Evolution of Mitochondrion-related Organelles in Metamonada

January 2019

Keitaro KUME

Evolution of Mitochondrion-related Organelles in Metamonada

A Dissertation Submitted to

the Graduate School of Life and Environmental Sciences,

the University of Tsukuba

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in Science

(Doctoral Program in Biological Sciences)

Keitaro KUME

Abstract

Evolution of mitochondria is one of most intensively studied areas of biology. Mitochondria are organelles found in nearly all of eukaryotes, and arose through an endosymbiotic event in which ancestral eukaryotes engulfed a bacterium closely related to extant α -proteobacteria. For the last two decades, unusual and divergent mitochondria have been reported from various lineages. Their organelles were shown to be of mitochondrial origin, but they lacked aerobic respiration functions and/or their own genomes. These reductive organelles are currently known as mitochondrion-related organelles (MROs). Metamonada, a major clade of eukaryotes, is an important lineage for studying the reductive evolution of mitochondria, because nearly all metamonads appear to contain MROs rather than typical mitochondria and their phylogenetic relationships have been clearly resolved. Additionally, the MROs of metamonad parasites, including *Trichomonas vaginalis*, *Giardia intestinalis*, and *Spironucleus salmonicida* have been well studied in the proteomics level.

Although various functions are known in typical mitochondria, previous studies of the evolution of MROs in metamonads have mainly focused on functions related to energy metabolism. In the first comparative transcriptome study, contamination of bacterial sequences in the metamonad transcriptome data sometimes interfered with correctly annotating mitochondria/MRO-related proteins, as mitochondria have a bacterial origin. In addition, absence of reliable methods for predicting whether a given protein is a mitochondrial/MRO protein or not was critical issue when discussing what the functions are the mitochondrion/MRO really has.

List of abbreviations

Abbreviations (Abb.)	Descriptions
ATP	Adenosine triphosphate
AUC	Area under the curve
CLO	Carpediemonas-like organism
DNA	Deoxyribonucleic acid
EST	Expressed sequence tag
GBM	Gradient boosting machine
HMM	hidden Markov model
ISC	Iron sulfur cluster
LGT	Lateral gene transfer
MRO	Mitochondrion-related organelle
Mt	Mitochondria
NGS	Next generation sequencing
RNA	Ribonucleic acid
ROC	Receiver operating characteristics
SVM	Support vector machine

AAT	Asparate aminotransferase
ACS	Acethyl CoA Synthase
AD	Arginine deminase
AK	Adenylate kinase
ALT	Alanine aminotransferase
ASCT	Acetate:succinate CoA transferase
Cpn60	Heat shock protein 60
CS	Cysteine syntase
Fdx	Ferredoxin
FDXR	Adrenodoxin-NADP+ reductase
Fe Hase	Fe only hydrogenase
Fxn	Frataxin
GCS H	Glycine cleavage system H-protein
GCS L	Glycine cleavage system L-Protein
GCS P	Glycine cleavage system P-protein
GCS T	Glycine cleavage system T-protein

Abb.	Descriptions
GDH	Glutamate dehydrogenase
GrpE	molecular chaperone GrpE
Grx5	Glutaredoxin-5
HCP	Hybrid cluster protein
HscB	molecular chaperone HscB
Hsp70	heat shock protein 70
HydE	Fe-hydrogenase Maturase E
HydF	Fe-hydrogenase Maturase F
HydG	Fe-hydrogenase Maturase G
IscA	Iron-sulfur cluster assembly
IscS	Cysteine desulfurase IscS
IscU	Iron-sulfur cluster assembly enzyme ISCU
MDH	Malate dehydrogenase
ME	Malic enzyme
MGL	Methionine gammma lyase
MPP	Mitochondrial processing peptidase
NFU	Iron-sulfur cluster scaffold protein NFU
NuoE	NADH:ubiquinone dehydrogenase E
NuoF	NADH:ubiquinone dehydrogenase F
OCDA	Ornithine cyclodeaminase
OCT	Ornithine carbamoyltransferase
PFO	Pyruvate:ferredoxin oxidoreductase
PSAT	Phosphoserine aminotransferase
Rbr	Rubrerythrin
Rxn	Rubredoxin
SCS	Succinyl coenzyme A (CoA) synthetase
SHMT	Serine hydroxymethyltransferase
SOD	Superoxide dismutase
Tnase	Tryptophanase
Trx	Thioredoxin
TrxP	Thioredoxin Peroxidase
TrxR	Thioledoxin reductase

Chapter 1. General introduction

1.1. Origin of mitochondria and its roles

Nearly all eukaryotes possess mitochondria, which are the double-membraned organelles. Although the endosymbiotic origin of mitochondria remains controversial (Andersson et al., 1998; Embley & Martin, 2006, Williams et al., 2007; Brindefalk et al., 2011; Gray, 2012; Pittis & Gabaldón, 2016; Martin et al., 2017; Martijn et al., 2018), it is widely accepted that a bacterium closely related to extant α-proteobacteria was engulfed by an ancestral eukaryotic host, giving rise to mitochondria. These organelles are responsible for various essential processes in the eukaryotic cells, such as aerobic energy metabolism, iron sulfur clusters (ISC) (Fe-S clusters) assembly, fatty acid metabolism, molecular chaperone system, anti-oxidant system, amino acid metabolism and apoptosis.

Mitochondria of extant species contain their own genome (mtDNA) typically composed of less than 100 genes coding for proteins and RNAs that function in mitochondria. Because most of the genes encoding mitochondrial proteins related to these biological processes are mainly located in the nuclear genome, these proteins must be translocated into the mitochondria by protein sorting and transport systems that recognize the mitochondrial targeting signal typically found in their N-terminal regions. This selective transport causes the concentration of mitochondrial proteins be in high concentration inside the mitochondria and maintains the efficiency of various enzymatic reactions.

The acquisition of mitochondria enabled aerobic respiration with high throughput energy production in ancestral eukaryotes, leading to the prosperity of eukaryotes. However, the ability of aerobic respiration is one of the benefits by mitochondria. In addition to aerobic energy metabolism mitochondria have been playing other essential roles in the process of eukaryotic cellular evolution through the above mentioned functions derived originally from α-proteobacteria.

1.2. Previous research for mitochondrial/mitochondrion-related organelle proteins in Metamonada

During the evolutionary process of eukaryotes, mitochondria have diverged extensively. Hydrogenosomes in *Trichomonas vaginalis* and mitosomes in *Giardia intestinalis* are typical examples of highly divergent mitochondria (Morrison et al., 2007; Jedelský et al., 2011; Schneider et al., 2011). These organelles are of mitochondrial origin, but they lack their own genomes and most of nuclear genome-encoded mitochondrial proteins related to the respiratory chain. These reductive organelles are currently referred to as mitochondrion-related organelles (MROs). Recently it was proposed that mitochondria/MROs should be classified into five functional types (Müller et al., 2012): aerobic mitochondria (Class 1), anaerobic mitochondria (Class 2), H_2-producing mitochondria (Class 3), hydrogenosomes (Class 4), and mitosomes (Class 5). In general, MROs are involved in Class 4 or Class 5. Various types of MROs have been identified in phylogenetically independent lineages which grow in micro-aerobic and anaerobic environments, indicating that these organelles arose independently several times throughout eukaryotic evolution (Roger et al., 2017). Metamonada is a large assemblage of flagellates adapted to microaerophilic/anaerobic environments. The monophyly and branching order of metamonads was robustly resolved by a recent phylogenomic analysis (Leger et al., 2017) (Figure 1-1). Notably, nearly all metamonads appear to possess MROs rather than typical aerobic mitochondria, indicating that analyses of metamonads can provide valuable information regarding the evolutionary process of MROs. Particularly, MROs of the ancestral metamonad lineage may have exhibited functions typical of those in the mitochondria, while those of derived lineages may have diverged towards reducing functions such as those in the mitosome of *Giardia* (Leger et al., 2017).

1.2.1. Metamonada

Metamonada is a major clade in Excavata, a large taxonomic group of eukaryotes (Adl et al., 2018). Metamonada consist of microaerophilic or anaerobic flagellates with various lifestyles, such as heterotrophic free-living, commensal, or parasitic. There are three sub clades of Metamonada, Preaxostyla, Parabasalia and Fornicata, exist with Preaxostyla as an early branching clade.

1.2.2. Preaxostyla

Transcriptome analyses were performed on two protists in Preaxostyla, *Trimastix marina* and *Paratrimastix pyriformi*. The data revealed the presence of mitochondrial protein homologs related to the functions of amino acid metabolism and pyruvate metabolism, suggesting that their putative MROs have these functions. However, the presence or absence of other mitochondrion derived functions in the MROs could not be concluded and their functions still remain unclear (Leger et al., 2017; Zubáčová et al., 2013). Notably, *Monocercomonoides* sp. was the first eukaryote which was reported to have neither mitochondria nor MRO. Genome and transcriptome data of *Monocercomonoides* sp. are available in a public database (Karnkowska et al., 2016). Although MRO was not identified morphologically, the presence of mitochondrion-related chaperon proteins such as CPN60 (chaperonin 60) suggested the secondary absence of MROs in the line leading to *Monocercomonoides* sp. (Karnkowska et al., 2016).

