represent potential innovations in the bilaterian lineage. In addition, we searched for homologs of
these proteins in the predicted mt-proteomes with HMMer using profiles downloaded from EggNOG (Huerta-Cepas et al., 2015). Only 42 human mt-proteins did not have a homolog in any nonbilaterian species studied. Several of these proteins were involved in the intrinsic pathway of apoptosis and its regulation, including Bcl2-associated agonist of cell death (BAD), BH3-interacting domain death agonist (BID), BH3-like motif-containing cell death inducer (BLID), Bcl-2-binding component 3 (BBC3), and Diablo homolog (DIABLO). Two other proteins were related to sperm mitochondrial processes: sperm mitochondrial-associated cysteine-rich protein (SMCP) (Hawthorne et al., 2006) and spermatogenesis-associated protein 19 (SPATA19) (Geng et al., 2017). In addition, proteins not found in nonbilaterian taxa included hypoxia and HIF-1-induced mt-protein (HUMMER), which is involved in mitochondrial transport and steroidogenesis (Jin et al., 2012; Li et al., 2009b), α-synuclein (SNCA), a presynaptic protein known to modulate mitochondrial fusion and morphology and implicated in Parkinsons disease (Bendor et al., 2013; Menges et al., 2017), and syntaphilin (SNPH), which is an outer-mitochondrial membrane protein responsible for immobilizing mitochondria in axons (Chen and Sheng, 2013; Lin et al., 2017; Kang et al., 2008) (Supporting Information File S2). Given that these proteins were not identified in any nonbilaterian species, it would be interesting to investigate their origin and evolution by analysis of bilaterian mt-proteomes.

2.4.3 Identification of predicted nonbilaterian mt-proteins with no ortholog in the reference mt-proteomes

The number of proteins predicted to be mitochondrial by MTS identification but lacking orthologs in human and yeast mt-proteomes (MTS-only proteins) varied between 45 in *H. populiferum* and 746 in *S. ciliatum* (Supporting Information S2). Some of these MTS-only proteins were orthologs of human non-mt-proteins, ranging from 4 in *H. populiferum* to 79 in *S. ciliatum*. Several of them were identified across multiple nonbilaterian mt-proteomes. For example, members of the histone H2A family were predicted to possess a MTS in five sponges, two cnidarians, and one ctenophore. Interestingly, proteins from the H2A family have been previously identified in mitochondrial membranes, with one member, H2AX, being possibly involved in the mitochondrial
transport (Choi et al., 2011). Furthermore, orthologs of the checkpoint protein HUS1 (HUS1) involved in DNA repair (Guan et al., 2007) were predicted to possess a MTS in six sponges and one cnidarian. Other nonmitochondrial human proteins predicted to possess an MTS in multiple nonbilaterian species included AP-1 complex subunit gamma-1 (AP1G1) (seven species), tRNA (uracil-5)-methyltransferase homolog A (TRMT2A), 40S ribosomal protein S24 (five species) and 60S ribosomal protein L21 (five species).

Next, we analyzed MTS-only proteins that did not possess an ortholog in any reference proteome (the complete human and yeast proteomes) but were found in more than one nonbilaterian species. The number of such proteins ranged from 14 in *K. iwatai* to 329 in *S. ciliatum*. Although none of these proteins were identified across all the nonbilaterian species in this study, alternative oxidase (AOX) was found in the majority of sponges (seven species), one cnidarian, and the placozoan. AOX provides an alternative branch point into the electron transport chain by accepting electrons from ubiquinol and reducing oxygen to water, thus bypassing complex III and complex IV. It has been previously characterized in other animals, plants, fungi, and protists but not in vertebrates (McDonald et al., 2009).

Finally, we identified MTS only proteins, which did not have orthologs in any other species in this study. The number of such species-specific proteins ranged from 11 in *H. populiferum* to 347 in *S. ciliatum*. Among the 347 proteins in *S. ciliatum*, TMHMM v.2.0 (Sonnhammer et al., 1998) predicted that 54 proteins possess at least one transmembrane helix, suggesting that at least some of them could be mitochondrial membrane proteins with novel functions. Domain centric GO-analysis of the protein domains of the 347 proteins from *S. ciliatum* revealed threonine-tRNA ligase activity, uridylyltransferase activity, polynucleotide adenylyltransferase activity, and tRNA binding as enriched molecular function terms. Some of the MTS-only proteins can be false-positive results of TargetP and MitoFates prediction, especially those with a low RC in TargetP and a low probability in MitoFates. Our analysis of human proteome indicates that 33% of proteins predicted to be mitochondrial by the joint TargetP plus MitoFates approach are not present in MitoCarta and thus likely false positives (see MATERIALS AND METHODS). Interestingly, some studies have
proposed that dual-targeted proteins, that is, those targeted to mitochondria and another location, possess weaker MTS (Ben-Menachem et al., 2011; Dinur-Mills et al., 2008). Thus, experimental analysis can help verify the mitochondrial localization of these proteins, as well as elucidate their functions. The large number of MTS-only proteins in the calcarean sponges makes them good candidates for future mitochondrial proteomic analyses.

2.4.4 Conservation of proteins involved in core mitochondrial functions

In addition to the analysis of the whole mt-proteome described earlier, we searched orthologous groups detected by Proteinortho for the specific proteins involved in several important mitochondrial pathways: TCA, OXPHOS, FE/S cluster assembly (FES), apoptosis, mitochondrial biogenesis (MtBio), as well as proteins from the mitochondrial contact site and cristae organizing system (MI-COS) and the mitochondrial ribosome. There was a substantial variation in the number of detected proteins both among analyzed pathways and among analyzed species (Figure 2.4 and Supporting Information File S3). For some pathways (TCA and FES), nearly all involved proteins were identified in the mt-proteomes of most nonbilaterian species, as well as the two choanoflagellate outgroups. For other pathways and complexes, for example, OXPHOS complexes II, III, IV, and V, the detection of protein orthologs was patchier. The number of nonbilaterian orthologs to human proteins involved in apoptosis was especially low. The number of predicted orthologs to human subunits varied also across analyzed species, where it correlated with the size of the predicted mt-proteome. Thus, species lacking some of the highly conserved TCA cycle proteins—H. populiferum, Corticium candelabrum, Ircinia fasciulata, and P. bachei—had the fewest identified mitochondrial orthologs in general, suggesting that their transcriptomes used to infer mt-proteomes were incomplete.

2.4.5 Analysis of mitochondrial targeting signals of mt-proteins

Import of mt-proteins via the MTS-pathway depends on the presence of a short (15-50 amino acid) sequence at the N-terminus of a peptide, usually enriched in hydrophobic and positively charged amino acids. While the N-termini of mt-proteins from human, mouse, and yeast have been
characterized (Calvo et al., 2017), no such analysis exists for any nonbilaterian animals. Hence, we analyzed the amino-acid composition, length, and net charge of all predicted MTS (Supporting Information File S4). The average size of the MTS ranged from 27 amino acids in *M. leidyi* to 32 amino acids in *Polypodium hydri forme*. The average net charge of the predicted MTS was around +5. The PCA of MTSs amino-acid composition revealed three distinct clusters: (i) mammals, (ii) choanoflagellates, and (iii) nonbilaterian animals + yeast. (Figure 2.5).

As reported above, the number of mt-proteins identified by MTS predictors varied more than 10-fold across analyzed nonbilaterian species (Supporting Information S2). To evaluate the source of this variation, that is, to identify whether these results are because of shortcomings of the prediction algorithms or whether they reflect the genuine absence of MTS in mt-proteins, we aligned and analyzed predicted targeting sequences of TCA cycle proteins. TCA cycle proteins provide a good model to evaluate the impact the MTS predictors have in detection of mt-proteins in nonbilaterian species because (i) they are well conserved in animals, (ii) they are matrix proteins involved in a crucial mitochondrial function, and (iii) all human TCA proteins possess a MTS. Although we identified nearly all TCA proteins in the inferred mt-proteomes, the number of TCA proteins with a detectable MTS varied. Variation was seen both with respect to specific TCA proteins and individual species (Figure 2.6). For instance, succinate dehydrogenase cytochrome b560 subunit (C560) was predicted to possess a MTS in 18 of the 19 nonbilaterian species detected, whereas citrate synthase (CISY) was predicted to possess a MTS in just 13 of the 21 nonbilaterian species. In the nonbilaterian species, we detected MTS in 20 of the 22 proteins in *Hydra vulgaris* and in just 10 of the 22 proteins in *Trichoplax adhaerens*.

In the majority of cases, the negative results of MTS prediction were either because the protein did not possess an MTS or the protein was a fragment and not because of the problem with the MTS predictors. As an example, Figure 2.7 depicts an N-terminal portion of alignment of fumarate hydratase (FUMH) for human and species where MTS was not detected. In only one of these proteins, the N-terminus of the protein is likely incomplete, whereas the rest appear to genuinely lack the MTS.
Further analysis is necessary to identify mechanism of mitochondrial import of TCA proteins, which lack a MTS. As these perform important functions in the TCA cycle, it is most likely that they are imported into the mitochondrial matrix. Biochemical studies, as performed in mitochondrial heme lyases (Diekert et al., 1999) would be useful to identify alternative signals for mitochondrial import, such as internal targeting signals and C-terminus targeting sequences (Diekert et al., 1999; Becker et al., 2012; Lee et al., 1999; Schmidt et al., 2010).

2.4.6 Analysis of mitochondrial protein domains

Protein domains are evolutionarily conserved functional modules and their analysis provides an additional approach to investigate functional diversity of the inferred mt-proteomes. Thus, we identified the number of different domains in the inferred mt-proteomes and analyzed their conservation and losses in animals (Supporting Information File S5).

While nearly all the reference mt-proteins possessed a known protein domain, the percentage of proteins with known domains varied in nonbilaterian species, ranging from 59% in *P. bachei* to 95% in *Haliclona amboinensis* and *T. adhaerens*. The total number of distinct protein domains also varied across the nonbilaterian mt-proteomes, from 431 in *H. populiferum* to 1,155 in *L. complicata* (Supplementary Materials File 5). While the number of protein domains detected and the number of mt-proteins detected were correlated ($r^2 = 0.81$), there was no correlation between the percentage of proteins with detected domains and the size of the mt-proteome ($r^2 = 0.02$).

The mean number of protein domains in the inferred mt-proteomes (852) was smaller than the number of protein domains in the reference human mt-proteome (1,362). Only 563 protein domains were shared across all the nonbilaterian groups under study and human. A domain centric GO analysis of these conserved domains revealed that they were involved in critical mitochondrial processes such as pyruvate dehydrogenase activity, molecular carrier activity, oxidoreductase activity, acting on the aldehyde or oxo group of donors, disulfide as acceptor, ATP-dependent peptidase activity, and NAD binding among others.
The top most abundant protein domain (protein domain present in the highest number of proteins) in all species was the mitochondrial carrier protein domain (Mito_carr), found in the mitochondrial carrier family (MCF)\textsuperscript{18}, consisting of inner mitochondrial membrane proteins that transport various solutes such as ADP/ATP, H+, coenzyme A, citrate, and so forth across the inner mitochondrial membrane (Palmieri, 2013).

We identified 120 protein domains present in the human mt-proteome but not identified in any other proteome in this study. Functional analysis of these domains using dcGO showed mitochondrial respiratory chain complex I as the top most abundant GO Cellular Component category, consistent with the current knowledge that the NADH dehydrogenase complex increased in complexity via subunit addition in animals and possesses various human-specific subunits (Gabaldón et al., 2005a). Among the most enriched GO biological process terms associated with these 120 protein domains were mitochondrial respiratory chain complex assembly, regulation of immune response, regulation of inflammatory response, cristae formation, and mitochondrial organization.

Forty nine protein domains were identified in mt-proteomes of yeast, choanoflagellates, and all lineages of nonbilaterian animals but not in human. dcGO analysis of these domains showed DNA-directed 5-3 RNA polymerase activity as the top most enriched GO Slim term. In addition, these protein domains included PAM17 (presequence translocase-associated motor [PAM] subunit 17), MitMem_reg (maintenance of mitochondrial structure and function), SMC_N (structural maintenance of chromosomes) among others.

Of the 2,274 protein domains identified in animals, 666 were present in at least one nonbilaterian mt-proteomes but not in human, yeast, or choanoflagellate mt-proteomes. Enriched GO biological process terms associated with these nonbilaterian specific protein domains included reproductive process, multicellular organism development, nervous system development, and localization. Of these 666 protein domains, MRP-L27 (Mitochondrial ribosomal protein L27) (11 species), Histone (10 species), and PAW (PNGase C-terminal domain, mannose-binding module PAW) (9 species) were present in most nonbilaterian species. Although some of these domains could be indicative of

\textsuperscript{18}MCF: mitochondrial carrier family
new functions in nonbilaterian mitochondria, many are likely to be derived from the proteins that have misidentified as mitochondrial (false positives) in the MTS analysis.

2.5 Conclusion

This study attempts the first characterization of mt-proteins in nonbilaterian animals, which represent most of the major lineages in the animal phylogenetic tree. Inclusion of these animals in the comparative analysis allowed us to obtain a broader picture on the conservation and evolution of mt-proteome in animals. We found that although the number of predicted mt-proteins varied widely across the species of nonbilaterian animals— from 454 in the myxozoan *K. iwatai* to 2,119 in the calcarean sponge *L. complicata*— 513 mitochondrial OGs were well conserved across all five major branches of animal phylogenetic tree.