1.2.3. Parabasalia

Genome and/or transcriptome sequence data were reported from hydrogenosome-containing parasites/commensals, *Tritrichomonas foetus*, *Trichomonas vaginalis* and *Pentatrichomonas hominis* in Parabasalia. More than four decades ago hydrogenosomes were discovered in *T. foetus* and *T. vaginalis,* but these were not recognized as MROs at that time (Lindmark & Müller 1973). Because *T. vaginalis* and *T. foetus* are important parasites in the medical or veterinary field, their biological characteristics and MRO features have been studied to a certain extent (Beltrán et al., 2013; Birkeland et al., 2010; Carlton et al., 2007; Franzén et al., 2009; Jedelský et al., 2011; Schneider et al., 2011). Based on biochemical, proteomic, genome and transcriptome analyses performed for both *T. foetus* and *T. vaginalis*, the trichomonad hydrogenosomes were shown to have lost their own genomes, parts of mitochondrial proteins and the ability to generate ATP by oxidative phosphorylation, whereas they possess Fe-S cluster assembly, amino acid metabolism and antioxidant systems (Schneider et al., 2011).

1.2.4. Fornicata

Fornicata consists of three taxonomic subgroups, diplomonads, retortamonads and *Carpediemonas*-like organisms (CLOs), but only diplomonads are monophyletic (Adl et al., 2018; Kolisko et al., 2010; Simpson 2003). Diplomonads include mammalian and fish parasites, such as *Giardia intestinalis*, *Spironucleus salmonicida*, *S. barkhanus*, *S. vortens,* and free-living flagellates classified to the genus *Trepomonas* or *Hexamita*. Morphological studies by electron microscopy showed that all of the fornicate organisms analyzed up to date do not contain typical mitochondria but do MROs:

previous studies have examined CLOs (Kolisko et al., 2010; Park et al., 2009; Yubuki et al., 2013; Yubuki et al., 2007), *G. intestinalis* (Tovar et al., 2003), *S. salmonicida* (Jerlström-Hultqvist et al., 2013), and *S. vortens* (Millet et al., 2013).

Genome and transcriptome analyses and proteomic analyses of MRO have been conducted for the human parasite *G. intestinalis* and a fish parasite *S. salmonicida*, indicating that these parasites possess highly derived MROs with reduced functions. Particularly, in the evolution leading to *Giardia*, the MRO (mitosome) lost most of its mitochondrial functions, and has only retained the function of the Fe-S cluster assembly (Jedelský et al., 2011; Morrison et al., 2007; Tovar et al., 2003).

Transcriptome data were reported for *Chilomastix cuspidata* and *Chilomatix caulleryi,* which are classified as retortamonads. Analyses of these data revealed that *C. cuspidata* MRO may function in amino acid metabolism and NADH reoxidation, while the MRO of *C. caulleryi*, a lumen-dwelling parasite, may have lost most of these functions during its evolution (Leger et al., 2017).

CLOs are a polyphyletic group and include *Carpediemonas membranifera*, *Ergobibamus cyprinoides*, *Aduncisulcus plauster*, *Kipferlia bialata* and *Dysnectes brevis*. Transcriptome analyses have been conducted for these organisms (Leger et al., 2017), and genomic analyses were performed for *K. bialata* (Tanifuji et al., 2018). Their MROs were shown to retain functions of at least amino acid metabolism, ATP synthesis, NADH reoxidation and H_2 production.

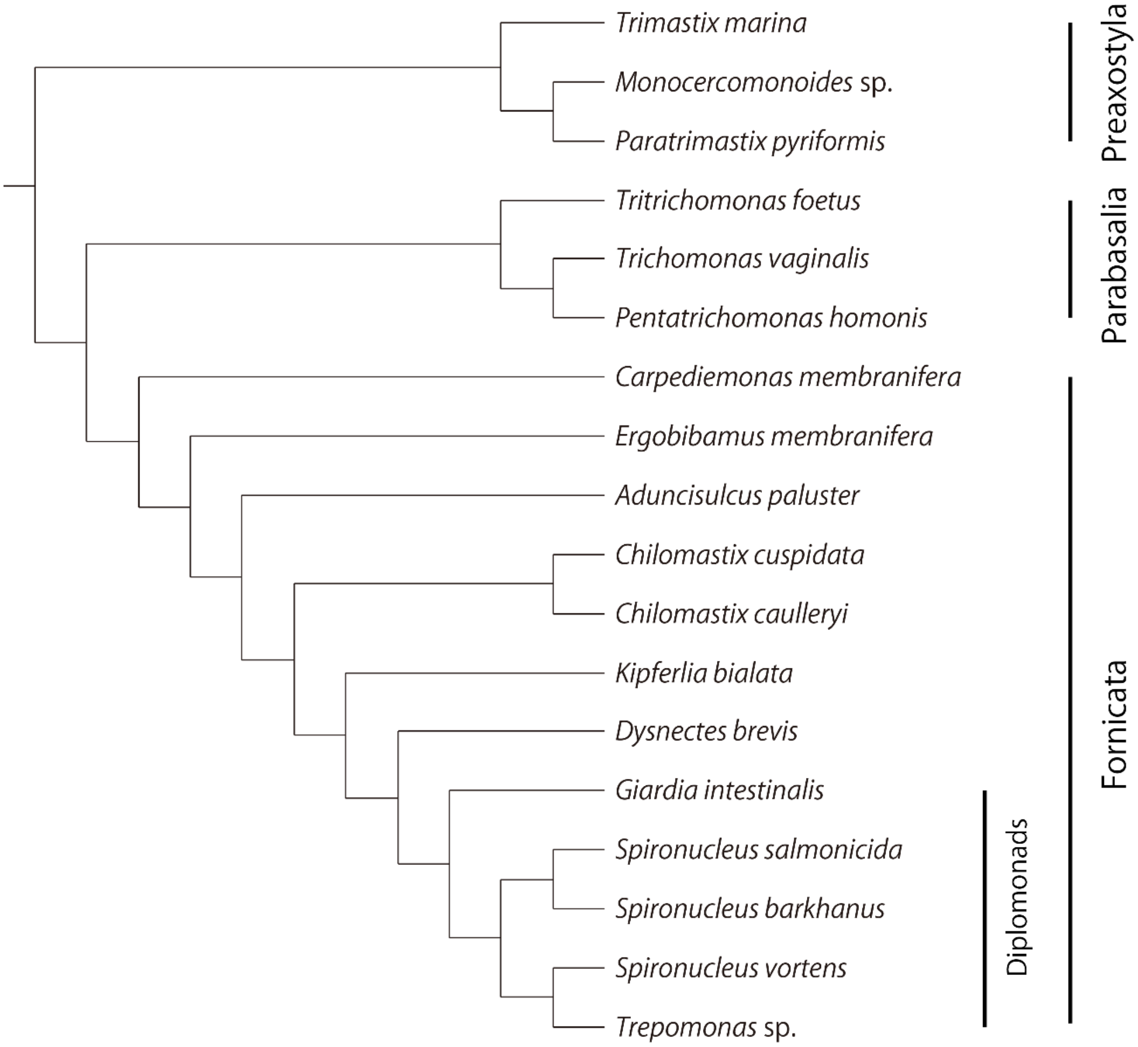
1.3. Application of bioinformatics method to the analysis of MRO proteome

While transcriptome analyses have been performed for many metamonads, proteome analyses of MRO have been conducted only for *T. foetus*, *T. vaginalis*, *G. intestinalis* and *S. salmonicida* (Table 1-1). Because parasites are important in the medical, veterinary and fishery fields, experimental procedures such as axenic culture, organelle purification, and biochemical analysis have been established. However, these methods have not been established yet for heterotrophic metamonads which must be cultured with bacterial feed. Developing these methods for a direct proteome analysis of MROs is very difficult because of the contamination with bacteria in the materials used for molecular analyses. Thus, it is necessary to distinguish mitochondrial/MRO proteins from other proteins using various bioinformatics methods.

Transcriptome data for heterotrophic metamonads such as *K. bialata* and *D. brevis* were generated from non-axenic cultures in a previous study by Leger et al. (2017). Bacterial contamination resulted in a low quality of assembly and a small amount of eukaryotic sequence data, preventing the detection of the presence or absence of each mitochondrial/MRO protein in the putative MRO proteome. To improve the quality of data, density gradient centrifugation was conducted to reduce bacterial contamination for the genome and transcriptome analyses of *K bialata*, resulting in the first report of a draft genome of heterotrophic metamonads (Tanifuji et al., 2018).