We found that most of the variation in the number of predicted mt-proteins was due to the MTS prediction approach. Interestingly, most of it appears to be because of biological (e.g., alternative targeting signals) rather than technical (failure to identify an MTS) reasons. In fact, analysis of identified MTS has shown that they are similar in length and charge across nonbilaterian species and that their amino composition is more similar to that in yeast rather than mammals and choanoflagellates.

About 42 human mt-proteins (∼2.5% of human mt-proteome) were not found in any nonbilaterian animals, suggesting that they have evolved within Bilateria. We have also identified several proteins that appear to be targeted to mitochondria in multiple nonbilaterian species but either are not known to have mitochondrial localization in humans, like histones and cytosolic ribosomal proteins, or not present in human mitochondria, like alternative oxidases (AOX). Finally, analysis of mt-protein domains has shown that 563 protein domains were shared across all the studied animal groups with the mitochondrial carrier protein domain (Mito_carr) being the most common in animal mt-proteins.
When interpreting results of this study, it is important to keep in mind several limitations of such in-silico approaches. In particular, there are several reasons why a mt-protein may not be found or recognized.

- The completeness and quality of the transcriptome would affect the number of mt-proteins and mt-protein domains detected.

- The protein may have evolved beyond recognition by homology searches, which could be possible in the fast evolving nonbilaterian genomes like those of the myxozoans.

- Due to the minimum size of the predicted proteins in our study (100 amino acids), proteins below 100 amino acids in length would not be identified.

Experimental identification of mt-proteins from the nonbilaterian phyla can be the next step in understanding the evolution of mt-proteomes in animals. An approach similar to the one performed in the MitoCarta analysis of mammalian mitochondria can be utilized, where several methods such as MS/MS and co-expression analysis were used, in addition to in silico techniques such as protein domain analysis, presequence detection, and homology (Calvo et al., 2015). This would be particularly useful to identify novel mt-proteins in nonbilaterian animals. Future research can also be focused on experimental characterization of presequences in nonbilaterian animals, as was performed in mammals and yeast (Calvo et al., 2017), to help study MTS evolution.

2.6 Acknowledgement

We gratefully acknowledge Dr. Carolyn Lawrence-Dill, Dr. Karin Dorman, Dr. Iddo Friedberg and Dr. Robert Jernigan for their comments and suggestions regarding the manuscript.

2.7 References


Figure 2.1: **Animal phylogeny with focus on nonbilaterian phyla: Cnidaria, Ctenophora, Placozoa, and Porifera.** Phylum Cnidaria has been traditionally classified into two groups: Anthozoa (sea anemones, gorgonians, soft and stony corals) and Medusozoa (hydroids, jellyfish, syphonophores), but now includes Myxozoa, obligate microscopic parasites, characterized by an extremely simplified morphology and massive genome reduction as well as Polypodiozoa, represented by a single parasite of fish. Sponges are subdivided into four groups (classes) based on their cellular organization, chemical composition of produced spicules as well as molecular phylogenetic analyses. Nonbilaterian lineages are shown within the gray box. Multiple hypotheses have been proposed for the phylogenetic relationships between sponges, cnidarians, ctenophores, placozoans, and bilaterians. The Coelenterata hypothesis has been depicted. This hypothesis states that Porifera is the sister group to all other animals, and Ctenophora and Cnidaria form Coelenterata. The selection of a particular hypothesis does not influence the results of this study.
Figure 2.2: Flow chart of the protocol used to infer mitochondrial proteomes in non-bilaterian animals. Details of each step are explained in the section Predicting Mitochondrial Proteomes.
Figure 2.3: The size of inferred mitochondrial proteomes in nonbilaterian animals and choanoflagellates. We have species on the x-axis and number of proteins on the y-axis. Different colors indicate proteins predicted by Proteinortho alone (purple), by the MTS-predictors (TargetP and MitoFates) alone (green), and by both methods (white). Species abbreviations are explained in Table 1. Also included in the image are the two reference mitochondrial proteomes: human and yeast (colored in black).
Figure 2.4: Proportion of human proteins with detected homologs in nonbilaterian mitochondrial proteomes. The following pathways/complexes are shown: Tricarboxylic Acid Cycle (TCA), FE/S cluster assembly (FES), oxidative phosphorylation complexes I-V (CI, CII, CIII, CIV, CV) and apoptosis. The median is represented by the bar (-) and the mean is represented with the (+).
Figure 2.5: **PCA of amino-acid composition of the presequences identified in the mitochondrial proteomes.** Prediction ellipses are such that with probability 0.95, a new observation from the same group will fall inside the ellipse.
Figure 2.6: **Analysis of proteins involved in TCA.** The upper matrix shows the presence/absence of the TCA cycle proteins in each analyzed species. Dark green: multiple co-orthologs detected by Proteinortho, Green: one ortholog detected and white: no ortholog detected. The lower matrix shows the presence/absence of predicted MTS in the TCA cycle proteins. Purple: MTS was predicted (in at least one protein, if multiple orthologs were detected) by at least one predictor, orange: no MTS was identified by either of the two predictors and white: no ortholog was detected by Proteinortho.
Figure 2.7: Alignment of fumarate hydratase (FUMH) sequences with no predicted MTS and that of human FUMH. Only the N-terminal end of the protein alignment is depicted in the image. For the human FUMH sequence (HS), the presequence predicted by MitoFates is underlined with a purple line.
Table 2.1: Species of nonbilaterian animals and choanoflagellates used in this study along with the name abbreviations used.

<table>
<thead>
<tr>
<th>Phylum</th>
<th>Species</th>
<th>Abbreviation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Porifera</td>
<td>Calcarea</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Leucosolenia complicata</td>
<td>LC</td>
<td>(Fortunato et al., 2014)</td>
</tr>
<tr>
<td></td>
<td>Sycon ciliatum</td>
<td>SC</td>
<td>(Fortunato et al., 2014)</td>
</tr>
<tr>
<td>Hexactinellida</td>
<td>Aphrocallistes vastus</td>
<td>AV</td>
<td>(Riego et al., 2014)</td>
</tr>
<tr>
<td></td>
<td>Hyalonema populaferum</td>
<td>HP</td>
<td>(Whelan et al., 2015)</td>
</tr>
<tr>
<td>Homoscleromorpha</td>
<td>Corticium candelabrum</td>
<td>CC</td>
<td>(Riego et al., 2014)</td>
</tr>
<tr>
<td></td>
<td>Oscarella pearsei</td>
<td>OP</td>
<td>(Ereskovsky et al., 2017)</td>
</tr>
<tr>
<td>Demospongiae</td>
<td>Haliclonia amboinensis</td>
<td>HA</td>
<td>(Guzman and Conaco, 2016)</td>
</tr>
<tr>
<td></td>
<td>Ephydatia muelleri</td>
<td>EM</td>
<td>(Peña et al., 2016)</td>
</tr>
<tr>
<td></td>
<td>Chondrilla nucula</td>
<td>CN</td>
<td>(Riego et al., 2014)</td>
</tr>
<tr>
<td></td>
<td>Ircinia fasciculata</td>
<td>IF</td>
<td>(Riego et al., 2014)</td>
</tr>
<tr>
<td></td>
<td>Tethya wilhelma</td>
<td>TW</td>
<td>(Francis et al., 2017)</td>
</tr>
<tr>
<td></td>
<td>Amphimedon queenslandica</td>
<td>AQ</td>
<td>(Fernandez-Valverde et al., 2015)</td>
</tr>
<tr>
<td></td>
<td>Halisarca dayardini</td>
<td>HD</td>
<td>(Fernandez-Valverde et al., 2015)</td>
</tr>
<tr>
<td>Cnidaria</td>
<td>Anthozoa</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nematostella vectensis</td>
<td>NV</td>
<td>(Putnam et al., 2007)</td>
</tr>
<tr>
<td></td>
<td>Acropora digitifera</td>
<td>AD</td>
<td>(Shinizato et al., 2011)</td>
</tr>
<tr>
<td>Medusozoa</td>
<td>Aurelia aurita</td>
<td>AA</td>
<td>(Fuchs et al., 2014)</td>
</tr>
<tr>
<td></td>
<td>Clytia hemisphaerica</td>
<td>CH</td>
<td>(Lapèbie et al., 2014)</td>
</tr>
<tr>
<td></td>
<td>Hydra vulgaris</td>
<td>HV</td>
<td>(Chapman et al., 2010)</td>
</tr>
<tr>
<td>Myxosporea</td>
<td>Kudoa irai</td>
<td>KI</td>
<td>(Chang et al., 2015)</td>
</tr>
<tr>
<td>Polypodiozoa</td>
<td>Polypodium hydriforme</td>
<td>PH</td>
<td>(Chang et al., 2015)</td>
</tr>
<tr>
<td>Ctenophora</td>
<td>Tentaculata</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mnemiopsis leidyi</td>
<td>ML</td>
<td>(Ryan et al., 2013)</td>
</tr>
<tr>
<td></td>
<td>Pleurobrachia backei</td>
<td>PB</td>
<td>(Moroz et al., 2014)</td>
</tr>
<tr>
<td>Placozoa</td>
<td>Trichoplax adhaerens</td>
<td>TA</td>
<td>(Srivastava et al., 2008)</td>
</tr>
<tr>
<td>Choanoflagellata</td>
<td>Monosiga brevicollis</td>
<td>MB</td>
<td>(King et al., 2008)</td>
</tr>
<tr>
<td></td>
<td>Salpingoeca rosetta</td>
<td>SO</td>
<td>(Fairclough et al., 2013)</td>
</tr>
</tbody>
</table>
CHAPTER 3. CAUSES AND CONSEQUENCES OF MITOCHONDRIAL PROTEOME SIZE-VARIATION IN ANIMALS

Modified from a manuscript under review in Mitochondrion

Viraj Muthye\(^1,2\) and Dennis Lavrov\(^1,2\)

\(^1\) Bioinformatics and Computational Biology Program, Iowa State University, 2437 Pammel Drive, Ames, Iowa 50011, USA
\(^2\) Department of Ecology, Evolution and Organismal Biology, Iowa State University, 241 Bessey Hall, Ames, Iowa 50011, USA

3.1 Abstract

Despite a conserved set of core mitochondrial functions, animal mitochondrial proteomes show a large variation in size. In this study, we analyzed the putative mechanisms behind and functional significance of this variation using experimentally-verified mt-proteomes of four bilaterian animals and two non-animal outgroups. We found that, of several factors affecting mitochondrial proteome size, evolution of novel mitochondrial proteins in mammals and loss of ancestral proteins in protostomes were the main contributors. Interestingly, gain and loss of conventional mitochondrial targeting signals was not a significant factor in the proteome size evolution.

3.2 Introduction

Mitochondria, membrane-bound organelles present in most eukaryotic organisms, are involved in a number of cellular processes, including oxidative phosphorylation, Fe/S cluster biosynthesis, amino-acid and lipid metabolism, and apoptosis (Szklarczyk and Huynen, 2010; Meisinger et al., 2008). These tasks typically require more than a thousand proteins, the vast majority of which are
encoded in the nuclear genome and imported into organelle (Becker et al., 2012; Chacinska et al., 2009; Gabaldón and Huynen, 2004). Thus, analysis of nuclear-encoded mitochondrial proteins (mt-proteins\(^1\)) is essential for understanding mitochondrial function and evolution.

Several approaches have been used to estimate the composition of mitochondrial proteomes (mt-proteomes\(^2\)), which can be roughly divided into those directly extracting proteins from mitochondria and those that infer their mitochondrial localization using a variety of analytical methods (Calvo and Mootha, 2010). The former methods include mass spectrometry (MS) and microscopy. The latter include identification of targeting signals, homology search, co-expression analysis, and phylogenetic profiling. Since all these techniques have their pros and cons (Calvo and Mootha, 2010), integrated approaches have been advocated and applied to elucidate some mt-proteomes, as in the case of human and mouse MitoCarta (Calvo et al., 2016). Despite these advances, the number of experimentally determined mt-proteomes remains small.

Mt-proteomes have been characterized in several eukaryotic species, including plants (Rao et al., 2017; Salvato et al., 2014; Lee et al., 2013; Huang et al., 2009), fungi (Cherry et al., 2012), animals (Calvo et al., 2016; Li et al., 2009a; Hu et al., 2015; White et al., 2011) and protists (Zhang et al., 2010; Gawryluk et al., 2014a; Seidi et al., 2018; Panigrahi et al., 2009). Within animals, mt-proteomes have been experimentally determined in a few model species: mammals (human, mouse and rat (Calvo et al., 2016; Smith and Robinson, 2016)), arthropod (Drosophila melanogaster) (Hu et al., 2015) and nematode (Caenorhabditis elegans) (Li et al., 2009a). These studies revealed a large size difference between mammalian mt-proteomes and those of protostome animals (nematodes and arthropods). Indeed, inferred mt-proteomes of D. melanogaster and C. elegans are about half and 65% in size, respectively, compared to those of human and mouse (Hu et al., 2015; Calvo et al., 2016; Smith and Robinson, 2016). Furthermore, our recent bioinformatic analysis has suggested an even larger variation in the sizes of non-bilaterian mt-proteomes (Muthye and Lavrov, 2018).

---

\(^1\)mt-proteins : mitochondrial proteins

\(^2\)mt-proteomes : mitochondrial proteomes
Given the large size difference in both experimentally determined and computationally predicted mt-proteomes, we asked three questions in the present study:

1. What is the mechanism behind the observed variation in mt-proteome size?

2. What is the functional significance of this variation?

3. How much of this variation can be attributed to gain/loss of Mitochondrial Targeting Signal (MTS<sup>3</sup>)?

The latter question concerns the predominant method of protein import into mitochondria - via the MTS/Presequence pathway. MTS refers to a targeting signal at the N-terminus end of the protein, usually within the first 90 amino-acid residues. The MTS are enriched in arginine and depleted in negatively charged residues, and form an amphipathic α-helical structure (Fukasawa et al., 2015; Emanuelsson et al., 2007) Proteins which harbor these MTS are transported into the organelle by TOM and TIM complexes in the mitochondrial outer and inner membranes, respectively. Since a large portion of the mt-proteins possess MTS, in silico identification of MTS is an important technique to predict protein localization.

To answer the above questions about the variation in mt-proteome composition, we analyzed the experimentally determined mt-proteomes of four bilaterian animals (Homo sapiens (Calvo et al., 2016), Mus musculus (Calvo et al., 2016), nematode Caenorhabditis elegans (Li et al., 2009a) and fruit-fly Drosophila melanogaster (Hu et al., 2015)) and two outgroups (Acanthamoeba castellanii (Gawryluk et al., 2014a) and yeast Saccharomyces cerevisiae (Cherry et al., 2012)). We hypothesized three distinct mechanisms that could contribute to the variation in the size and content of animal mt-proteomes:

<sup>3</sup>MTS : mitochondrial targeting signal
1. duplication, and/or loss of ancestral mt-proteins

2. re-localization of ancestrally non-mt-proteins to mitochondria

3. evolution of novel mt-proteins

To understand the contribution of each of these factors to the evolution of animal mt-proteomes, we subdivided all mt-proteins into four categories - ancestral proteins, proteins which underwent mitochondrial neolocalization, novel animal proteins and species-specific proteins - and analyzed the contribution of each of these categories to the mt-proteome size variation. In addition, we investigated the functional implications of this observed variation in mt-proteome size. Finally, for each of the four categories of mt-proteins, we identified the proportion of proteins possessing a MTS. Our results showed that the size-increase in mammalian mt-proteomes is primarily due to the evolution of novel animal mt-proteins in mammals and loss of ancestral mt-proteins in protostomes. In addition, some of the size difference appears to be an artifact of better mt-proteome annotation in deuterostomes. Interestingly, we found that the majority of neolocalized and novel animal mt-proteins did not possess a detectable MTS, suggesting that other import pathways play an important role in protein import of animal mt-proteins.

3.3 Materials and Methods

3.3.1 Assembling animal mt-proteomes

Four experimentally determined animal mt-proteomes were used in this study: those from *Homo sapiens*, *Mus musculus*, *Caenorhabditis elegans* and *Drosophila melanogaster*. The rat mt-proteome was not used in this analysis to avoid an over-representation of mammals. Human and mouse mt-proteomes were assembled using data from MitoCarta v2.0 (Calvo et al., 2016) and IMPI (Integrated Mitochondrial Protein Index) vQ2 2018 (Smith and Robinson, 2016). The *C. elegans* mt-proteome was downloaded using data from Jing et al. (Li et al., 2009a). The *D. melanogaster* mt-proteome was downloaded using data from the iGLAD database (Hu et al., 2015).
In addition, two non-metazoan outgroups were used: *Acanthamoeba castellani* and *Saccharomyces cerevisiae*. The *A. castellani* mt-proteome was downloaded from Harvard Dataverse (Gawryluk et al., 2014a,b) and the yeast mt-protein was downloaded from the Saccharomyces Genome Database (https://www.yeastgenome.org/) (Cherry et al., 2012).

The complete proteomes for three animal species: *Homo sapiens*, *Mus musculus*, and *Drosophila melanogaster* and the protist *Acanthamoeba castellanii* were downloaded from Uniprot (TheUniProt-Consortium, 2016). The complete yeast proteome was downloaded from the Saccharomyces Genome Database (Cherry et al., 2012). The complete proteome for *Caenorhabditis elegans* was downloaded from WormBase (Harris et al., 2009).

### 3.3.2 Identification of orthologous groups

Proteinortho v5.16b (Lechner et al., 2011) with the default parameters (e-value:1e-05, similarity:0.25, percentage-identity:0.25, purity:1, connectivity:0.1) was used to identify groups of orthologous proteins (OGs<sup>4</sup>) between complete proteomes of six eukaryotic species. From the resulting sets of OGs, OGs containing mt-proteins were extracted. The OGs and the proteins contained within the OGs were subdivided into three categories, based on the presence/absence and sub-cellular localization (mitochondrial/ non-mitochondrial) of outgroup mt-proteins within individual OGs. This is described in more detail in section 3.4.1.1.

### 3.3.3 Identification of Mitochondrial Targeting Signals (MTS)

MTS were identified using MitoFates (Fukasawa et al., 2015) (metazoan option) and TargetP v1.1 (Emanuelsson et al., 2007) (Non-Plant option). TargetP outputs a statistic (Reliability Class (RC)<sup>5</sup>) for each prediction, which is an indication of the strength of that prediction, with RC1 indicating the strongest and RC5 indicating the weakest prediction. MitoFates, on the other hand, provides a probability for each prediction. In this study, a protein was considered to possess a MTS if either TargetP (RC1-5) or MitoFates (any prediction probability) identified a MTS.

---

<sup>4</sup>OG: Orthologous Group

<sup>5</sup>RC: Reliability Class
3.3.4 Functional analysis of mt-proteins

We used Gene Ontology (GO) analysis, pathway analysis and protein-protein interaction network analysis to assign biological function to mt-proteins. PantherDB v14.0 (Mi et al., 2016) was used for GO analysis of mt-proteins from all four animal species, while WormBase (Chen et al., 2005) and FlyMine (Lyne et al., 2007) were used for analysis of *C. elegans* and *D. melanogaster* mt-proteins respectively. PantherDB was used to perform the Panther Over-representation test. For a list of given genes, the over-representation test identifies GO terms which are over-represented/under-represented when compared to a list of reference genes. In addition to GO analysis, tissue-enrichment and phenotype-enrichment analysis for genes from *C. elegans* was carried out using WormBase (Angeles-Albores et al., 2018, 2016).

Functional annotation of proteins is greatly aided by knowledge of their interaction-partners: protein-protein interaction networks (PPI). We used two tools, which utilize PPI information to identify biological processes and pathways enriched in different sets of mt-proteins: ConsensusPathDB Release 34 (Kamburov et al., 2010) and StringDB (Szklarczyk et al., 2016). Default parameters were used for ConsensusPathDB (minimum overlap with input list:2 and p-value:0.01). ConsensusPathDB was used for functional annotation of mammalian mt-proteins, since it houses data for human, mouse and yeast. StringDB incorporates PPI data from multiple sources and identifies functional associations between query proteins. This tool was used to identify functionally-associated clusters of species-specific mt-proteins, particularly from *C. elegans* and *D. melanogaster*.

3.3.5 Data availability

Results of these analyses and scripts used are available on Open Science Framework (Muthye, V., & Lavrov, D. (2019, July 31). Data for Causes and Consequences of animal mt-proteome size variation. https://doi.org/10.17605/OSF.IO/A49YW)
3.4 Results

3.4.1 Evolution of animal mt-proteomes

3.4.1.1 Four categories of animal mt-proteins

We used Proteinortho v5.16b to identify groups of orthologous proteins (OGs) in four animals (Homo sapiens, Mus musculus, Caenorhabditis elegans, Drosophila melanogaster) and two outgroups (Acanthamoeba castellani and Saccharomyces cerevisiae). 1909 OGs had a mt-protein from at least one animal species. These were divided into three categories (Figure 3.1, Figure 3.2):

- Category I: OGs containing at least one mt-protein from a given animal and a mt-protein from at least one outgroup. mt-proteins within these OGs were catalogued as ancestral mt-proteins.

- Category II: OGs containing at least one mt-protein from a given animal and a non-mt-protein from at least one outgroup, but no mt-protein from either outgroups. mt-proteins within these OGs were catalogued as neolocalized mt-proteins.

- Category III: OGs containing at least one mt-protein from a given animal and no protein (mitochondrial or non-mitochondrial) from either outgroups. mt-proteins within these OGs were catalogued as novel animal mt-proteins.

Some animal mt-proteins were not a part of any OG, i.e. lacking identifiable orthologs in all other species used in the study. These were catalogued as category IV Species-specific mt-proteins. Below we describe the composition of each of the four categories of animal mt-proteins.

3.4.1.2 Category I: Ancestral mt-proteins

In total, 528 OGs containing ancestral mt-proteins were identified. Nearly half of them (205/528) had at least one ancestral mt-protein from all four bilaterian species (Figure 3.3A). In addition, 239 OGs in this category were comprised of mt-proteins from both mammals but lacked a mitochondrial
ortholog from either one (149) or both protostomes (90). Interestingly, the majority of OGs, lacking just one protostome mt-protein (52/74 in C. elegans and 60/75 in D. melanogaster) possessed a non-mitochondrial protein from that species, suggesting possible misannotation. Furthermore, 42 out of 90 OGs with an ancestral mt-protein from both mammals contained non-mitochondrial orthologs in both protostomes. We identified 45 and 68 proteins from C. elegans and D. melanogaster, which were present in Category I OGs and possessed a MTS, but were not annotated as mt-proteins. At the same time, 31 out of 90 OGs with an ancestral mt-protein from both mammals had no protostome protein (either mitochondrial or non-mitochondrial), indicating a loss in protostomes.

3.4.1.3 Category II: Neolocalized mt-proteins

We identified 462 OGs with proteins belonging to category II (Figure 3.3B). 36 OGs contained category II proteins from all four bilaterian animals, suggesting mitochondrial neocalization prior to the protostome-deuterostome split. Additionally 66 OGs possessed mt-proteins from both mammals and one of the two protostome animals. 29/42 and 16/24 of these OGs contained an orthologous protein identified as non-mitochondrial in the other protostome species, either D. melanogaster or C. elegans, respectively. These may represent further examples of yet-to-be-identified mt-proteins. We identified 164 OGs with a category II protein from both mammals, but no mt-protein from either protostome. 123 of these 164 OGs comprised of category II proteins from both mammals and only non-mitochondrial proteins from either one or both protostomes suggesting neocalization in mammals. The remaining 41/164 OGs contained a category II protein from both mammals and no ortholog from either of the protostomes, implying both loss and neocalization in their evolution. Finally, in 23 category II OGs, there were mt-proteins from both protostomes, and a non-mt-protein from both mammals, suggesting a neocalization in protostomes.

3.4.1.4 Category III: Novel animal mt-proteins

In total, novel animal mt-proteins were identified in 919 OGs, of which only 86 (around 9%) had category III proteins from all four animals (Figure 3.3C). More than half of the total category
III OGs (494/919) contained either a novel mitochondrial protein from both mammals but no orthologs from either of the protostomes (342 OGs) or a novel mt-protein from both mammals and a non-mitochondrial protein from either one or both protostomes (152 OGs). By contrast, only 21 OGs possessed either a novel mt-protein from both protostomes but no mammalian orthologs (10 OGs) or a novel mt-protein from both protostomes and a non-mitochondrial protein from both mammals (11 OGs).

### 3.4.1.5 Category IV: Species-specific mt-proteins

13, 30, 124 and 275 category IV proteins (proteins with no identifiable orthologs) were detected in mouse, human, *D. melanogaster* and *C. elegans* respectively (Figure 3.3D). The large difference in the number of species-specific proteins between protostome vs. mammalian species can be explained by the difference in their estimated divergence times: ~555 MYA (protostomes) (Parfrey et al., 2011) vs. ~88 MYA (mammals) (Kumar et al., 2017).

### 3.4.1.6 Contribution of the four categories towards the size-increase of mammalian mt-proteomes

The mammalian mt-proteomes were, on average, 645 proteins larger than the protostome mt-proteomes. We quantified the extent to which each category of proteins contributed to this difference by calculating its contribution factor (CF)\(^6\). For each category, the contribution factor was defined as

\[
CF = \frac{\text{average number of mammal proteins} - \text{average number of protostome proteins}}{645}
\]

A contribution factor of 1 would indicate that the size-increase of mammalian mt-proteomes can be explained completely by that category, while 0 would indicate that category has no effect. A negative value would indicate a higher number of protostome proteins in that category compared

---

\(^6\text{CF: Contribution Factor}\)
to mammals. Since category IV is a special case of category III, we combined the contribution factors for both categories into a single score. The contribution factors were 0.17, 0.27 and 0.56 for category II, I and (III+IV) respectively.

In addition to neolocalization and evolution of novel mt-proteins, we also hypothesize that duplication of genes encoding mt-proteins could also result in an increase in size of mammalian mt-proteomes. To estimate the contribution of gene-duplication towards mitochondrial proteome size-variation, we calculated the contribution factor per category as follows:

$$\text{diff} = \text{number of proteins} - \text{number of OGs}$$

$$\text{CF} = \frac{\text{average mammal diff} - \text{average protostome diff}}{645}$$

The contribution factors for gene-duplication in the three categories I-III were low (0.04 (category I), -0.02 (category II) and 0.02 (category III) respectively).

### 3.4.2 Functional analysis of mt-proteins

#### 3.4.2.1 Expansion of pathways in mammalian mitochondria

On average, mammals had 538 more novel animal mt-proteins (category III) compared to protostomes. In mammals, majority of these novel animal mt-proteins were present in both human and mouse (494 OGs). Several proteins from these 494 OGs were additions to the mitochondrial inner (44) and outer (54) membranes, locations of critical mitochondrial functions such as energy metabolism and metabolite and protein import. Panther Over-representation test showed that these mammal-specific novel animal mt-proteins were involved in mitochondrial translation, apoptosis, complex I assembly and oxidative phosphorylation. “Thermogenesis”, “Butanoate metabolism”, “Apoptosis” and “Oxidative phosphorylation” were the top enriched pathways detected by ConsensusPathDB.