Most previous studies (Jedelský et al., 2011; Rada et al., 2011; Schneider et al., 2011) used prediction software for mitochondrial proteins such as TargetP (Emanuelsson et al., 2007), TPpred2 (Savojardo et al., 2014) and Mitofates (Fukasawa

217

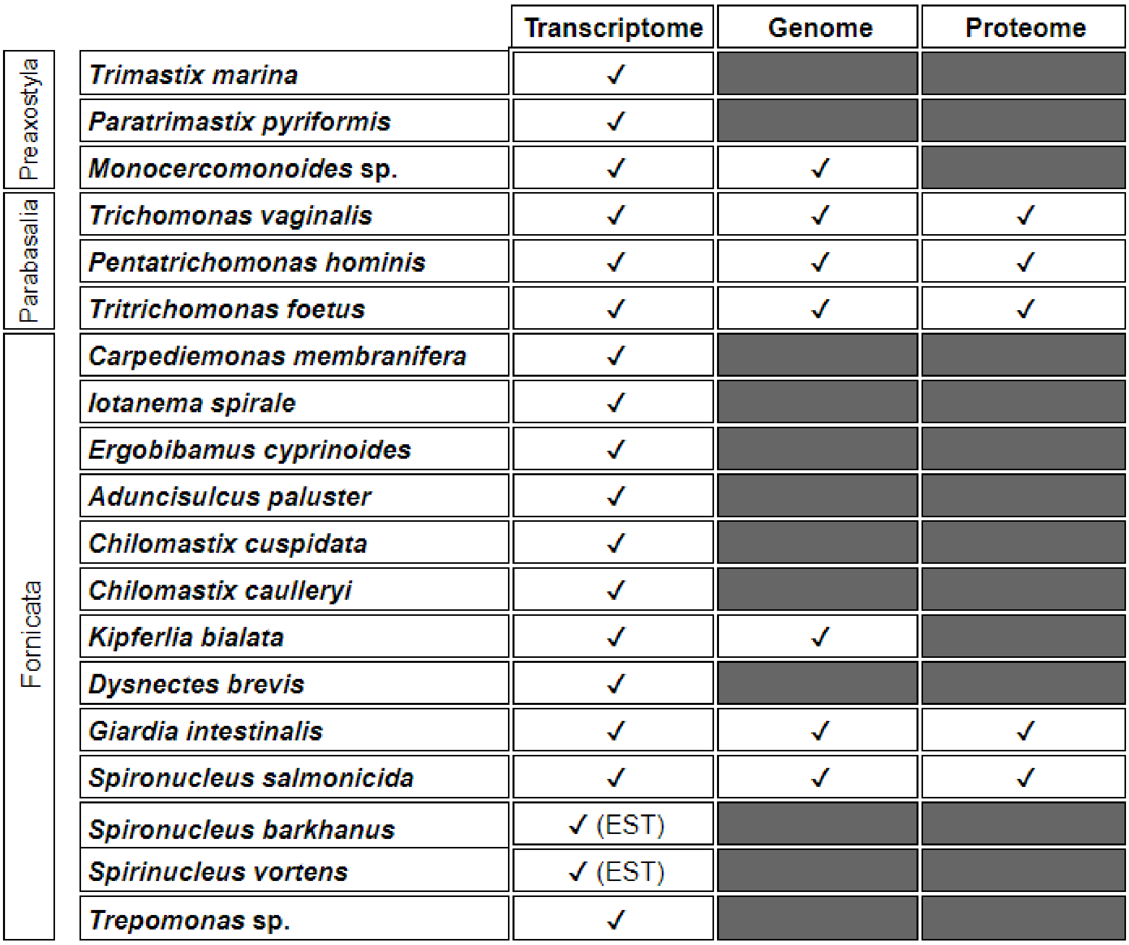


218

Figure 1-1: Phylogenetic tree of Metamonada (Leger et al., 2017 modified into cladogram tree; 159 proteins, 39,089 sites, 94 taxa, CAT-GTR + Γ model.)

219

Table 1-1: List of previous omics studies for metamonads. A check mark (✓) indicates that the corresponding omics analysis was performed, and its data are available. A grey cell indicates that the data are unavailable or do not exist.



		Transcriptome	Genome	Proteome
Preaxostyla	*Trimastix marina*	✓	■	■
	Paratrimastix pyriformis	✓	■	■
	Monocercomonoides sp.	✓	✓	■
Parabasalia	*Trichomonas vaginalis*	✓	✓	✓
	Pentatrichomonas hominis	✓	✓	✓
	Tritrichomonas foetus	✓	✓	✓
Fornicata	*Carpediemonas membranifera*	✓	■	■
	Iotanema spirale	✓	■	■
	Ergobibamus cyprinoides	✓	■	■
	Aduncisulcus paluster	✓	■	■
	Chilomastix cuspidata	✓	■	■
	Chilomastix caulleryi	✓	■	■
	Kipferlia bialata	✓	✓	■
	Dysnectes brevis	✓	■	■
	Giardia intestinalis	✓	✓	✓
	Spironucleus salmonicida	✓	✓	✓
	Spironucleus barkhanus	✓ (EST)	■	■
	Spirinucleus vortens	✓ (EST)	■	■
	Trepomonas sp.	✓	■	■

Chapter 3. NommPred: Prediction of Mitochondrial and Mitochondrion-related Organelle Proteins of Non-model Organisms

3.1. Introduction

Mitochondria are separated from other cellular components by a double membrane, resulting in the concentration of mitochondrial proteins inside the membrane. In general, the functions of an organelle are determined by the protein repertoire of the organelle. Therefore, the estimation of the function of mitochondria needs to determine the repertoire of mitochondrial proteins, most of which are nuclear encoded, expressed in cytosol, and finally transported into mitochondria (Gonczarowska-Jorge et al., 2017).

To determine a repertoire of mitochondrial proteins, the proteomic analysis of mitochondria is essential. For model organisms, experimental methods for the proteomic analysis of mitochondria have already been established during their long research histories (Kumar et al., 2002; Sickmann et al., 2003; Reinders et al., 2006; Cherry et al., 2012; Chen et al., 2010); however, for non-model organisms, there are no general strategies for the proteomic analysis of mitochondria. Even in non-model organisms, information on the amino acid sequences of proteins is indirectly obtained from the nucleotide sequences of the genome or transcriptome analysis, and these are useful tools for studying the cellular and molecular biological research subjects of non-model organisms of which proteins are difficult to treat directly during experiments. Recently, high throughput sequencing, the so-called next-generation sequencing (NGS), has allowed us to easily obtain the entire genome or transcriptome data even from non-model organisms at a low cost and in a short time. Therefore, transcriptome analysis is performed for the entire cell extracts of non-model organisms including mitochondria and the other cellular components, and the mitochondrial proteins are predicted by using an amino acid sequence-based computational method instead of purifying mitochondria and determining the repertoire of mitochondrial proteins directly. Such a bioinformatics approach needs to discriminate mitochondrial proteins from all the proteins that are deduced from the entire cell transcriptome data.

A machine learning approach has been often used to classify mitochondrial/non-mitochondrial proteins. Various software programs based on machine learning are available; these programs predict whether an input protein sequence is a mitochondrial protein. For example, TPpred3 (Savojardo et al., 2015) and Mitofates (Fukasawa et al., 2015) are prediction software programs based on support vector machines, whereas TargetP (Emanuelsson et al., 2007) is a software program based on neural network techniques.

Most of the current prediction software programs, including TPpred3, Mitofates, and TargetP, are trained only with the data derived from model organisms, which belong to the taxonomic groups, Metazoa, Fungi, or Embryophyta, and these programs are designed for application to the proteins of model organisms and their relatives. Model organisms have been studied experimentally at an enormous cost because of their basic biological, medical, or industrial importance. This has resulted in the accumulation of vast biochemical experimental data of protein localization to cellular compartments including mitochondria.

On the other hand, in the case of non-model organisms, except for those that are closely related to the known model organisms, very few experimental data are available because of the shortage of basic experimental procedures although they exhibit most parts of the eukaryotic diversities (Adl et al., 2018). Hereafter, I refer to such non-model organisms that do not belong to Metazoa, Embryophyta, and Fungi as non-model organisms. Therefore, for the study of the mitochondrial proteins derived from non-model organisms, the sequence data of genome or transcriptome that are produced by using the NGS approach are mainly used to predict the proteins that would be mitochondrially localized. In general, the prediction tools designed for model organisms are usually applied for these analyses; however, these tools do not necessarily guarantee accuracy of prediction because the N-terminal sequence features important for the prediction of the mitochondrial proteins could be far divergent in non-model organisms compared to those of the model organisms. In particular, in the case of the prediction of MRO protein, the prediction tools currently available are highly inaccurate (Makiuchi & Nozaki, 2014). Therefore, in general, for predicting of mitochondrial/MRO proteins in non-model organisms, the consensus of the results from multiple predictors is considered to avoid false predictions. However, this cannot be validated.

To resolve this problem, here, I propose a software program, NommPred (non-model organismal mitochondrial/MRO protein predictor), which predicts the mitochondrial/MRO proteins derived from non-model organisms. To develop this software, I prepared a dataset including the mitochondrial or MRO proteins derived widely from non-model organisms and adopted a gradient boosting machine (GBM) (Friedman, 1999; Friedman et al., 2000; Friedman, 2002) as a classifier. GBM, which is one of the ensemble classifiers, was used instead of the support vector machine (Cortes & Vapnik, 1995), which was adopted in the previous predictors Mitofates and TPpred3. NommPred could resolve the problem due to the inconsistency between the origins of the training and input data when predicting the mitochondrial/MRO proteins of non-model organisms. The performance of NommPred was shown to be superior to Mitofates, which was demonstrated to be the best among the alternative methods, in predicting the mitochondrial/MRO proteins derived from non-model organisms. Therefore, NommPred is the best predictor for the mitochondrial/MRO proteins of non-model organisms.

3.2. Materials and Methods

3.2.1. Scheme of NommPred

A flowchart and a message sequence chart of the newly developed software, NommPred, are illustrated in Figures 3-1 and 3-2, respectively. The software takes as input both the protein sequence in FASTA format (Definition is available from: www.ncbi.nlm.nih.gov/books/NBK53702/) and organismal information from which the protein sequence is derived. The feature of each protein was extracted based on Mitofates' algorithm to create a 920-dimensional feature vector (Figure 3-1). The vector is subjected to the GBM predictor (Mit Predictor for mitochondrial proteins or MRO Predictor for MRO proteins as described below), and the predictor outputs the prediction results and probabilities.