Novel mammalian mt-proteins could represent either novel biological pathways or additions to existing pathways. KEGG analysis revealed that 86% (231/268) of pathways identified in at least one animal species were present in all four animals. Thus, most of the mammal-specific mt-proteins
were additions to existing animal mitochondrial pathways. In nearly half (118/268) of the identified pathways, the average number of mammalian proteins was at least double the number of invertebrate proteins. Majority of them were signaling pathways, like “JAK-STAT signaling pathway”, “Prolactin signaling pathway” and “ErbB signaling pathway”.

3.4.2.2 Loss of proteins involved in amino-acid metabolism in protostomes

528 OGs included ancestral mt-proteins, of which 31 OGs contained ancestral mt-proteins from both mammals, and no orthologs in either protostome, thus representing loss in protostomes. 24 proteins from these 31 OGs were mapped to at least 1 pathway by ConsensuspathDB including metabolism of amino acids and derivatives (8) and urea cycle (3). Other ancestral proteins lost in protostomes included Monocarboxylate transporter 1 (SLC16A1), Magnesium transporter (MRS2) and Protein adenylyltransferase SelO (SELENOO).

3.4.2.3 Protostome-specific mt-proteins

We identified 507 protostome-specific mt-proteins: novel animal proteins (category III) and species-specific proteins (category IV) in the two analyzed protostome mt-proteomes. Only 21 of them were present in both species. 322 C. elegans proteins did not have any mitochondrial orthologs in other species and 275 of them had neither mitochondrial nor nuclear orthologs (category IV). Similarly, 164 D. melanogaster proteins did not have mitochondrial orthologs in other species, of which 124 lacked both mitochondrial and nuclear orthologs (category IV).

Interestingly, many of these 275 category IV proteins from C. elegans were involved in nematode-specific functions, like nematode larval development, embryo development and protection against microbes. Phenotype-enrichment analysis identified “innate immune response variant” and “avoids bacterial lawn” as the top enriched terms. StringDB identified 55 clusters of mt-proteins with at least 2 interacting members. The largest cluster contained 16 proteins involved in “lipid metabolism” and “nutrient reservoir activity”. The next largest cluster of 9 proteins were possible sperm mt-proteins. The remaining three clusters comprised of DNA-binding proteins, proteins
involved in regulation of transcription, mitochondrial membrane proteins and proteins involved in ATP hydrolysis coupled proton transport.

In *D. melanogaster*, 124 mt-proteins did not have an ortholog in any of the other species in this study. As per PantherDB, the top enriched protein classes of these *D. melanogaster*-specific mt-proteins were “oxidoreductases” (21) and “proteases” (14). FlyMine identified pathways for 41/124 *D. melanogaster* proteins. These proteins were involved primarily in metabolic pathways, with 11 involved in oxidative phosphorylation and 9 in lipid metabolism. StringDB identified 14 clusters of mt-proteins with at least 2 interacting members. The largest cluster possessed 10 proteins involved in sperm mitochondrial processes. These include Sperm-Leucylaminopeptidase 3, Sperm-Leucylaminopeptidase 5, Sperm-Leucylaminopeptidase 8 and Loopin-1.

### 3.4.3 Role of MTS in mt-proteome evolution

The proportion of mt-proteins with a MTS ranged from 41% to 56% among the four animal species. However, because of the difference in the mitochondrial proteome size, the absolute number of proteins with MTS was nearly twice as large in *H. sapiens* compared to *C. elegans* and *D. melanogaster*. We noticed a significant difference in the proportion of proteins possessing a MTS among the four categories of proteins described above (Figure 3.4). The average proportion of proteins with MTS ranged from 30% in category II to 66% in category I. For category I, III and IV, the proportion of proteins with MTS was lowest in *C. elegans* and highest in *D. melanogaster* (Table 3.2).

### 3.5 Discussion

One might expect that the centrality of mitochondria in multiple crucial cellular processes in animals would result in a large core of conserved animal mt-proteins with few lineage-specific additions/deletions. This expectation is reinforced by the near-identical mtDNA gene-content of the four animals used in this study (Anderson et al., 1981; Okimoto et al., 1992; Bibb et al., 1981; Garresse, 1988). By contrast, our comparative analysis of the four animal mt-proteomes paints a more
dynamic picture of the organelar proteome - a small core of ancestral conserved mt-proteins and a large number of lineage-specific differences (mammal-specific, nematode-specific or arthropod-specific) (Figure 3.2).

On average, while just 19% of each animal mt-proteome was conserved in all four animals, ∼ 42% did not have mitochondrial orthologs in other species in this study, i.e. lineage-specific mt-proteins. There are two possible interpretations for these results: 1] observed differences in mt-proteome reflect morphological and ecological differences among the animal species used for the study, or 2] observed differences can be explained by our inability to detect orthologous proteins in faster evolving protostome genomes. While both of these interpretations can be valid to some extent, functional analysis of observed differences favour the former one. For example, a large number of mammal-specific mt-proteins were involved in signaling pathways, a finding consistent with increased morphological and cellular complexity of mammals. The importance of mitochondria as signaling organelle in mammals is also well-supported by recent studies (Chandel, 2015; Tait and Green, 2012).

By contrast, in the nematode, interactions with its surrounding microbiome have influenced some of changes in the mt-proteome, with several species-specific mt-proteins involved in protection of the nematode from bacteria and fungi. *C. elegans* is a free-living nematode which grows on decomposing materials containing a high number of micro-organisms (Schulenburg and Félix, 2017). These microbes play a critical role in the life of the nematode, as food, competitors or even as pathogens and parasites. Interestingly, mitochondria themselves may be targeted by pathogenic microbes, and several proteins have been identified which up-regulate protective pathways following such mitochondrial disruption (Liu et al., 2014).

In addition to lineage-specific gains, another factor contributing to the variation in size and content of animal mt-proteomes was the loss of ancestral mt-proteins in protostomes. Indeed, ecdysozoans are known to have undergone extensive gene-loss (Albalat and Cañestro, 2016). For instance, in both protostomes, enzymes involved in amino-acid metabolism, such as those involved in arginine biosynthesis were lost, suggesting that these enzymes were dispensable in *C. elegans*
and *D. melanogaster*. These enzymes have been lost independently in multiple eukaryotic lineages (Payne and Loomis, 2006). The dual-role of these enzymes in the urea-cycle in vertebrates seems to have been responsible for the conservation of the arginine biosynthesis pathway in vertebrates.

Our analysis showed that most of the novel genes encoding mt-proteins are lineage-specific (mammal-specific, *C. elegans*-specific or *D. melanogaster*-specific), whereas comparatively few genes evolved before the protostome-deuterostome split. This result contrasts previous studies on the comparative analysis of animal genomes, which showed that most of the genomic novelty originated in the last common metazoan ancestor, and at the origin of bilaterian animals with few lineage-specific changes (Paps and Holland, 2018; Simakov and Kawashima, 2017). This discrepancy can be due to the limited taxonomic sampling of our study. If some novel mt-proteins were acquired either in the common ancestor of all metazoans or in the lineage leading to bilaterian animals and then lost in Ecdysozoa, the pattern of presence/absence of such proteins would be indistinguishable from the gain of these proteins in deuterostomes. The other major protostome group, lophotrochozoans, did not experience a gene-loss as severe as the ecdysozoans (Zmasek and Godzik, 2011). Characterization of mt-proteomes from this group can help differentiate whether a mt-protein was gained in mammals or lost in protostomes. In deuterostomes, both mt-proteomes belong to a single class (Mammalia) from a single phylum (Chordata). Also, experimental characterization of mt-proteomes from more animal lineages, such as lophotrochozoans and non-bilaterian animals, would aid in analysis of the metazoan mt-proteome. Additionally, it is also possible that some of the variation among mt-proteomes is due to the technical problems with orthology detection. While generating OGs, Proteinortho might miss some of the faster-evolving orthologs in *C. elegans* or *D. melanogaster*. This might result in overestimation of the number of lineage-specific mt-proteins.

Nearly, 1/5th of the animal mt-proteins are mitochondrial neolocalized proteins. Interestingly, majority of the mitochondrial recruitments did not happen via gaining a MTS. MTS play an important role in import of ancestral mt-proteins, but majority of mt-proteins from other classes
lacked MTS. This suggests that 1) alternate targeting pathways or non-canonical MTS facilitate import of a sizeable portion of animal mt-proteins and 2) an MTS-only approach to identify novel mt-proteins would miss a large number of proteins in this category. To date, several alternate pathways which import proteins in the mitochondria have been identified. For example, all of the mitochondrial outer-membrane proteins and majority of the intermembrane-space proteins contain non-canonical signals within the mature protein (Chacinska et al., 2009). In some proteins, like DNA helicase Hmi1p in yeast, the targeting signal is present at the C-terminal end of the protein instead of the canonical N-terminal end (Lee et al., 1999). Identification of such alternate targeting signals and development of tools to identify these signals in proteins can further our understanding of mt-protein import.

Some additional factors need to be considered while interpreting results of this study, like differences in methods used in characterization of mt-proteomes and variation in completeness of individual animal mt-proteomes. The four animal mt-proteomes were characterized using different experimental and bioinformatic approaches. Each of these methods have their pros and cons. For instance, MS based methods face the problem of co-purifying contaminants and missing low-abundance mt-proteins. On the other hand, bioinformatic methods like detection of MTS, would miss non-canonical targeting signals. Mammalian mt-proteomes represent the most complete mt-proteomes of the four animals, because both experimental and computational techniques were applied simultaneously to identify mt-proteins. However, even the best studied mt-proteomes (human and mouse) are suggested to be only around 85% complete (Calvo and Mootha, 2010).

3.6 Conclusion

Animal mt-proteomes vary with respect to size and content. Using computational techniques, we investigated the causes and functional significance of this variation in size and content of animal mt-proteomes. We found that several factors contributed to the above-mentioned size difference: evolution of novel mt-proteins, loss of ancestral mt-proteins, neolocalization of non-mt-proteins to mitochondria and duplication of genes encoding mt-proteins. Evolution of novel mt-proteins was
the primary reason for the increase in size of mammalian mt-proteomes. Functional analysis of these novel mammalian mt-proteins suggest an increased role of mitochondria in cellular signaling. In addition to novel biological functions, these novel mammalian mt-proteins were also additions to existing mitochondrial pathways and complexes. Loss of ancestral mt-proteins in protostomes, like those involved in amino-acid metabolism, also resulted in the size-variation in animal mt-proteomes. Surprisingly, gain of MTS was not a significant contributor to the size-variation. In fact, majority of novel and neolocalized mt-proteins in animals lack an identifiable MTS, suggesting either an increased role of alternate mt-protein import pathways or presence of non-canonical targeting signals which are not identified by existing methods. Experimental characterization of mt-proteomes from other animal lineages, from both bilaterian and non-bilaterian phyla, and development of tools to identify non-canonical mitochondrial targeting signals will help further our understanding of animal mt-proteome evolution.

3.7 Acknowledgement

We gratefully acknowledge Dr. Carolyn Lawrence-Dill, Dr. Karin Dorman, Dr. Iddo Friedberg and Dr. Robert Jernigan for their comments and suggestions regarding the manuscript.

3.8 References


3.9 Tables and Figures

Figure 3.1: Decision tree used to subdivide animal mt-proteomes into four categories.
Figure 3.2: Composition of experimentally determined animal mt-proteomes. Different categories of proteins are shown by colors. The radius of the pie-chart is proportional to the overall size of the mt-proteome. The animal silhouettes were taken from PhyloPic (http://phylopic.org/). The four experimentally determined animal mt-proteomes varied in size from 838 proteins (D. melanogaster) to 1700 proteins (human). Taxonomic groups with characterized mt-proteomes are marked with a black box.
Figure 3.3: **VENN diagram of Orthologous Groups (OGs) and proteins from mt-proteomes of the four animals belonging to the four categories of mt-proteins** (explained in section 3.4.1.1) A-C] VENN diagram of OGs containing at least 1 mt-protein from categories I, II and III in the four bilaterian animals, as described in Section 3.4.1.1. A] Category I: ancestral mt-proteins, B] Category II: proteins which underwent mitochondrial neolocalization in animals, C] Category III: novel animal mt-proteins. D] VENN Diagram of Category IV species-specific mt-proteins from the four bilaterian animals.
Figure 3.4: Percentage of proteins with MTS in all the four categories in animal mt-proteomes.
Table 3.1: Number of OGs and mt-proteins found in each category of bilaterian animal mt-proteome. The four different categories of animal mt-proteins are outlined in Section 3.4.1.1.