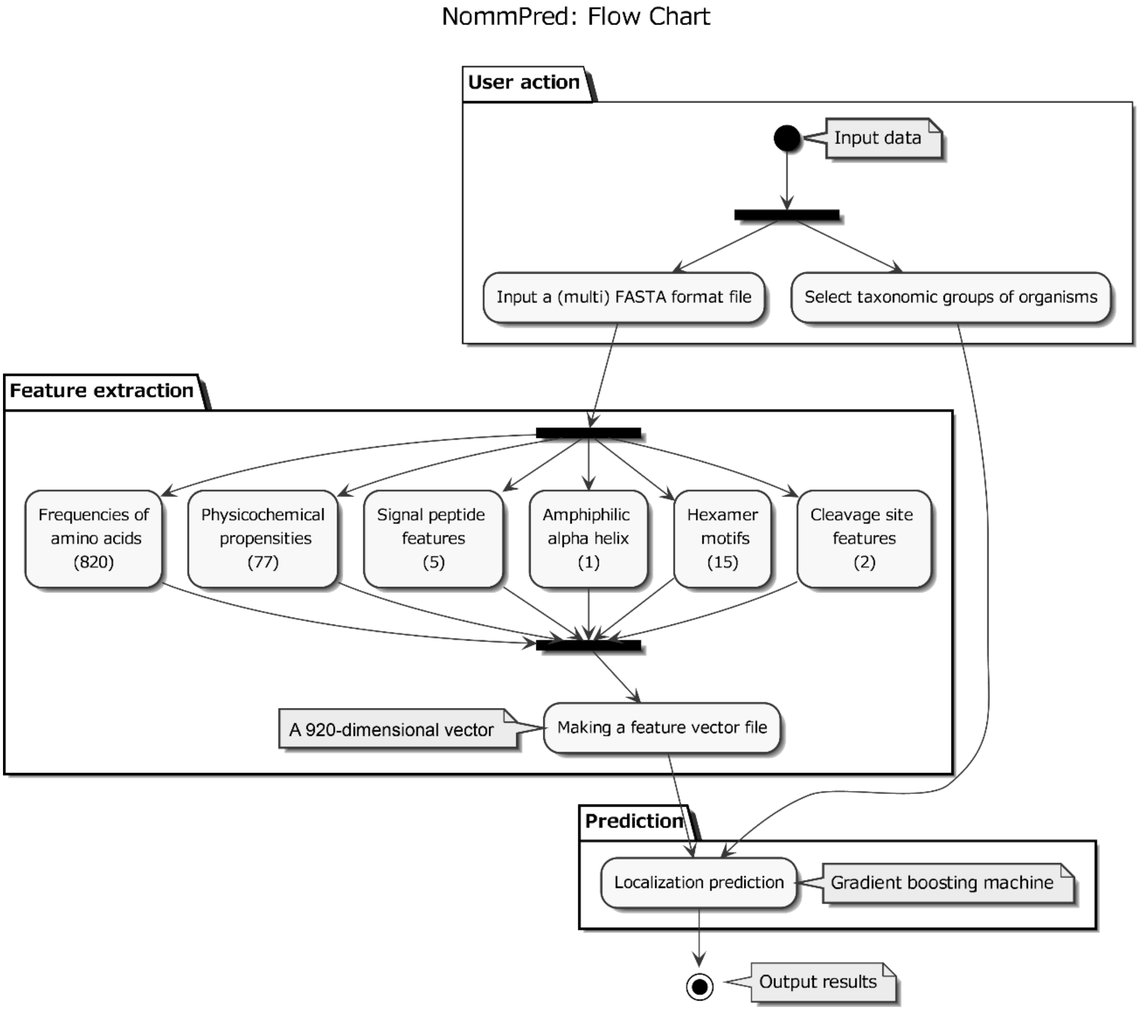

NommPred: Flow Chart
User action
Input data
Input a (multi) FASTA format file
Select taxonomic groups of organisms
Feature extraction
Frequencies of amino acids (820)
Physicochemical propensities (77)
Signal peptide features (5)
Amphiphilic alpha helix (1)
Hexamer motifs (15)
Cleavage site features (2)
A 920-dimensional vector
Making a feature vector file
Prediction
Localization prediction
Gradient boosting machine
Output results

Figure 3-1: Flowchart of NommPred. The closed circle represents the starting point of the program, and the closed circle surrounded by a larger open circle represents the endpoint. The user input data (Input data) include the protein sequence in FASTA format and information of the protein sequence origin (taxonomic group). The input data are classified into (the first black bar in User action step) protein sequence, which is used for feature extraction, and organismal information, which is used for the selection of an appropriate GBM Predictor: Mit Predictor, MRO Predictor, or others. In the feature extraction step, the 920 calculated features (Table 3-1) are integrated, and a 920-dimensional feature vector is obtained as the output. In the figure, only six feature categories are depicted with the number of individual features. This vector is subjected to a selected GBM Predictor as the input data, and then the prediction result is shown (Output Results).

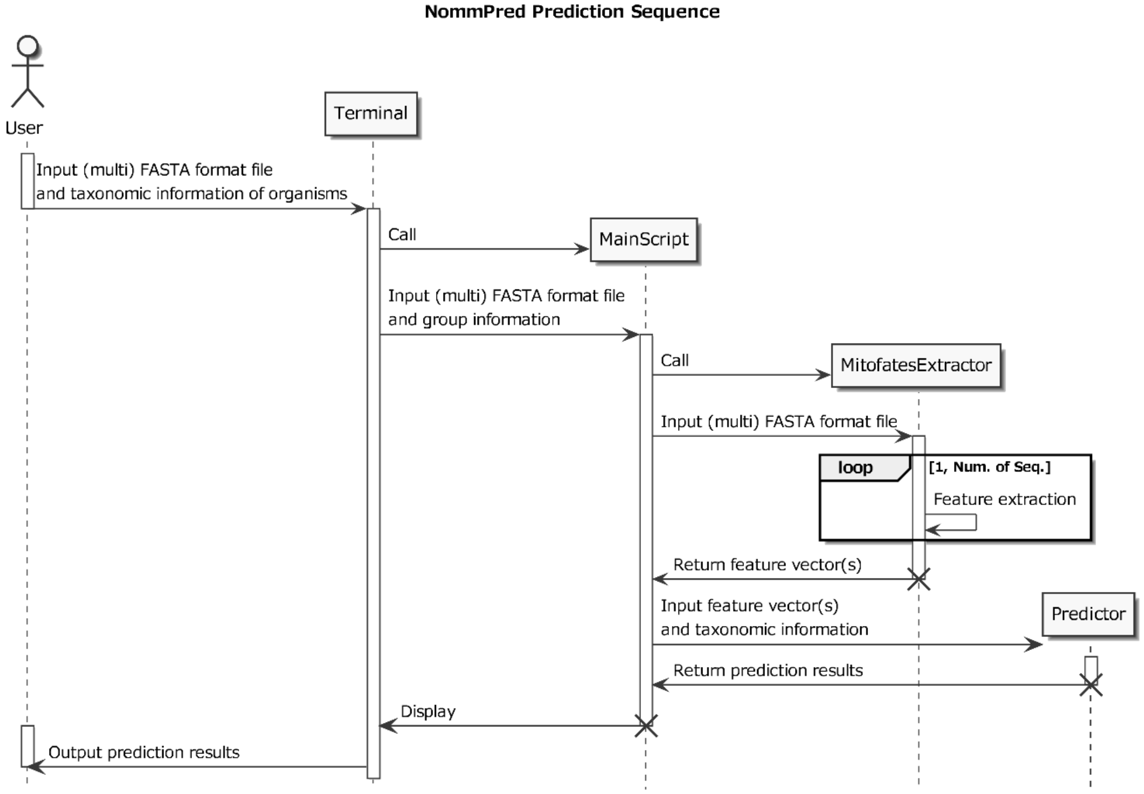
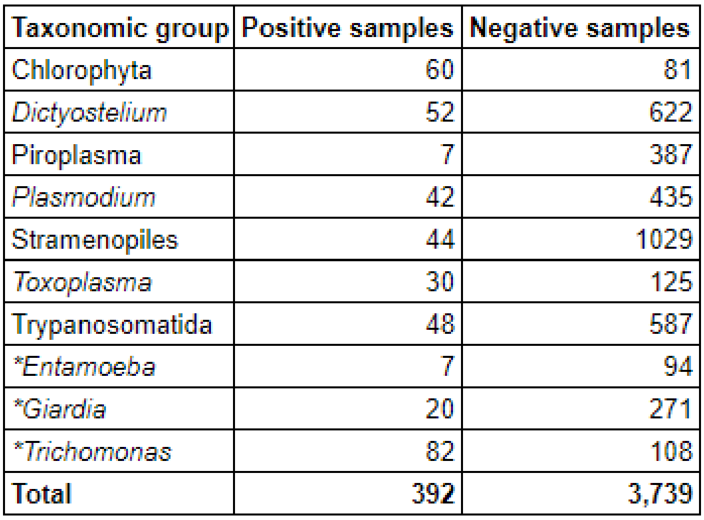

NommPred Prediction Sequence
User
Terminal
Input (multi) FASTA format file and taxonomic information of organisms
Call
MainScript
Input (multi) FASTA format file and group information
Call
MitofatesExtractor
Input (multi) FASTA format file
loop
[1, Num. of Seq.]
Feature extraction
Return feature vector(s)
Input feature vector(s) and taxonomic information
Predictor
Return prediction results
Display
Output prediction results

Figure 3-2: Message sequence chart of NommPred. The software for NommPred is console user interface (CUI), and it runs on the terminal. The software accepts a protein sequence file in multi FASTA format and a text file with information of the origins of the sequences and outputs the prediction results at last.