<table>
<thead>
<tr>
<th></th>
<th>Ancestral</th>
<th>Neolocalized</th>
<th>Novel</th>
<th>Species-specific</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OG mt-protein</td>
<td>OG mt-protein</td>
<td>OG mt-protein</td>
<td>OG mt-protein</td>
</tr>
<tr>
<td><strong>C. elegans</strong></td>
<td>322</td>
<td>344</td>
<td>192</td>
<td>233</td>
</tr>
<tr>
<td><strong>D. melanogaster</strong></td>
<td>317</td>
<td>332</td>
<td>134</td>
<td>140</td>
</tr>
<tr>
<td><strong>H. sapiens</strong></td>
<td>485</td>
<td>535</td>
<td>303</td>
<td>312</td>
</tr>
<tr>
<td><strong>M. musculus</strong></td>
<td>450</td>
<td>491</td>
<td>274</td>
<td>282</td>
</tr>
</tbody>
</table>

Table 3.2: Proportion of proteins possessing a MTS from individual categories of the bilaterian animal mt-proteomes

<table>
<thead>
<tr>
<th></th>
<th>Category I</th>
<th>Category II</th>
<th>Category III</th>
<th>Category IV</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C. elegans</strong></td>
<td>63.95%</td>
<td>15.88%</td>
<td>47.06%</td>
<td>29.09%</td>
</tr>
<tr>
<td><strong>D. melanogaster</strong></td>
<td>68.07%</td>
<td>28.57%</td>
<td>54.55%</td>
<td>57.26%</td>
</tr>
<tr>
<td><strong>H. sapiens</strong></td>
<td>66.73%</td>
<td>40.06%</td>
<td>48.48%</td>
<td>40.00%</td>
</tr>
<tr>
<td><strong>M. musculus</strong></td>
<td>67.01%</td>
<td>37.59%</td>
<td>48.16%</td>
<td>38.46%</td>
</tr>
</tbody>
</table>
4.1 Abstract

Comparative analysis of animal mitochondrial proteomes faces two challenges: the scattering of data on experimentally-characterized animal mitochondrial proteomes across several databases, and the lack of data on mitochondrial proteomes from the majority of metazoan lineages. In this study, we developed two resources to address these challenges: 1] the Metazoan Mitochondrial Proteome Database (MMPdb), which consolidates data on experimentally-characterized mitochondrial proteomes of vertebrate and invertebrate model organisms, and 2] MitoPredictor, a novel machine-learning tool for prediction of mitochondrial proteins in animals. MMPdb allows comparative analysis of animal mitochondrial proteomes by integrating results from orthology analysis, prediction of mitochondrial targeting signals, protein domain analysis, and Gene Ontology analysis. Additionally, for mammalian mitochondrial proteins, MMPdb includes experimental evidence
of localization from MitoMiner and the Human Protein Atlas. MitoPredictor is a Random Forest classifier which uses orthology, mitochondrial targeting signal prediction and protein domain content to predict mitochondrial proteins in animals.

4.2 Introduction

Mitochondria are involved in various cellular functions, including energy production, amino-acid and lipid metabolism, apoptosis, Fe/S cluster biosynthesis, innate immunity and signaling (Szklarczyk and Huynen, 2010; Meisinger et al., 2008). Mitochondrial proteins (mt-proteins)\(^1\) responsible for these biological processes are contributed by two genomes- the mitochondrial genome (mt-genome)\(^2\) and the nuclear genome. However, the extent of contribution is asymmetric, with nearly 99% of mt-proteins encoded in the nuclear genome and imported into the organelle (Becker et al., 2012; Chacinska et al., 2009; Gabaldón and Huynen, 2004). Thus, characterization of nuclear-encoded mt-proteins- i.e. the mitochondrial proteome\(^3\)- is critical for understanding mitochondrial function and evolution.

While the evolution and conservation of the mt-proteome has been studied in some eukaryotic lineages, the mt-proteome within animals remains relatively unexplored. Two major challenges are encountered while carrying out comparative analysis of animal mt-proteomes: 1] data for experimentally-characterized mt-proteomes in animals are scattered across several different databases and 2] majority of animal lineages lack an experimentally-characterized mt-proteome. In this study, we developed two resources to help address these challenges: 1] the Metazoan Mitochondrial Proteome Database, which consolidated information on experimentally-characterized animal mt-proteomes, and 2] MitoPredictor, a novel machine-learning tool for prediction of mt-proteins in animals.

Data for experimentally-characterized mt-proteomes form vertebrate and invertebrate animals are currently scattered across several databases. Some of these databases house information for a

\(^{1}\text{mt-proteins: mitochondrial proteins}\)

\(^{2}\text{mt-genome: mitochondrial genome}\)

\(^{3}\text{mt-proteome: mitochondrial proteome}\)
single species, such as the GLAD (Gene List Annotation for Drosophila) (Hu et al., 2015) database (Drosophila melanogaster), HMPdb (https://bioinfo.nist.gov/) (human), MitoProteome (Cotter et al., 2004) (human), and MitoPhenome (Scharfe et al., 2009) (human), while others contain information from multiple species, like MitoMiner v4.0 (human, mouse and rat) (Smith and Robinson, 2016). Some mt-proteomes have not been deposited in any database, like the Caenorhabditis elegans mt-proteome from the study by Jing et al (Li et al., 2009a). No one single database includes mt-proteins from both vertebrate (human and mouse) and invertebrate (C. elegans and D. melanogaster) model organisms.

In this study, we created the Metazoan Mitochondrial Proteome database (MMPdb), where we consolidated data on experimentally-verified mt-proteomes from four animals: human, mouse, C. elegans and D. melanogaster, and two outgroup species: the protist Acanthamoeba castellanii (Gawryluk et al., 2014a) and yeast Saccharomyces cerevisiae (Cherry et al., 2012). To facilitate the comparative analysis of animal mt-proteomes, MMPdb includes the following information for all proteins in these organisms: 1] orthology predictions by Proteinortho v5.16b (Lechner et al., 2011), 2] Mitochondrial Targeting Signal (MTS) predictions by TargetP v1.1 (Emanuelsson et al., 2007) and MitoFates (Fukasawa et al., 2015), 3] protein-domain content and 4] associated Gene Ontology (GO) terms. Additionally, for mammalian proteins, MMPdb includes experimental evidence of mitochondrial localization from MitoMiner v4.0 and Human Protein Atlas v19 (Thul and Lindskog, 2018). MMPdb is available as an R Shiny applet for visual exploration of the animal mt-proteomes.

The number of experimentally-verified mt-proteomes in animals has not kept pace with the total number of genomes/transcriptomes sequenced. Even well-studied model organisms like Danio rerio (zebrafish) lack comprehensive data on mt-proteomes. Additionally, efforts in experimental characterization of mt-proteomes in animals have been restricted to Bilateria, which represents just one of the five main branches in the animal phylogenetic tree. Thus, a practical solution for most animal species lacking such experimental data is to predict mt-proteins bioinformatically. To help

---

4MMPdb: Metazoan Mitochondrial Proteome Database
5MTS: mitochondrial targeting signal
6GO: Gene Ontology
address this issue, we developed a novel machine-learning tool: MitoPredictor. MitoPredictor is a Random Forest-based (Breiman, 2001) tool for predicting mt-proteins in animals, which uses three sources of information: 1] orthology to known mt-proteins, 2] the predicted presence of MTS and 3] protein-domain content.

In the present work, we introduce and demonstrate the utility of these two resources for facilitating comparative analysis and visual exploration of animal mt-proteomes. First, we describe the MMPdb, an R Shiny-based database, which functions as a single portal to the data on experimentally-characterized mt-proteomes in animals. Second, we introduce MitoPredictor, a novel machine-learning tool, for accurate prediction of mitochondrial localization of animal proteins. Furthermore, we show that MitoPredictor outperforms SubCons, an existing ensemble method for prediction of protein subcellular localization.

4.3 Materials and Methods

4.3.1 The Metazoan Mitochondrial Proteome Database

The Metazoan Mitochondrial Proteome Database (MMPdb) is coded in R and contains information for six eukaryotic species: Homo sapiens, Mus musculus, Caenorhabditis elegans, Drosophila melanogaster, Acanthamoeba castellanii and Saccharomyces cerevisiae. The information in MMPdb is organized in four datasets: 1] Orthology, 2] Mitochondrial Targeting Signal (MTS), 3] Protein domain, and 4] Gene Ontology. The first three datasets were generated in our companion study [Chapter 3]. In addition, we used Pannzer v2.0 (Protein ANNotation with Z-scoRE)(Törönen et al., 2018) to identify GO terms associated with each protein for the Gene Ontology dataset. Each of the four datasets is available as a CSV (comma-delimited) file, and represents one tabset in the R Shiny applet. Below, we describe the four datasets in more detail.

1. Orthology dataset

We used Proteinortho v5.16b (Lechner et al., 2011) to identify orthologous groups (OGs)\(^7\) in the complete proteomes of the four animals and the two outgroups mentioned above [Chapter 3].

\(^7\)OG: Orthologous Group
3]. The results of this analysis formed the basis of the orthology dataset. Each OG was assigned a unique identifier, in the form “OG[XYZ]”, where “XYZ” is the OG number.

2. MTS dataset

In our comparative analysis of mt-proteomes of animals [Chapter 3], we used TargetP v1.1 and MitoFates to predict the presence of MTS in proteins from the six eukaryotic species mentioned above. Here, we used those predictions to create the MTS dataset. From the TargetP predictions, attributes included in the MTS dataset were: 1] the prediction of sub-cellular localization of protein (mitochondria, secretory pathway, other) and the Reliability Class (RC)\(^8\) (a statistic ranging from 1-5, representing the strength of the prediction, with 1 being the strongest prediction and 5 being the weakest). From the MitoFates predictions, we included : 1] the probability of prediction of MTS (ranging from 0-1), and 2] the Mitochondrial Processing Peptidase (MPP)\(^9\) cleavage site.

3. Protein domain dataset

Protein domains are structural and functional units of proteins, associated with distinct biochemical functions. The protein domain dataset contains a list of all domains identified in mitochondrial and non-mitochondrial proteins of the six eukaryotes. This dataset contains, both the PFAM accession number (e.g. “PF00153”) as well as the PFAM identifier (e.g. “Mito_carr”) (Bateman et al., 2004).

4. Gene Ontology dataset

The Gene Ontology dataset contains GO terms associated with all proteins from the six eukaryotic species. Pannzer v2.0 was used to identify GO terms, with default parameters (Törönen et al., 2018). Pannzer v2.0 uses four tools for Gene Ontology analysis (RM3, ARGOT, JAC, HYGE). ARGOT (Annotation Retrieval of Gene Ontology Terms) (Fontana et al., 2009) predictions were used for assigning GO terms to the eukaryotic proteins since it has been reported to be one of the top predictors in their benchmark studies.

\(^8\)RC: reliability class

\(^9\)MPP: mitochondrial processing peptidase
In addition to the tabset-specific attributes described above, some attributes are shared across all the four tabsets:

1. Known subcellular localization of the protein (mitochondrial/non-mitochondrial)

2. Length of the protein in amino-acids

3. For mammalian mt-proteins, experimental evidence of mitochondrial localization from MitoMiner v4.0: from mass-spectrometry studies and from GFP-analysis. Mitochondrial localization evidence from the Human Protein Atlas is also included for the human mt-proteins.

4. For human proteins, tissue-enrichment data from Human Protein Atlas is included. According to the Human Protein Atlas, tissue-enriched genes, are genes which have at least four-fold higher mRNA level in a particular tissue, compared to any other tissues.

The four datasets described above are input files for an R Shiny applet. This R Shiny applet allows for visualization, analysis and download of information and sequences from the four datasets. The app can be run by giving the R command `runApp("[directory where the app files are stored]")` or by using the “Run App” button in RStudio (RStudio Team, 2015). Source code and data for MMPdb can be downloaded directly from the GitHub repository (https://github.com/virajmuthye/database).

4.3.2 MitoPredictor

MitoPredictor is a novel machine-learning tool for prediction of mt-proteins in animals (Figure 4.1). It uses three sources of information: 1] orthology to known mt-proteins, 2] the predicted presence of an MTS and 3] protein domain content. MitoPredictor uses publicly available software (Proteinortho v5.16b, TargetP v1.1, MitoFates, CD-HIT (Fu et al., 2012), HMMer (Eddy, 2011)) and databases (Pfam-A). Below, we describe MitoPredictor in more detail.
4.3.2.1 Step I: Processing input query proteins

The input for MitoPredictor is a set of protein sequences, in FASTA format, from a genome or transcriptome of an animal, for which mt-proteins are not known. The input proteome is referred to as the “query” proteome, and all the proteins contained within it are referred to as the “query” proteins. First, all proteins smaller than 100 amino-acids are removed. Then, CD-HIT is used to remove redundancy in the proteome, which is often the result of mis-assembly. CD-HIT clusters proteins based on sequence similarity, and returns one protein sequence (the longest sequence) as the representative sequence for the cluster. The sequence identity cut-off used is 98%, but can be changed by the user for a more strict or lenient approach. The resulting set of query proteins is used for step II (Orthology) and step IV (Protein domain). TargetP and MitoFates require a complete N-terminus for accurate prediction. Thus, for step III (MTS), only query proteins beginning with a methionine at position 1 are used as input for TargetP and MitoFates, so as to exclude potential N-terminus fragments and reduce the number of false-positives in MTS predictions.

4.3.2.2 Steps II-IV: Feature extraction

After processing the input query proteins (step I), MitoPredictor analyzes each query protein for orthology with known mt-proteins (step II), presence of MTS (step III), and presence of protein domains common in mt-proteins (step IV). In steps II-IV, MitoPredictor extracts seven features for prediction of mt-proteins. Below, we provide a brief description each step.