3.2.2. Dataset construction

The dataset used for the training and test is shown in Table 3-1. The mitochondrial or MRO proteins are treated as positive samples and the others as negative samples. The sequence data were obtained from UniProt (UniProt C, 2017; www.uniprot.org/), GiardiaDB (Hagen et al, 2011; Aurrecoechea et al, 2009; giardiadb.org/giardiadb/), TrichDB (Aurrecoechea et al, 2009; trichdb.org/trichdb/), and ApiLoc (Woodcroft et al, 2011; apiloc.biochem.unimelb.edu.au/apiloc/apiloc). Although these databases sometimes annotate mitochondrial or MRO proteins based on computational prediction, I used only those proteins whose localization was confirmed experimentally (e.g., Westernbloting, immunoblotting, or fluorescence microscope analysis) to mitochondria or MROs by investigating the literature. Then, I applied protein sequence redundancy reduction by using the BLASTClust program from the NCBI BLAST packages (Altschul et al, 1990). I adopted the criteria of being redundant at > 95% sequence identity. Finally, I prepared 392 positive mitochondrial or MRO protein sequences and 3,739 negative sequences. I classified the entire dataset into mitochondrial and MRO datasets, Mit and MRO. Then, I created a predictor for each dataset; one is the predictor for the mitochondrial protein trained with the mitochondrial proteins of 7 non-model organismal taxonomic groups (Mit Predictor), whereas the other is the predictor for the MRO protein trained with the MRO proteins of three non-model organismal taxonomic groups that possess MRO (groups marked with asterisks in Table 3-1) (MRO Predictor), because these two datasets were expected to be apparently different in the N-terminal sequence features of the mitochondrial/MRO protein sequences. The N-terminal sequence features of the MRO proteins are generally considered to be extremely divergent from those of the mitochondrial proteins.

Table 3-1: Entire dataset used for training and test. If the taxonomic group corresponds exactly to the genus, the name of the genus is represented in italic form. “Stramenopiles” is not a formal taxonomic rank but is generally used for the name of the group. The “Positive samples” column lists the number of sequences of the mitochondrial or MRO proteins. The “Negative samples” column lists the number of sequences of the non-mitochondrial or non-MRO proteins. The groups that possess MRO are represented with asterisks.



Taxonomic group	**Positive samples**	**Negative samples**
Chlorophyta	60	81
Dictyostelium	52	622
Piroplasma	7	387
Plasmodium	42	435
Stramenopiles	44	1029
Toxoplasma	30	125
Trypanosomatida	48	587
**Entamoeba*	7	94
**Giardia*	20	271
**Trichomonas*	82	108
Total	**392**	**3,739**

3.2.3. Feature extraction

For the extraction of features, I used the method described in Fukazawa et al. (2015). The feature of each protein was extracted to create a 920-dimensional feature vector. Extracted features and its details are shown in Table 3-2.

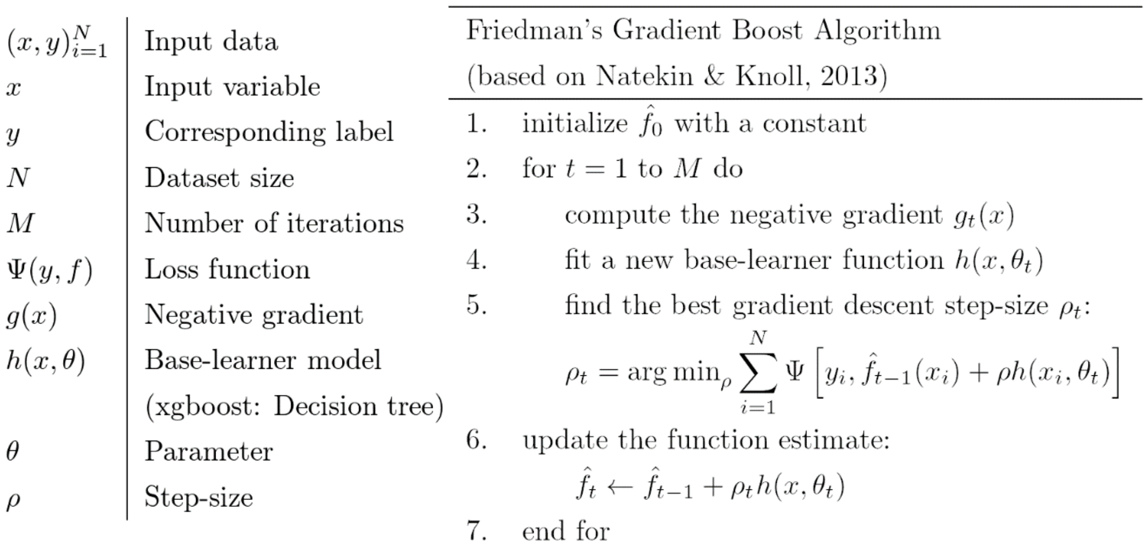
Table 3-2: List of features. For more details, refer to Mitofates (Fukasawa et al., 2015).

Feature category	Number of features	Features	Description
Frequencies of amino acids	820	Monopeptide: X_1 Dipeptide: X_1X_2 Skip-two dipeptides: X_1xxX_2	Normalized frequencies for each of the possible combinations of the standard 20 amino acids in 30 N-terminal residues of the input sequences. X_i (i = 1,2) and small x represents standard 20 amino acids, A C D E F G H I K L M N P Q R S T V W Y.
Physicochemical propensities	77	Segment scores Whole score	90 N-terminal residues of the input sequence are divided into six segments. Segment scores are calculated for each of the six segments. The total score is the sum of the segment scores. Each score is computed for: the mean of 1) hydrophobicity, 2) α-helical, or 3) β-strand periodicity, or the density of 4) positive charge, 5) negative charge, 6) serine, 7) threonine, 8) proline, 9) glycine, 10) amphiphilic, or 11) aromatic residues.
Signal peptide features	5	SP scores	Each score is computed in the putative signal peptide region defined by a sliding widow method search within 90 N-terminal residues for: the density of 1) positive charge, 2,3) two kinds of the density of negative charge, or the mean of 4) hydrophobicity, and 5) cleavage site preference residues.

Amphiphilic alpha helix	1	PA score: $\frac{\mu_H-\mu_C r\cos A}{n}$	For 30 N-terminal residues of the input sequence, the segments between 10 and 20 residues are generated by the sliding window method. For each segment, the score is computed from μ_H (magnitude of a hydrophobic moment vector) and μ_C (magnitude of a charge moment vector) by a formula as shown in the left column, and the best segment score is picked up as the PA score. *n* is the size of the window, *r* is the ratio parameter between μ_H and μ_C, and A is the angle between the two vectors.
Hexamer motifs	15	Motif scores: $-\log_{10}(p)$ Total score	Motifs are the 14 hexamer motifs that are significantly and frequently observed in the mitochondrial proteins compared to the non-mitochondrial ones ($p < 10^{-5}$.) The total score is the sum of each motif score.
Cleavage site features	2	Cleavage scores	For 100 N-terminal residues of the input sequence, 10 residue segments are generated by the sliding window method. For each segment, the cleavage score is calculated as the sum of position weighted matrix (PWM) scores for 10 residues, and the best and the second-best cleavage scores are picked up. The PWM is a given matrix in the Mitofates program.

3.2.4. Training and prediction method

I adopted GBM, one of the ensemble learning algorithms, and created predictors using xgboost (Chen & Guestrin, 2016) package in R (R Core Team, 2018) for the Mit and MRO datasets (Mit Predictor and MRO predictor). GBM reconstructs the unknown functional dependence $x \xrightarrow{f} y$ with estimate $\hat{f}(x)$; x is the explanatory input variables, y is the corresponding label. The scheme of the algorithm is shown in Figure 3-3 (based on Natekin & Knoll, 2013). Xgboost choices decision tree as the base-learner.



$(x, y)_{i=1}^{N}$	Input data
x	Input variable
y	Corresponding label
N	Dataset size
M	Number of iterations
$\Psi(y, f)$	Loss function
$g(x)$	Negative gradient
$h(x, \theta)$	Base-learner model (xgboost: Decision tree)
θ	Parameter
ρ	Step-size

```
Friedman's Gradient Boost Algorithm
(based on Natekin & Knoll, 2013)
1.  initialize $\hat{f}_0$ with a constant
2.  for $t = 1$ to $M$ do
3.      compute the negative gradient $g_t(x)$
4.      fit a new base-learner function $h(x, \theta_t)$
5.      find the best gradient descent step-size $\rho_t$:
            $\rho_t = \arg\min_\rho \sum_{i=1}^{N} \Psi\left[y_i, \hat{f}_{t-1}(x_i) + \rho h(x_i, \theta_t)\right]$
6.  update the function estimate:
        $\hat{f}_t \leftarrow \hat{f}_{t-1} + \rho_t h(x, \theta_t)$
7.  end for
```

Figure 3-3: The scheme of the algorithm of GBM.

I searched for optimal values of logical variables employed in the xgboost algorithm. Parameters for tree boosting, learning rate (*eta*), maximum depth of a tree (*max_depth*), minimum sum of instance weight (*min_child_weight*), maximum delta step (*max_delta_step*), and gamma were tuned with grid search, and finally I determined to set the default values for these variables. In addition, I optimized the parameter of the number of trees to the model by cross-validation. For other parameters, I used the default value.

3.2.5. Performance measures

To evaluate the performances of both the NommPred predictors—Mit Predictor and MRO Predictor—a receiver operating characteristics (ROC) curve and a ROC area under the curve (AUC) (Bradley, 1997) were used. In the R system, the ROC curve was drawn by plotting the true positive rate (y-axis) against the false positive rate (x-axis) for different cut-off values, and the ROC AUC was drawn based on the ROC curve.

To evaluate the robustness of the ROC AUC measures, I randomly divided the Mit or MRO dataset into three subsets (three-fold cross-validation), and I used two of them for the training data, and the other for the test data. This process was repeated 100 times (Figure 3-4).

To compare NommPred with a previous predictor, Mitofates, I used the same test data as that of NommPred for Mitofates to evaluate its performance. In this performance comparison, I carried out the paired *t* test and Wilcoxon signed rank test to evaluate the difference between the means of these 100 paired ROC AUC scores.

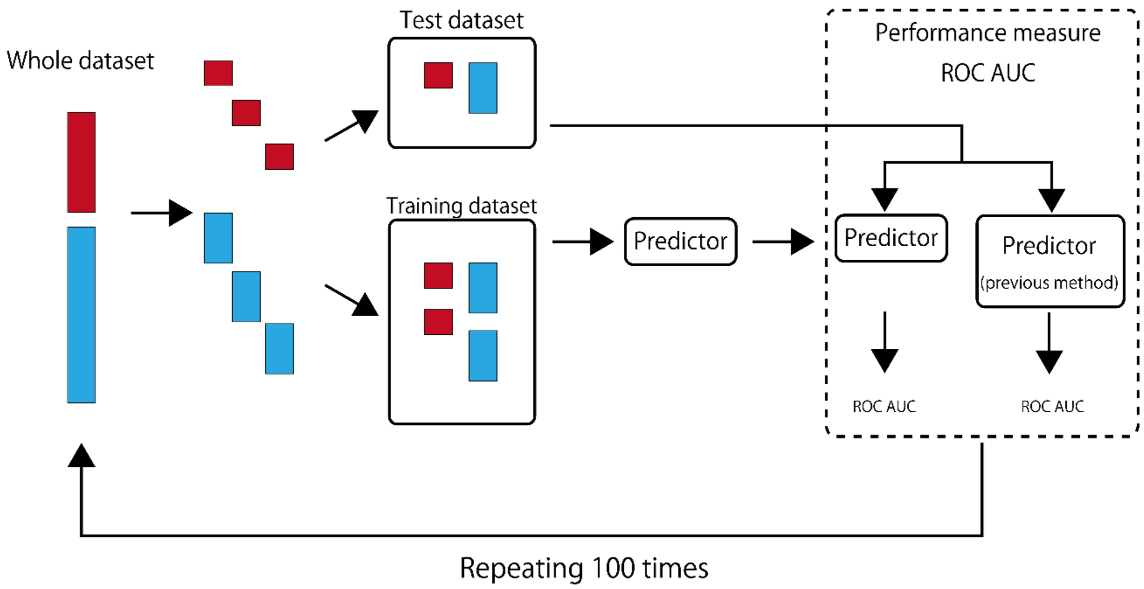


Figure 3-4: The scheme of the three-fold cross validation.

3.3. Results

3.3.1. Performance comparison analysis

Prediction of mitochondrial proteins

I carried out the performance comparison analysis between NommPred and a previous method, Mitofates. A dataset including the mitochondrial and non-mitochondrial proteins of seven non-model organismal taxonomic groups was used for the preparation of the training and test datasets (as described in the Materials and Methods section), resulting in the creation of Mit Predictor. Performance measure scores are listed in Table 3-3 and shown by boxplot in Figure 3-5.

Table 3-3: Comparison of the mean ROC AUC scores between NommPred and Mitofates. In NommPred mitochondrial proteins were predicted by Mit Predictor, while MRO proteins were by MRO Predictor. 100 randomly generated datasets (*n* = 100) of mitochondrial or MRO proteins were used for cross-validation (see Materials and Methods).

Prediction target	**ROC AUC**		***p value***	
	NommPred	**Mitofates**	**Paired *t* test**	**Wilcoxon signed rank test**
Mitochondrial protein	0.9463	0.9080	1.62E-42	0.00E+00
MRO protein	0.9041	0.8021	6.86E-40	0.00E+00

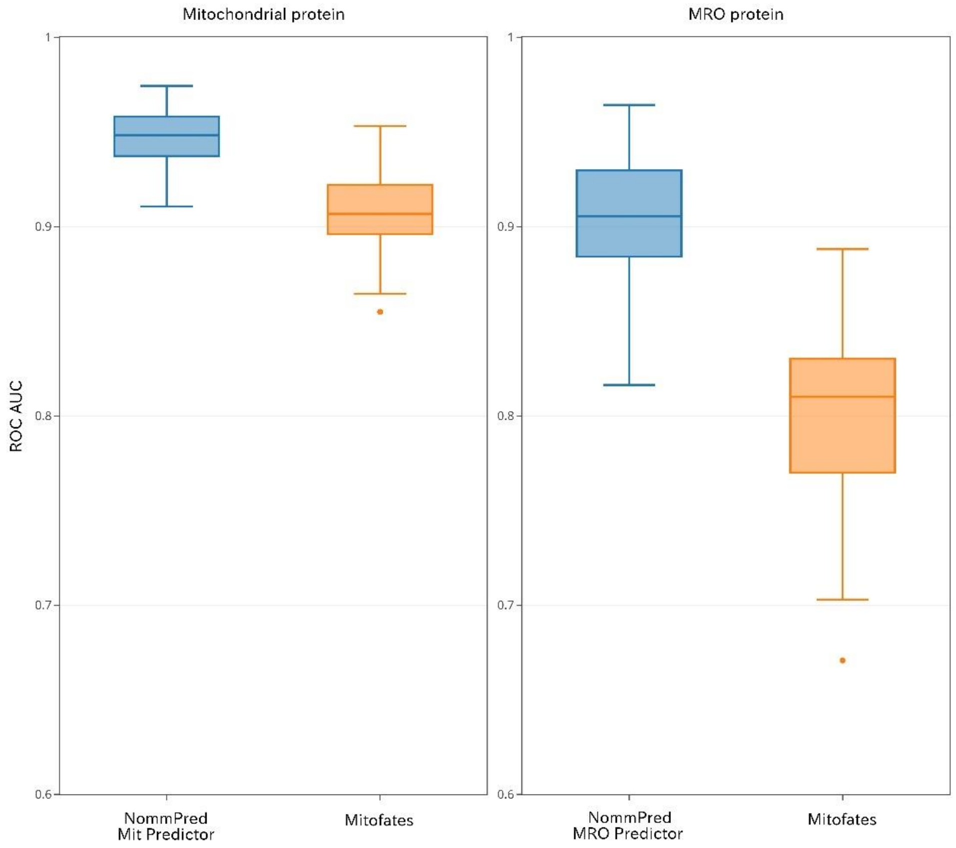


Figure 3-5: Boxplots showing the performance of the predictors of the mitochondrial and MRO proteins. The ROC AUC scores of 100 randomly generated datasets (y-axis) of the two predictors are plotted for, NommPred (Mit Predictor or/and MRO Predictor) in NommPred, and Mitofates (x-axis), are plotted. Lines within the boxplot indicate the median, the lower/higher quartile (Q1/Q3), and lower/higher whiskers.

For the mean ROC AUC scores (sample size *n* = 100), Mitofates achieved 0.9080, whereas the performance of Mit Predictor of NommPred was superior with a value of 0.9463 (Table 3-3). Moreover, the difference between the two mean ROC AUC scores was significant (paired *t* test: *p* value = 1.618×10^{-42}, Wilcoxon signed rank test: *p* value = ~ 0).

Generally, the ROC AUC score ranging between 0.5 and 0.7 is regarded as less accurate, between 0.7 and 0.9 as moderately accurate, and more than 0.9 as highly accurate (Fischer et al., 2003). Based on these criteria, Mitofates still showed sufficient accuracy in the prediction of the mitochondrial proteins derived from non-model organisms. However, for the prediction of those proteins, Mit Predictor with a higher ROC AUC score was preferred.

Prediction of MRO proteins

As described in the Materials and Methods section, I classified the entire dataset into two—Mit and MRO (Table 3-1). The MRO dataset including the MRO and non-MRO proteins of three non-model organismal groups was used for the preparation of the training and test datasets (described in the Materials and Methods section), resulting in the creation of MRO Predictor. I carried out a similar comparison analysis between the performance of MRO Predictor and that of Mitofates for the prediction of the MRO proteins. The performance measure scores are listed in Table 3-3.

Mitofates achieved a mean ROC AUC score (sample size *n* = 100) of 0.8021, whereas the performance of the MRO predictor of NommPred was far better with a mean value of 0.9041 (paired *t* test: *p* value = 6.855×10^{-40}, Wilcoxon signed rank test: *p* value = ~ 0) (Table 3-3). Based on these results, MRO Predictor of NommPred is suitable for the MRO proteins.