1. Step II Identification of mitochondrial orthologs in the query proteome

In this step, MitoPredictor identifies query proteins that are orthologous to reference mt-proteins. Proteomes of five eukaryotic species: *Homo sapiens, Mus musculus, Caenorhabditis elegans, Drosophila melanogaster*, and *Saccharomyces cerevisiae* are used for this analysis, and are labelled as “reference proteomes”. The mt-proteome of *Acanthamoeba castellanii* is not used in this approach to reduce the time required for orthology-detection. Proteinortho v.5.16b is used to identify groups of orthologous proteins (OGs) from the five reference proteomes and the query proteome. For each query protein, an “OrthoScore” (OS) is assigned-
1 if the query protein is orthologous to any reference mt-protein and 0 if the query protein is not orthologous to any reference mt-protein.

2. **Step III Prediction of MTS in the query proteome**

   In step III, MitoFates and TargetP v1.1 are used for prediction of MTS in the query proteins. TargetP is run using the “Non-plant” option and MitoFates is run using the “Metazoa” option. From the prediction results of TargetP and MitoFates, the following features are extracted:

   (a) **TargetP**: 1] mTP (probability of possessing a MTS), 2] sTP (probability of possessing a signal peptide), and 3] other (probability that the protein is localized to any other subcellular location)

   (b) **MitoFates**: 1] probability of possessing a MTS and 2] net-charge.

3. **Step IV Protein domain analysis**

   In addition to homology-search and MTS-prediction, protein domain information has been used for prediction of mt-proteins in animals (Kumar et al., 2018). The availability of experimentally-characterized mt-proteomes from multiple animals provides information about the association of protein domain content and protein subcellular localization, which is used by MitoPredictor in step IV to predict mt-proteins in the query proteome. In this step, MitoPredictor identifies protein domains in the query proteins, and assigns a Domain Score (DS) to each query protein. Below, we describe how this feature (DS) is calculated.

   For all the reference animal proteins, protein domain annotations were performed using the Pfam-A database (release-32.0). For each protein domain \( d \) identified in majority of the reference animal proteomes (at least three species), the domain score \( D_{ds} \) was calculated as follows:

\[
D_{ds} = \frac{P_{mito}}{P_{total}}
\]
where $P_{\text{mito}}$ and $P_{\text{total}}$ refer to the number of mitochondrial and total (mitochondrial + non-mitochondrial) proteins containing the protein domain $d$ in species $s$. The $D_{ds}$ ranges from 0 to 1, where a score of 0 indicates that the domain $d$ is present only in non-mt-proteins in species $s$, and a score of 1 means that the domain $d$ is present only in mt-proteins in the species $s$. Next, the final domain score for each domain was calculated as the average of the top three $D_{d(s)}$ scores:

$$DS_d = \frac{1}{3} \sum_{s=1}^{3} D_{d(s)}$$

where $D_{d(s)}$ are the domain scores for domain $d$, sorted from largest to smallest. Here, we only consider the top three $D_{d(s)}$ scores for the domain $d$. The $D_{d(s)}$ ranges from 0 to 1, where a score of 1 indicates that all proteins possessing the domain $d$, in the majority of the animals used in this study, are mt-proteins. Conversely, an $D_{d(s)}$ of 0 means that all the proteins possessing the domain $d$, in the majority of animals used in this study, are non-mt-proteins.

A protein domain library is constructed, which contains the protein domain and its associated $DS_d$ score.

In step IV, MitoPredictor identifies protein domains in query proteins and assigns a Domain Score (DS) to query proteins. If a query protein possesses a single protein domain $d$ present in the MitoPredictor protein domain library, the Domain Score (DS) for that query protein is the $DS_d$ associated with that domain $d$. If a query protein contains multiple protein domains from the MitoPredictor protein domain library, then the highest $DS_d$ score is assigned as the Domain Score (DS) for the query protein.

### 4.3.2.3 Step V: Prediction of query mt-proteins using a Random Forest classifier

In step V, MitoPredictor uses a Random Forest classifier to predict mt-proteins in the query proteome, based on the seven features extracted in steps II-IV (Table 4.1). The final result is the generation of a “Final Matrix” file. This final matrix includes orthology, MTS, domain and
known/prediction localization information for all the proteins in both the reference and query proteomes, and is the input file for the R Shiny applet. The R Shiny applet allows for visualization and exploration of the predicted query mt-proteome. A folder named “stats” is created, which includes basic information regarding the predicted mt-proteome.

4.3.2.4 MitoPredictor model development and testing

1. Datasets

We used two datasets during the development and demonstration of MitoPredictor.

(a) **Golden dataset**: The “golden dataset” refers to the set of 1,225 human proteins with experimentally-verified subcellular localization, which was compiled in the study by Salvatore et. al, 2017 (Salvatore et al., 2017). These proteins were used as an independent test set in this study to evaluate the performance of the mt-protein predictors.

(b) **MitoPredictor dataset**: All the seven features, listed in Table 4.1, were extracted for all the proteins from the four reference animal species used in MitoPredictor: human, mouse, *C. elegans*, and *D. melanogaster*. CD-HIT was used to cluster the complete proteomes of the four reference animal species at 70% similarity. The longest protein sequence was retained as the representative protein sequence. From the remaining set of proteins, the golden dataset proteins, along with their orthologs, were removed. This final set of proteins is referred to as the “MitoPredictor” dataset.

2. **Model selection and implementation**

Prediction of mt-proteins in the query proteome represents a supervised binary classification problem. Here, we compared the performance of three widely-used ML algorithms: Random Forest (RF)\(^\text{10}\), Logistic Regression (LR)\(^\text{11}\) and Support Vector Machine (SVM)\(^\text{12}\). To check the performance of the three ML algorithms, the entire MitoPredictor dataset was used. For each algorithm, 10-fold cross validation was carried out. Commonly-used performance metrics were used: F1 Score, Matthews Correlation

\(^{10}\text{RF}: \text{Random Forest}\)

\(^{11}\text{LR}: \text{Logistic Regression}\)

\(^{12}\text{SVM}: \text{Support Vector Machine}\)
Coefficient (MCC)\textsuperscript{13}, Area Under the Receiver Operating Characteristic Curve (AUROC)\textsuperscript{14} and Area Under Precision-Recall Curve (AUPRC)\textsuperscript{15}. The average values of these performance metrics were used to select the best-performing ML algorithm. Our results showed that the performance of the algorithms were very similar (Supplementary Materials 4.8.1). For MitoPredictor, we selected the Random Forest model because it had a slightly higher MCC and F1 score compared to the other two algorithms. Details regarding model evaluation are given in Supplementary Materials 4.8.2.

3. Comparison of MitoPredictor to existing subcellular localization predictors

To evaluate the performance of MitoPredictor, we compared its performance to that of SubCons, an ensemble predictor of protein subcellular localization (Salvatore et al., 2017), as well as TargetP and MitoFates, which rely solely on the predicted presence of an N-terminus MTS to infer mitochondrial localization. SubCons utilizes prediction results from four subcellular localization predictors (CELLO2.5 (Yu et al., 2006), LocTree2 (Goldberg et al., 2012), MultiLoc2 (Blum et al., 2009) and SherLoc2 (Briesemeister et al., 2009)). For comparing the performance of both predictors, the golden dataset was used, since proteins from the golden dataset, and their orthologs, were removed from the training and testing datasets from both predictors. First, we compared the performance of the two predictors, as trained by their respective authors, on the golden dataset.

We also compared the performance of features from both MitoPredictor and SubCons on our Random Forest classifier. For that, we extracted the features used in the SubCons algorithm, which are the scores of the four individual predictors (CELLO2.5, LocTree2, MultiLoc2 and SherLoc2) for nine subcellular localizations (nucleus, cytoplasm, mitochondrion, extracellular space, plasma membrane, peroxisome, endoplasmic reticulum, Golgi apparatus, and lysosome). For proteins which lacked SherLoc2 predictions in the SubCons output, SherLoc2 was used to extract those features. We evaluated the performance of the following sets of

\textsuperscript{13}MCC: Matthews Correlation Coefficient
\textsuperscript{14}AUROC: Area Under the Receiver Operating Characteristic Curve
\textsuperscript{15}AUPRC: Area Under Precision-Recall Curve

To use the most recent localization labels (mitochondrial/non-mitochondrial) for human proteins, we updated the subcellular localization labels of the golden dataset proteins with the labels from MitoCarta v2.0 (Calvo et al., 2016) and IMPI Q2 2018 (Smith and Robinson, 2016).

4.3.2.5 Data availability

Source code for MitoPredictor can be downloaded from our GitHub repository (https://github.com/virajmuthye/mitopredictor).

4.4 Results and Discussion

4.4.1 The Metazoan Mitochondrial Proteome Database

The Metazoan Mitochondrial Proteome Database (MMPdb) consolidates data on experimentally-characterized mt-proteomes of bilaterian animals from different sources to facilitate comparative analysis of animal mt-proteins. MMPdb integrates data on the mt-proteomes of four animals: Homo sapiens, Mus musculus, Caenorhabditis elegans and Drosophila melanogaster, along with two outgroups: Acanthamoeba castellanii and Saccharomyces cerevisiae. MMPdb is organized in four tabsets: MTS, Orthology, Domain and Gene Ontology (Figure 4.2A). Each of the four tabsets can be filtered by the following criteria (Figure 4.2B):

1. Species

2. Known subcellular localization of protein (mitochondrial or non-mitochondrial)

3. Presence/absence of MTS as predicted by TargetP and MitoFates
In each tabset, a bar-graph visualizes the number of proteins from each species selected by the user (Figure 4.2D). The user can download either the sequences of the selected proteins in a FASTA format, or the entire dataset for the selected proteins in a CSV (comma-separated value) format. Additionally, for mammals only, each of the four tabsets can be further filtered by experimental evidence of mitochondrial localization from MitoMiner v4.0. For example, a user can fetch all proteins from mammals which have been identified as mitochondrial by at least one mass-spectrometry study or GFP-analysis. Human proteins can be further filtered by tissue-enrichment data from the Human Protein Atlas. For instance, a user can download all human testis-enriched mt-proteins for further analysis.

The orthology tabset enables users to analyze proteins belonging to a specified orthologous group (OG). For instance, this tabset can be used to fetch the mitochondrial inner membrane ADP/ATP translocase proteins from the six eukaryotic species (Figure 4.3A). The user can explore the various attributes of the proteins in the OG, such as MTS characteristics, protein-domain content, and localization evidence. The About tabset provides an easy-to-use search-tool for mapping protein identifiers to OG identifiers. Proteins from any selected OG(s) can be downloaded in a FASTA format for further analysis.

The MTS tabset allows for the exploration of the MTS prediction results from TargetP and MitoFates. One potential use of this tabset is to identify candidate novel species-specific mitochondrial proteins: non-mitochondrial proteins which lack orthologs in remaining species, and possess an MTS. In animals, the number of such proteins ranges from 74 in human to 295 in *C. elegans* (Figure 4.3B). For mammals, a user can further filter the database based on the experimental evidence of mitochondrial localization, like GFP-analysis and mass-spectrometry studies.

The protein domain tabset in MMPdb can be used to select proteins containing a specific domain. For example, a user can extract all proteins containing the mitochondrial carrier domain (“Mito_carr”) (Figure 4.3C). This domain is present in the mitochondrial carrier proteins, which are involved in the import of metabolites across the inner-mitochondrial membrane. The result of this query shows that the number of mammalian mt-proteins with the domain is double that
of the invertebrate proteins. However, a large number of proteins with the domain, 22 and 24 in *C. elegans* and *D. melanogaster* respectively, are annotated as “non-mitochondrial”. It is likely that these annotations are not correct since only mt-proteins contain the “Mito_carr” domain in analyzed mammalian species. Additionally, this tabset also provides a bar-graph displaying the number of different protein domain-combinations for a specific domain. For example, one can see that the “Pkinase” (protein kinase domain) exists in just 1 and 3 domain combinations in *C. elegans* and *D. melanogaster* mt-proteins respectively, compared to 10 and 13 combinations in human and mouse (data not shown).

The Gene Ontology tabset allows users to analyze proteins, which are mapped to a particular user-specific GO ID. For example, a user can analyze animal proteins, which are mapped to the GO term “mitochondrial inner membrane” (GO:0005743) (Figure 4.3D). This query reveals that the number of mitochondrial inner-membrane proteins in mammals is nearly double the number of mitochondrial inner-membrane proteins in the two invertebrates. This difference in the number of inner-membrane proteins between the two mammals and the two invertebrates can be explained by either a gain of proteins in mammals (as in the case of Complex I of the electron transport chain (Gabaldón et al., 2005b)) or by a loss of proteins in the two invertebrates. Indeed, both *C. elegans* and *D. melanogaster* are known to have experienced extensive gene-loss (Albalat and Cañestro, 2016).

It is important to note that the utility of this database extends beyond the analysis of animal mt-proteins. All four tabsets contain information for both mitochondrial and non-mitochondrial proteins from the six eukaryotes. Thus, a user can explore and download orthologs of non-mitochondrial proteins from the six eukaryotes. TargetP predictions include the probability of possessing a signal peptide, *i.e.* the protein is a part of the secretory pathway. Similar to the mt-proteome, a user can analyze the predicted “secretome” (set of all secreted proteins) in animals.
4.4.2 MitoPredictor

MitoPredictor is a novel machine-learning tool for prediction of mt-proteins in animals. MitoPredictor uses orthology, predicted presence of MTS and protein domain content to predict mt-proteins. Proteins from the reference animal species (*Homo sapiens*, *Mus musculus*, *Caenorhabditis elegans* and *Drosophila melanogaster*) were used to train and test a Random Forest classifier for prediction of mt-proteins in the query proteome. The golden dataset was used as an independent test set to evaluate the performance of MitoPredictor and compare its performance to SubCons.

First, we compared the performance of MitoPred and SubCons, as trained by their respective authors, on the golden dataset. Additionally, we compared the performance of the both MitoPredictor and SubCons to TargetP and MitoFates. Both MitoPredictor and SubCons outperformed TargetP and MitoFates, which use only the predicted presence of an N-terminus MTS to infer mitochondrial localization (Table 4.2). MitoPredictor outperformed SubCons, and had a higher MCC (0.898 (MitoPredictor) v/s 0.737 (SubCons)) and F1 score (0.916 (MitoPredictor) v/s 0.784 (SubCons)).

In MitoPredictor, seven features are used to predict mt-proteins using a Random Forest classifier (Table 4.1). We evaluated the importance of each feature using the mean decrease in Gini values. Orthology analysis played the most important role in mt-protein prediction in MitoPredictor, since the largest mean decrease in Gini value was observed for the OrthoScore (OS) feature. Interestingly, the second-highest mean decrease in Gini was observed for the Domain Score (DS), a feature we developed for this study (Figure 4.4).

Next, we evaluated the performance of features from MitoPredictor and SubCons in our Random Forest classifier using the same training and testing dataset. From the MitoPredictor dataset, a training dataset, comprising of randomly selected 2,000 mitochondrial and non-mitochondrial proteins, and a testing dataset, comprising of randomly selected 760 mitochondrial and non-mitochondrial proteins was generated. The training and testing datasets were mutually exclusive, *i.e.* no protein was a part of both datasets. The features used for prediction of protein subcellular localization from both MitoPredictor and SubCons were extracted for proteins from the training
Our Random Forest model was used to evaluate the performance of prediction features from MitoPredictor and SubCons on the golden dataset. We report that features from MitoPredictor outperformed features from SubCons (Table 4.3). Additionally, we examined whether the performance of MitoPredictor could be improved by addition of features from SubCons. We evaluated the performance of the following combinations of features: 1] only MitoPredictor, 2] MitoPredictor and Cello2.5, 3] MitoPredictor and LocTree2, 4] MitoPredictor and MultiLoc2, 5] MitoPredictor and SherLoc2, and 6] MitoPredictor and all features from SubCons. None of the features from SubCons, when added to MitoPredictor, improved the performance of MitoPredictor (Supplementary Materials 4.8.3).

The three categories of features used in prediction of mt-proteins in MitoPredictor, i.e. MTS prediction, orthology-analysis and protein domain content, have their individual pros and cons, with respect to protein subcellular-localization prediction. For example, MTS prediction is an important approach to predict mt-proteins, since the MTS pathway is the predominant pathway of mitochondrial protein import. However, only around 60-70% of mt-proteins possess a detectable MTS (Wiedemann and Pfanner, 2017; Fukasawa et al., 2015). Proteins lacking the canonical N-terminus MTS are involved in critical mitochondrial functions, and include all mitochondrial outer-membrane proteins and majority of the mitochondrial inner-membrane proteins (Wiedemann and Pfanner, 2017). Additionally, MTS predictors also generate false-positives in predictions. Thus, an MTS-only approach would not only miss important mt-proteins (false-negatives) but also result in several mis-annotations (false-positives).

In MitoPredictor, we identified OrthoScore (OS) as the most important feature involved in mt-protein prediction. However, several proteins have been known to undergo lineage-specific neolocalization, i.e. localized to mitochondria in one lineage, and to another subcellular location in another lineage. In such a situation, an orthology-alone approach would lead to an erroneous prediction of mt-proteins. Protein domain content has been shown to be useful in prediction of mt-proteins. Even in MitoPredictor, it was the second-most important feature used in the classifier. In this study, we see that while each category of features has its pros and cons, when used together,
they provide a powerful tool for prediction of mt-proteins. There are several advantages of Mito-
Predictor, when compared to SubCons. MitoPredictor requires fewer software to be downloaded
and installed prior to use, compared to SubCons. MitoPredictor is trained on experimentally-
characterized mt-proteins from both invertebrate and vertebrate animals, where as SubCons is
trained on only human proteins. Additionally, output from MitoPredictor is specifically geared
towards facilitating comparative analysis of animal mt-proteomes. In addition to generating a file
of predicted mt-proteins in FASTA format, MitoPredictor also provides an easy-to-use R Shiny
applet for visual exploration of the predicted proteins. The organization of the final R Shiny applet
is similar to MMPdb. MitoPredictor also includes important information regarding the predicted
mt-proteome in a folder called “stats”. This includes a list of potential species-specific mt-proteins,
i.e. predicted mt-proteins in the query proteome which have no ortholog to any other reference
protein. MitoPredictor also outputs lists of mt-proteins from each reference species, with and with-
out orthologs to the predicted query mt-proteins. These lists can be used as input lists for several
tools such as PantherDB (Mi et al., 2016) and ConsensusPathDB (Kamburov et al., 2010) to analyze
conservation and potential loss of mitochondrial proteins in the predicted query mt-proteome.

4.5 Acknowledgement

We want to thank Dr. Karin Dorman, Dr. Iddo Friedberg, Dr. Carolyn Lawrence-Dill, and Dr.
Robert Jernigan for their suggestions and improvements to the work. We would also like to thank
Akshay Yadav for his comments and suggestions on MitoPredictor. We want to thank the review-
ers for their insights and comments, which helped improve the manuscript. We wish to sincerely
acknowledge Dr. Arne Elofsson and Dr. Marco Salvatore for their assistance with SubCons.
4.6 References


4.7 Tables and Figures

Figure 4.1: Overflow of MitoPredictor. In feature extraction, the number of features extracted in each step are mentioned in parentheses. The list of features used in MitoPredictor is also listed in Table 1.
Figure 4.2: **General structure of the MMPdb R Shiny applet.** Here the orthology tabset is depicted. A] The available tabsets of MMPdb are shown here. The About tabset provides an overview of the database and also contains instructions on usage of the remaining tabsets. B] The different criteria for filtering the database can be selected here. C] Primary search query field. For orthology tabset, the unique OG identifier is the primary search query. The OG identifier can be easily mapped to the protein identifier in the About tabset. Here, the OG identifier for the mitochondrial ADP/ATP translocase ("OG76") is entered. For domain tabset and gene-ontology tabset, the primary search query is the domain name (e.g. "Mito_carr") and GO ID (e.g. "GO:0005739" for mitochondrion) respectively. D] A bar-graph depicting the number of proteins selected from each species after filtering. E] Data-table for proteins selected from each species after filtering. Key for interpretation of column names is provided in the sidebar panel [B] (not shown in this image).
Figure 4.3: Utility of MMPdb. A] Orthology tabset was used to fetch ADP/ATP translocase proteins using the unique Orthologous Group identifier (“OG76”) as the query. B] MTS tabset was used to fetch all non-mitochondrial proteins from all six species, which are predicted to possess an MTS and have no ortholog in any other species in this study. C] Domain tabset was used to fetch all mt-proteins possessing a mitochondrial carrier protein domain, using the domain name “Mito_carr” as the query. D] Gene ontology tabset was used to identify all mt-proteins which were mapped to the GO term “inner mitochondrial membrane” by Pannzer v2.0. Images of the graph are taken directly from the MMPdb applet.
Figure 4.4: Mean decrease in Gini values for each of the seven features from MitoPredictor used in the Random Forest classifier. A large mean decrease in Gini value indicates higher importance of the feature. Here, the largest mean decrease was observed for the OrthoScore (OS), followed by the Domain Score (DS).
Table 4.1: List of features extracted from each step of MitoPredictor (Figure 4.1).

<table>
<thead>
<tr>
<th>Step</th>
<th>Features extracted</th>
</tr>
</thead>
<tbody>
<tr>
<td>II</td>
<td>Orthology</td>
</tr>
<tr>
<td></td>
<td>1) OrthoScore (OS)</td>
</tr>
<tr>
<td>III</td>
<td>MTS prediction</td>
</tr>
<tr>
<td></td>
<td>2) Probability of protein possessing a MTS from TargetP (mTP)</td>
</tr>
<tr>
<td></td>
<td>3) Probability of protein possessing a signal peptide from TargetP (sTP)</td>
</tr>
<tr>
<td></td>
<td>4) Probability that protein localizes to any other subcellular location from TargetP (other)</td>
</tr>
<tr>
<td></td>
<td>5) Probability of protein possessing a MTS from MitoFates (mfpred)</td>
</tr>
<tr>
<td></td>
<td>6) Net-charge from MitoFates (charge)</td>
</tr>
<tr>
<td>IV</td>
<td>Protein domain analysis</td>
</tr>
<tr>
<td></td>
<td>7) Domain Score (DS)</td>
</tr>
</tbody>
</table>

Table 4.2: Prediction performance of MitoPredictor, SubCons, TargetP, and MitoFates on the golden dataset. The predictors were used, as trained by their respective authors, to predict mitochondrial proteins in the golden dataset. The following performance metrics were calculated: F1-score, AUROC: Area Under the Receiver Operating Characteristic Curve, AUPRC: Area Under Precision-Recall Curve, MCC: and Matthews Correlation Coefficient.

<table>
<thead>
<tr>
<th></th>
<th>MitoPredictor</th>
<th>SubCons</th>
<th>TargetP</th>
<th>MitoFates</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCC</td>
<td>0.898</td>
<td>0.737</td>
<td>0.529</td>
<td>0.657</td>
</tr>
<tr>
<td>F1 score</td>
<td>0.916</td>
<td>0.784</td>
<td>0.628</td>
<td>0.688</td>
</tr>
<tr>
<td>AUROC</td>
<td>0.959</td>
<td>0.845</td>
<td>0.686</td>
<td>0.759</td>
</tr>
<tr>
<td>AUPRC</td>
<td>0.991</td>
<td>0.913</td>
<td>0.800</td>
<td>0.834</td>
</tr>
</tbody>
</table>
Table 4.3: Performance of features from MitoPredictor and SubCons on the golden dataset on our Random Forest classifier.

<table>
<thead>
<tr>
<th></th>
<th>MitoPredictor</th>
<th>SubCons</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCC</td>
<td>0.899</td>
<td>0.675</td>
</tr>
<tr>
<td>F1 score</td>
<td>0.918</td>
<td>0.743</td>
</tr>
<tr>
<td>AUPRC</td>
<td>0.965</td>
<td>0.874</td>
</tr>
<tr>
<td>AUROC</td>
<td>0.992</td>
<td>0.933</td>
</tr>
</tbody>
</table>

4.8 Supplementary Materials

4.8.1 Selection of machine-learning algorithm

Table 4.4: Comparing the performance of the three machine-learning algorithms (RF: Random Forest, LR: Logistic Regression and SVM: Support Vector Machine). For each algorithm, 10-fold cross validation was performed. The following performance metrics were calculated: F1, AUROC: Area Under the Receiver Operating Characteristic Curve, AUPRC: Area Under Precision-Recall Curve, MCC: and Matthews Correlation Coefficient.

<table>
<thead>
<tr>
<th></th>
<th>RF</th>
<th>LR</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCC</td>
<td>0.944</td>
<td>0.939</td>
<td>0.933</td>
</tr>
<tr>
<td>F1</td>
<td>0.972</td>
<td>0.970</td>
<td>0.967</td>
</tr>
<tr>
<td>AUPRC</td>
<td>0.987</td>
<td>0.987</td>
<td>0.948</td>
</tr>
<tr>
<td>AUROC</td>
<td>0.990</td>
<td>0.990</td>
<td>0.966</td>
</tr>
</tbody>
</table>
4.8.2 Evaluation of the Random Forest model

- Randomization experiments

First, we assessed the impact of randomization in generating training and testing datasets on model performance. We created 500 random training and testing datasets. Each training dataset contained 2,000 randomly selected mitochondrial and non-mitochondrial proteins. Each testing dataset contained randomly selected 760 mitochondrial and non-mitochondrial proteins. For each iteration, we trained the Random Forest model on the training dataset, and tested model performance on the testing dataset. For each iteration, performance metrics listed in section 7.1.1 were calculated. We examined the distribution of values for these performance metrics. Results of this randomization analysis show that there is very little variance in performance of MitoPredictor on the random datasets (Figure 4.5).
Figure 4.5: **Box-plot depicting distribution of performance metrics for the randomization analysis**. In this analysis, for 500 iterations, training and testing datasets were generated by randomly selecting 2000 mitochondrial and non-mitochondrial proteins from the MitoPredictor dataset as the training dataset and 760 mitochondrial and non-mitochondrial proteins as the testing dataset. The following performance metrics were evaluated at each iteration: AUROC: Area Under the Receiver Operating Characteristic Curve, AUPRC: Area Under Precision-Recall Curve, MCC: Matthews Correlation Coefficient.
• **10-fold cross-validation**

We performed 10-fold cross-validation to analyze our model performance. 10-fold CV was performed using the entire MitoPredictor dataset. The entire dataset was partitioned into 10 partitions. For 10 iterations, one of the 10 partitions was used as the testing dataset, while the remaining nine partitions were used as the training dataset. Performance metrics listed in section 7.1.1 were calculated. Our results showed that there was very little variation in the values of the performance metrics (Figure 4.6).

![Box-plot depicting distribution of performance metrics for the 10-Fold Cross-Validation analysis](image)

**Figure 4.6:** Box-plot depicting distribution of performance metrics for the 10-Fold Cross-Validation analysis. 10-fold CV was performed using the entire MitoPredictor dataset. The entire dataset was partitioned into 10 partitions. For 10 iterations, one of the 10 partitions was used as the testing dataset, while the remaining nine partitions were used as the training dataset. For each of the 10 iterations, the following performance metrics were evaluated: AUROC: Area Under the Receiver Operating Characteristic Curve, AUPRC: Area Under Precision-Recall Curve, MCC: Matthews Correlation Coefficient
4.8.3 Evaluation of prediction performance of MitoPredictor and SubCons features

Table 4.5: Evaluation of prediction performance of features from MitoPredictor and SubCons in the Random Forest classifier. For MitoPredictor, the seven features used in prediction of mt-proteins was the input for the Random Forest classifier. The input features for SubCons, and the individual predictors used in the SubCons algorithm, are listed in Section 2.2.4. The combination of features used are listed in the Table above. For each combination of features, 10-fold cross-validation was carried out and the average value of each performance metric (MCC, F1, AUPRC, AUROC) was calculated.

<table>
<thead>
<tr>
<th>Features from MitoPredictor and features from</th>
<th>CELLO2.5</th>
<th>LocTree2</th>
<th>MultiLoc2</th>
<th>SherLoc2</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCC</td>
<td>0.944</td>
<td>0.939</td>
<td>0.944</td>
<td>0.944</td>
<td>0.943</td>
</tr>
<tr>
<td>F1 score</td>
<td>0.972</td>
<td>0.969</td>
<td>0.972</td>
<td>0.972</td>
<td>0.972</td>
</tr>
<tr>
<td>AUPRC</td>
<td>0.985</td>
<td>0.986</td>
<td>0.986</td>
<td>0.987</td>
<td>0.985</td>
</tr>
<tr>
<td>AUROC</td>
<td>0.989</td>
<td>0.989</td>
<td>0.99</td>
<td>0.99</td>
<td>0.989</td>
</tr>
</tbody>
</table>
Figure 4.7: ROC curve showing the performance of all features from MitoPredictor and SubCons in the Random Forest classifier.
CHAPTER 5. GENERAL CONCLUSION

Mitochondria are involved in diverse cellular processes, like energy production (Hatefi, 1985), Fe/S cluster biosynthesis (Stehling and Lill, 2013), amino-acid metabolism (King, 2007), apoptosis (Wang and Youle, 2009), and cellular signaling (Tait and Green, 2012). Proteins responsible for these processes are contributed by the mitochondrial genome and the nuclear genome. Interestingly, the only mitochondrial function that needs proteins encoded in the mitochondrial genome is oxidative phosphorylation (Race et al., 1999). The vast majority of mitochondrial processes rely solely on proteins contributed by the nuclear genome, i.e. the mitochondrial proteome. Some of these processes, like Fe/S cluster biosynthesis, are well-conserved in animals. Others represent changes to the mitochondrial functional repertoire in specific animal lineages, like the gain of innate-immunity in vertebrates (West et al., 2011), or the loss of arginine biosynthesis in ecdysozoans (Payne and Loomis, 2006). Therefore, we can learn about the conservation, loss and gain of mitochondrial functions in animals, by performing a comparative analysis of animal mitochondrial proteomes.

For such an analysis, we can compare the existing experimentally-characterized mitochondrial proteomes in animals. In animals, mitochondrial proteomes have been experimentally-characterized in multiple species-human, mouse, Caenorhabditis elegans and Drosophila melanogaster. However, all four animals represent only one of the five major lineages of animal phylogeny- Bilateria. Thus, our current understanding of the animal mitochondrial proteome excludes the remaining four major lineages of animals- Phyla Porifera, Cnidaria, Ctenophora, and Placozoa- referred to as “non-bilaterian animals”. Thus, to obtain a broad picture of animal mitochondrial proteome evolution, one practical solution is to use computational techniques to predict mitochondrial proteins in animals lacking such experimental data. In this dissertation, we carry out a comparative analysis of experimentally-characterized mitochondrial proteomes in bilaterian animals (Chapter 3), and use computational techniques to predict and characterize the mitochondrial proteomes of non-bilaterian
animals (Chapter 2) (Muthye and Lavrov, 2018). Additionally, to facilitate comparative analysis of mitochondrial proteomes in animals, we developed two tools: 1] the Metazoan Mitochondrial Proteome Database (MMPdb) and 2] MitoPredictor, a novel machine-learning tool to predict mitochondrial proteins in animals (Chapter 4).

In Chapter 2, we present the first comparative analysis of the inferred mitochondrial proteomes of non-bilaterian animals (Muthye and Lavrov, 2018). We use two bioinformatic techniques: ortholog-detection and mitochondrial targeting signal (MTS)$^1$-prediction, to predict mitochondrial proteomes of animals from all the four non-bilaterian phyla. We observe a large variation in the size and content of the inferred mitochondrial proteomes from the non-bilaterian animals; there is nearly a 5X difference between the largest and smallest inferred mitochondrial proteome. Much of this variation is due to the number of proteins predicted by the presence of MTS alone; there is a 16X difference between the highest and lowest number of MTS-alone proteins in the non-bilaterian species. Some of the proteins predicted by the MTS-approach alone potentially represent mitochondrial functions specific to non-bilaterian animals. For instance, several non-bilaterian species possess a sulfide-resistant alternative oxidase, (AOX)$^2$, which provides an alternate respiratory pathway and is known to be lost in vertebrates. While the role of AOX in plants is well-studied (Fernie et al., 2004; Watling et al., 2006; Onda et al., 2008; Saha et al., 2016), the function of AOX in animals is much less understood. It has been proposed that AOX helps marine animals acclimatize to changing oxygen-levels, or any conditions which may impair accurate assembly of the canonical respiratory complexes (McDonald and Gospodaryov, 2018). We also identify potential instances of mitochondrial neolocalization in non-bilaterian animals, i.e. non-mitochondrial proteins in mammals with MTS in several non-bilaterian species, like histones and cytosolic ribosomal proteins. Conversely, we find that nearly 2.5% of the human mitochondrial proteome lacks orthologs in any non-bilaterian species, representing potential bilaterian mitochondrial innovations.

Thus, the exclusion of mitochondrial proteomes from non-bilaterian species would result in underestimating the diversity of the mitochondrial proteome in animals and prevent us from obtaining

---

$^1$MTS: Mitochondrial Targeting Signal  
$^2$AOX: alternate oxidase
a complete image of the evolution of the mitochondrial proteome in Metazoa. Our results reinforce the need for experimental-characterization of mitochondrial proteomes from non-bilaterian animals. Such analyses would help identify some of the potential “hidden diversity” in animal mitochondrial proteomes. We propose two interesting candidates for such an exercise- the calcarean sponge *Sycon ciliatum* and the myxozoan *Kudoa iwatai*. *S. ciliatum* has the largest number of proteins with a predicted MTS, but no orthologs to any human or yeast protein (746). Some of these would be false-positives resulting from MTS prediction. However, experimental techniques like Mass-spectrometry and GFP-analysis, would help identify novel mitochondrial proteins in this species. Conversely, the obligate endoparasite *K. iwatai* has the smallest inferred mitochondrial proteome in animals (just 26% the size of the human mitochondrial proteome). Experimental analysis of mitochondrial proteins from this animal would help understand the effect of parasitism on mitochondrial function.

The lack of experimentally-characterized mitochondrial proteomes from any non-bilaterian animal prevented us from further exploring the causes and functional implications of size-variation of the animal mitochondrial proteome. In **Chapter 3**, we answer those questions, by carrying out a comparative analysis of mitochondrial proteomes of four vertebrate and invertebrate model organisms: human, mouse, *C. elegans* and *D. melanogaster*. We find that the bilaterian animal mitochondrial proteome is a dynamic entity, with a small core of conserved animal mitochondrial proteins, and a large number of lineage-specific gains and losses. We report that several factors are responsible for the observed variation in size and content of the animal mitochondrial proteomes, of which the loss of conserved animal mitochondrial proteins in ecdysozoans and the gain of novel mitochondrial proteins in mammals were the main causes. We find that both morphological and ecological factors seem to affect the composition of the mitochondrial proteome. For instance, in mammals, an increase in organismal complexity resulted in tissue-specific OXPHOS subunits and also potentially necessitated an increased role of mitochondria in cellular-signaling. Conversely, both invertebrates are known to have experienced drastic gene-loss. We find that several conserved mitochondrial proteins were lost in both invertebrates, like those involved in arginine biosynthesis.
Interestingly, while nearly 1/5th of each animal mitochondrial proteome underwent mitochondrial neolocalization, majority of these proteins did so without acquiring a canonical MTS. In fact, while the majority of the well-conserved animal mitochondrial proteins possess an MTS, most of the neolocalized and novel mitochondrial proteins do not. These results indicate the importance of alternate mitochondrial targeting signals and/or pathways in the import of neolocalized and novel animal mitochondrial proteins. Additionally, our results have important implications for the use of MTS-prediction in identifying mitochondrial proteins. Prediction of MTS is a widely-used method for identifying novel mitochondrial proteins. Thus, an MTS-only approach would potentially miss several novel mitochondrial proteins, which is the goal of MTS-approach in the first place.

Our results from chapters 2 and 3 demonstrate the importance and utility of carrying out comparative analysis of mitochondrial proteomes in animals. However, researchers attempting to do so encounter two major challenges: 1) the data on the experimentally-characterized mitochondrial proteomes of animals are scattered across multiple databases, and 2) mitochondrial proteomes have not been experimentally-characterized for most animal phyla. Chapter 4 presents two tools we developed to address these challenges: 1) the Metazoan Mitochondrial Proteome Database, a database where we consolidate information on the animal mitochondrial proteomes, and 2) MitoPredictor, a novel machine-learning tool for prediction of mitochondrial proteins in animals.

No one single database houses data on the mitochondrial proteomes of vertebrate and invertebrate animals. The goal of MMPdb is to provide a user-friendly database for visual exploration and comparative analysis of existing animal mitochondrial proteomes. MMPdb facilitates comparative analysis of animal mitochondrial proteomes by integrating results from orthology analysis, MTS prediction, protein domain annotation and GO annotation. For mammalian proteins, evidence of mitochondrial localization is also added from the MitoMiner v4.0 database (Smith and Robinson, 2018). Additionally, for human proteins, tissue-specific data is included from the Human Protein Atlas (Uhlen et al., 2010). An R Shiny applet allows for visualization and analysis of the data contained in MMPdb. The applet provides an easy-to-use interface for researchers not comfortable working on the command-line. Thus, users can focus on asking interesting questions regarding
animal mitochondrial proteome evolution rather than spending most of their time consolidating subcellular localization information for animal mitochondrial proteins.

While MMPdb allows for analysis of existing data on animal mitochondrial proteomes, the majority of animal species lack an experimentally-characterized mitochondrial proteomes. A practical solution for addressing this challenge is to use bioinformatic tools to predict mitochondrial proteomes for species lacking such data. For this, we developed a novel machine-learning tool- MitoPredictor, a Random Forest classifier for prediction of mitochondrial proteins in animals. MitoPredictor, is an ensemble predictor, which uses three sources of information for prediction of mitochondrial proteins: orthology-predictions by Proteinortho Lechner et al. (2011), MTS predictions by TargetP (Emanuelsson et al., 2007) and MitoFates (Fukasawa et al., 2015) and protein domain content.

We show that MitoPredictor outperforms SubCons (Salvatore et al., 2017), a widely-used ensemble predictor of subcellular localization, as well as the two MTS-predictors: TargetP and MitoFates. MitoPredictor is trained on the most up-to-date subcellular localization data for proteins from both vertebrate and invertebrate model organisms, whereas SubCons is trained exclusively on human mitochondrial proteins. MitoPredictor is more user-friendly than SubCons, and requires far fewer tools to be downloaded compared to SubCons. We find that orthology and protein domain content play an important role in predicting mitochondrial proteins in our Random Forest model. The output from MitoPredictor is aimed to facilitate comparative analysis of mitochondrial proteomes from the four reference animals (human, mouse, C. elegans and D. melanogaster) and the predicted mitochondrial proteome. The output from MitoPredictor is divided into two parts: 1] an easy-to-use R Shiny applet, which contains orthology information, MTS-prediction and protein domain content of the predicted mitochondrial proteome, and 2] a folder, called “Stats”, including basic information regarding the predicted mitochondrial proteome. The Stats folder contains lists of reference animal proteins, both with and without orthologs in the query mitochondrial proteome. These lists of proteins can be used directly as input files for tools like PantherDB (Mi et al., 2016) and ConsensusPathDB (Kamburov et al., 2010), to analyze conservation and loss of mitochondrial proteins in the query mitochondrial proteome. Additionally, MitoPredictor also outputs a list of
all query proteins which are predicted to possess an MTS by both TargetP and MitoFates but have no orthologs in the reference animal proteomes. The output includes a list of protein domains contained with each of these potential species-specific proteins, which can be used in software like dcGO (Fang and Gough, 2012) for functional annotation.

Future directions

In this dissertation, we provide a broad picture of the evolution of mitochondrial proteomes in animals by carrying out comparative analysis of mitochondrial proteomes from bilaterian and non-bilaterian animals. One major limitation in such an analysis is the lack of experimental data from the majority of animal lineages. Currently, such information is limited to only three animal phyla (Chordata, Nematoda, and Arthropoda). Even well-studied bilaterian model species, like Danio rerio (zebrafish), lack a complete experimentally-characterized mitochondrial proteome. The experimental-characterization of mitochondrial proteomes from multiple phyla would greatly improve our understanding of the evolution of mitochondrial function in animals. Additionally, analysis of N-terminii of mitochondrial proteins from multiple animal species can improve our understanding of mitochondrial targeting signals, and help researchers develop better tools for predicting mitochondrial proteins.

5.1 References