3.4. Discussions

I succeeded in developing NommPred, the predictors for the mitochondrial and MRO proteins derived from diverse non-model organisms, except for those belonging to Metazoa, Embryophyta and Fungi. Previously, the protein sequence data derived from non-model organisms were subjected to the predictor trained only by using the data from model organisms. NommPred could resolve the problem resulted from such inconsistency between the origins of the training data (model organisms) and the input data (non-model organisms).

3.4.1. Performance comparison analysis

The results of the statistical analysis (Table 3-3) clearly supported the superiority of NommPred in the performance of predicting the mitochondrial proteins of non-model organisms when compared to the existing best method, Mitofates. In particular, NommPred is the first software that is expected to be used for predicting the MRO proteins. NommPred would be useful for the prediction of metabolic pathways relating to the mitochondria/MROs from non-model organisms, the NGS data of which can be available. Since there is no other predictor suitable for the prediction of MRO proteins, MRO predictor in NommPred is useful tool to search for putative MRO proteins.

In this study, I retrieved almost all protein sequence data whose cellular localization were experimentally verified to mitochondria/MROs from various sequence databases. However, the origins of the sequence data of mitochondrial/MRO proteins in the entire dataset (Table 3-1) are biased for those of the parasitic organisms. Therefore, taxon sampling of our dataset is still very sparse. The accumulation of more data of the mitochondrial/MRO proteins of non-model organisms, especially from the free-living ones whose localization was confirmed experimentally, is essential to further improve the predictors presented in this work. I should continuously make efforts toward updating the dataset to provide more accurate predictors. Although NommPred may still have some problems that need to be improved in the future, I hope it will be helpful for the prediction of the mitochondrial/MRO proteins of non-model organisms.

Acknowledgements

I would like to express the deepest appreciation to my supervisor, Professor Tetsuo Hashimoto (University of Tsukuba) and Professor Yuji Inagaki (University of Tsukuba) for their support and encouragement during my PhD program period. I am deeply grateful to Professor Toshiyuki Amagasa (University of Tsukuba) and Professor Hiroyuki Kitagawa (University of Tsukuba) for accepting me as a student of master’s program in computer science using Graduate School Dual Degree Program of university of Tsukuba, and contributing to develop my prediction software. Assistant Professor Ryoma Kamikawa (Kyoto University) for training in the earlier stage on my career. Doctor Goro Tanifuji (National Museum of Nature and Science), Assistant Professor Takuro Nakayama (Tohoku University) and Doctor Yoshiyuki Ishitani gave me a lot of insight comments and suggestions.

Finally, I want to thank all the member of Molecular Evolution of Microbes (Hashimoto & Inagaki laboratory at University of Tsukuba) and Kitagawa Data Engineering (University of Tsukuba) for their help.

References

1. Adl SM, Bass D, Lane CE, et al. Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. J Eukaryot Microbiol. 2018. doi:10.1111/jeu.12691.
2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. A basic local alignment search tool. J Mol Biol. 1990;215:403–410. doi:10.1016/S0022-2836(05)80360-2.
3. Andersson SG, Zomorodipour A, Andersson JO, et al. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature. 1998;396(6707):133-40. doi:10.1038/24094.
4. Arthur D & Vassilvitskii S. k-means++: The advantages of careful seeding. Proceedings of the Eighteenth Annual ACM﹣SIAM Symposium on Discrete Algorithms. 2007;1027-35.
5. Aurrecoechea C, Brestelli J, Brunk BP, et al. GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*. Nucleic Acids Res. 2009;37(Database issue):D526-30. doi:10.1093/nar/gkn631.
6. Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 2011;27(12):1691-2. doi:10.1093/bioinformatics/btr174.
7. Beltrán NC, Horváthová L, Jedelský PL, et al. Iron-induced changes in the proteome of *Trichomonas vaginalis* hydrogenosomes. PLoS One. 2013;8(5):e65148. doi:10.1371/journal.pone.0065148.
8. Benson DA, Cavanaugh M, Clark K, et al. GenBank. Nucleic Acids Res. 2018;46(D1):D41-D47. doi:10.1093/nar/gkx1094.

9. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 1997;30(7):1145-59. doi:10.1016/S0031-3203(96)00142-2.

10. Brindefalk B, Ettema TJ, Viklund J, Thollesson M, Andersson SG. A phylometagenomic exploration of oceanic alphaproteobacteria reveals mitochondrial relatives unrelated to the SAR11 clade. PLoS One. 2011;6(9):e24457. doi:10.1371/journal.pone.0024457.

11. Camacho C, Coulouris G, Avagyan V, et al. Blast+: architecture and applications. BMC bioinformatics. 2009;10(1):421. doi:10.1186/1471-2105-10-421.

12. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972-3. doi:10.1093/bioinformatics/btp348.

13. Carlton JM, Hirt RP, Silva JC, et al. Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. Science. 2007;315(5809):207-12. doi:10.1126/science.1132894.

14. Chen T, Guestrin C. A scalable tree boosting system. eprint arXiv. 2016. doi:10.1145/2939672.2939785.

15. Chen X, Li J, Hou J, Xie Z, Yang F. Mammalian mitochondrial proteomics: insights into mitochondrial functions and mitochondria-related diseases. Expert Rev Proteomics. 2010;7(3):333-45. doi:10.1586/epr.10.22.

16. Cherry JM, Hong EL, Amundsen C, et al. *Saccharomyces* genome database: the genomics resource of budding yeast. Nucleic Acids Res. 2012;40(D1):D700-5. doi:10.1093/nar/gkr1029.

17. Cortes C & Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273-97. doi:10.1007/BF00994018.

18. Eddy SR. A New Generation of Homology Search Tools Based on Probabilistic Inference. Genome Inform. 2009;23(1):205-11. doi:10.1142/9781848165632_0019.

19. Emanuelsson O, Brunak S, von Heijne G, Nielsen H. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc. 2007;2(4):953-71. doi:10.1038/nprot.2007.131.

20. Embley TM & Martin W. Eukaryotic evolution, changes and challenges. Nature. 2006;440(7084):623-30. doi:10.1038/nature04546.

21. Fang YK, Chien KY, Huang KY, et al. Responding to a Zoonotic Emergency with Multi-omics Research: Pentatrichomonas hominis Hydrogenosomal Protein Characterization with Use of RNA Sequencing and Proteomics. OMICS. 2016;20(11):662-669. doi:10.1089/omi.2016.0111.

22. Finn RD, Bateman A, Clements J, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42(Database issue):D222-30. doi:10.1093/nar/gkt1223.

23. Fischer JE, Bachmann LM, Jaeschke R. A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. Intensive Care Med. 2003;29(7):1043-51. doi:10.1007/s00134-003-1761-8.

24. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting. Ann Statist. 2000;28(2):337-407. doi:10.1214/aos/1016218223.

25. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Statist. 1999;29(5):1189-232. doi:10.1214/aos/1013203451.

26. Friedman JH. Stochastic gradient boosting. Comput Stat Data An. 2002;38(4):367-78. doi:10.1016/S0167-9473(01)00065-2.

27. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012. 28(23):3150-2. doi:10.1093/bioinformatics/bts565.

28. Fukasawa Y, Tsuji J, Fu SC, Tomii K, Horton P, Imai K. MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites. Mol Cell Proteomics. 2015;14(4):1113-26. doi:10.1074/mcp.M114.043083.

29. Gonczarowska-Jorge H, Zahedi RP, Sickmann A. The proteome of baker's yeast mitochondria. Mitochondrion. 2017;33:15-21. doi:10.1016/j.mito.2016.08.007.

30. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644-52. doi:10.1038/nbt.1883.

31. Gray MW. Mitochondrial Evolution. Cold Spring Harb Perspect Biol. 2012;4(9): a011403. doi:10.1101/cshperspect.a011403.

32. Haas BJ, Papanicolaou A, Yassour M, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8(8):1494-512. doi:10.1038/nprot.2013.084.

33. Hagen KD, Hirakawa MP, House SA, et al. Novel structural components of the ventral disc and lateral crest in *Giardia intestinalis*. PLoS Negl Trop Dis. 2011;5(12):e1442. doi:10.1371/journal.pntd.0001442.

34. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol. 2018;35(2):518-522. doi:10.1093/molbev/msx281.

35. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and

AUGUSTUS. Bioinformatics. 2015;32(5):767-769. doi:10.1093/bioinformatics/btv661.

36. Hrdy I, Hirt RP, Dolezal P, et al. *Trichomonas* hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature. 2004;432(7017):618-22. doi:10.1038/nature03149.

37. Jedelský PL, Doležal P, Rada P, et al. The minimal proteome in the reduced mitochondrion of the parasitic protist *Giardia intestinalis*. PLoS One. 2011;6(2):e17285. doi:10.1371/journal.pone.0017285.

38. Jerlström-Hultqvist J, Einarsson E, Xu F, et al. Hydrogenosomes in the diplomonad *Spironucleus salmonicida*. Nat Commun. 2013;1.897916667. doi:10.1038/ncomms3493.

39. Jerlström-Hultqvist J, Franzén O, Ankarklev J, et al. Genome analysis and comparative genomics of a Giardia intestinalis assemblage E isolate. BMC Genomics. 2010;0.835416667. doi:10.1186/1471-2164-11-543.

40. Jones P, Binns D, Chang HY, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236-40. doi:10.1093/bioinformatics/btu031.

41. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14(6):587-589. doi:10.1038/nmeth.4285.

42. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. J Mol Biol. 2016;428(4):726-731. doi:10.1016/j.jmb.2015.11.006.

43. Karnkowska A, Vacek V, Zubáčová Z, et al. A Eukaryote without a Mitochondrial Organelle. Curr Biol. 2016;26(10):1274-84. doi:10.1016/j.cub.2016.03.053.

44. Katoh K & Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772-80. doi:10.1093/molbev/mst010.

45. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4): R36. doi:10.1186/gb-2013-14-4-r36.

46. Kolisko M, Silberman JD, Cepicka I, et al. A wide diversity of previously undetected free-living relatives of diplomonads isolated from marine/saline habitats. Environ Microbiol. 2010;12(10):2700-10. doi:10.1111/j.1462-2920.2010.02239.x.

47. Kumar A, Agarwal S, Heyman JA, et al. Subcellular localization of the yeast proteome. Genes Dev. 2002;16(6):707-19. doi:10.1101/gad.970902.

48. Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. Front Genet. 2013;0.33125. doi:10.3389/fgene.2013.00237.

49. Leger MM, Kolisko M, Kamikawa R, et al. Organelles that illuminate the origins of *Trichomonas* hydrogenosomes and *Giardia* mitosomes. Nat Ecol Evol. 2017;1(4):0092. doi:10.1038/s41559-017-0092.

50. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. Nucleic Acids Res. 2011;39(Database issue):D19-21. doi:10.1093/nar/gkq1019.

51. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. doi:10.1093/bioinformatics/btp352.

52. Lindmark DG & Müller M. Hydrogenosome, a cytoplasmic organelle of the anaerobic flagellate *Tritrichomonas foetus*, and its role in pyruvate metabolism. J Biol Chem. 1973;248(22):7724-8.

53. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. 2014;42(15):e119. doi:10.1093/nar/gku557.

54. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 2005;33(20): 6494–6506. doi:10.1093/nar/gki937.

55. Makiuchi T, Nozaki T. Highly divergent mitochondrion-related organelles in anaerobic parasitic protozoa. Biochimie. 2014;100:3-17. doi:10.1016/j.biochi.2013.11.018.

56. Martijn J, Vosseberg J, Guy L, Offre P, Ettema TJG. Deep mitochondrial origin outside the sampled alphaproteobacteria. Nature. 2018;557(7703):101-105. doi:10.1038/s41586-018-0059-5.

57. Martin WF, Roettger M, Ku C, Garg SG, Nelson-Sathi S, Landan G. Late Mitochondrial Origin Is an Artifact. Genome Biol Evol. 2017;9(2):373-379. doi:10.1093/gbe/evx027.

58. Millet CO, Williams CF, Hayes AJ, Hann AC, Cable J, Lloyd D. Mitochondria-derived organelles in the diplomonad fish parasite *Spironucleus vortens*. Exp Parasitol. 2013;135(2):262-73. doi:10.1016/j.exppara.2013.07.003.

59. Morrison HG, McArthur AG, Gillin FD, et al. Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. Science. 2007;317(5846):1921-6. doi:10.1126/science.1143837.

60. Müller M, Mentel M, van Hellemond JJ, et al. Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. Microbiol Mol Biol Rev. 2012;76(2):444-95. doi:10.1128/MMBR.05024-11.

61. Natekin A & Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot. 2013;7:21. doi:10.3389/fnbot.2013.00021.

62. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268-74. doi:10.1093/molbev/msu300.

63. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27(5):824-834. doi:10.1101/gr.213959.116.

64. Park JS, Kolisko M, Heiss AA, Simpson AG. Light microscopic observations, ultrastructure, and molecular phylogeny of Hicanonectes teleskopos n. g., n. sp., a deep-branching relative of diplomonads. J Eukaryot Microbiol. 2009;56(4):373-84. doi:10.1111/j.1550-7408.2009.00412.x.

65. Peña-Diaz P & Lukeš J. Fe–S cluster assembly in the supergroup Excavata. J Biol Inorg Chem. 2018;23(4):521-541. doi:10.1007/s00775-018-1556-6.

66. Pittis AA & Gabaldón T. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. Nature. 2016;531:101–104. doi:10.1038/nature16941.

67. R Core Team. R: A language and environment for statistical computing. 2018;https://www.R-project.org/.

68. Rada P, Doležal P, Jedelský PL, et al. The core components of organelle biogenesis and membrane transport in the hydrogenosomes of *Trichomonas vaginalis*. PLoS One. 2011;6(9):e24428. doi:10.1371/journal.pone.0024428.

69. Reinders J, Zahedi RP, Pfanner N, Meisinger C, Sickmann A. Toward the complete yeast mitochondrial proteome: multidimensional separation techniques for mitochondrial proteomics. J Proteome Res. 2006;5(7):1543-54. doi:10.1021/pr050477f.

70. Roger AJ, Muñoz-Gómez SA, Kamikawa R. The Origin and Diversification of Mitochondria. Curr Biol. 2017;27(21):R1177-R1192. doi:10.1016/j.cub.2017.09.015.

71. Savojardo C, Martelli PL, Fariselli P, Casadio R. TPpred2: improving the prediction of mitochondrial targeting peptide cleavage sites by exploiting sequence motifs. Bioinformatics. 2014;30(20):2973-4. doi:10.1093/bioinformatics/btu411.

72. Savojardo C, Martelli PL, Fariselli P, Casadio R. TPpred3 detects and discriminates mitochondrial and chloroplastic targeting peptides in eukaryotic proteins. Bioinformatics. 2015;31(20):3269-75. doi:10.1093/bioinformatics/btv367.

73. Schneider RE, Brown MT, Shiflett AM, et al. The *Trichomonas vaginalis* hydrogenosome proteome is highly reduced relative to mitochondria, yet complex compared with mitosomes. Int J Parasitol. 2011;41(13-14):1421-34. doi:10.1016/j.ijpara.2011.10.001.

74. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2017;Volume 34, Issue 17, Pages i884–i890. doi:10.1093/bioinformatics/bty560.

75. Sickmann A, Reinders J, Wagner Y, et al. The proteome of *Saccharomyces cerevisiae* mitochondria. Proc Natl Acad Sci U S A. 2003;100(23):13207-12. doi:10.1073/pnas.2135385100.

76. Simpson AG. Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota). Int J Syst Evol Microbiol. 2003;53(Pt 6):1759-77. doi:10.1099/ijs.0.02578-0.

77. SRA Toolkit Development Team. http://ncbi.github.io/sra-tools/.

78. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008. doi:10.1093/bioinformatics/btn013.

79. Stanke M, Schöffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics. 2006;0.334722222. doi:10.1186/1471-2105-7-62.

80. Suzuki S, Kakuta M, Ishida T, Akiyama Y. GHOSTX: an improved sequence homology search algorithm using a query suffix array and a database suffix array. PLoS One. 2014;9(8):e103833. doi:10.1371/journal.pone.0103833.

81. Takishita K, Kolisko M, Komatsuzaki H, et al. Multigene phylogenies of diverse Carpediemonas-like organisms identify the closest relatives of 'amitochondriate' diplomonads and retortamonads. Protist. 2012;163(3):344-55. doi:10.1016/j.protis.2011.12.007.

82. Tanifuji G, Takabayashi S, Kume K, et al. The draft genome of *Kipferlia bialata* reveals reductive genome evolution in fornicate parasites. PLoS One. 2018;13(3):e0194487. doi:10.1371/journal.pone.0194487.

83. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res. 2008;18(12):1979-90. doi:10.1101/gr.081612.108.

84. Thomas PD, Campbell MJ, Kejariwal A, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003;13(9):2129-41. doi:10.1101/gr.772403.

85. Tovar J, León-Avila G, Sánchez LB, et al. Mitochondrial remnant organelles of *Giardia* function in iron-sulphur protein maturation. Nature. 2003;426(6963):172-6. doi:10.1038/nature01945.

86. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9): 1105–1111. doi:10.1093/bioinformatics/btp120.

87. UniProt C. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45(D1):D158-69. doi:10.1093/nar/gkw1099.

88. Vacek V, Novák LVF, Treitli SC, et al. Fe-S Cluster Assembly in Oxymonads and Related Protists. Mol Biol Evol. 2018;35(11):2712-2718. doi:10.1093/molbev/msy168.

89. van Grinsven KW, Rosnowsky S, van Weelden SW, et al. Acetate:succinate CoA-transferase in the hydrogenosomes of *Trichomonas vaginalis*: identification and characterization. J Biol Chem. 2008. 283(3):1411-8. doi:10.1074/jbc.M702528200.

90. Waterhouse RM, Seppey M, Simão FA, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2017;35(3):543–548. doi:10.1093/molbev/msx319.

91. Williams KP, Sobral BW, Dickerman AW. A robust species tree for the alphaproteobacteria. J Bacteriol. 2007;189(13):4578-86. doi:10.1128/JB.00269-07.

92. Wilson D, Pethica R, Zhou Y, et al. SUPERFAMILY - Comparative Genomics, Datamining and Sophisticated Visualisation. Nucleic Acids Res. 2009;37(Database issue):D380-6. doi:10.1093/nar/gkn762.

93. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):R46. doi:10.1186/gb-2014-15-3-r46.

94. Woodcroft BJ, Scanlon KL, Doyle M, Speed T, Ralph SA. ApiLoc - A database of published protein sub-cellular localisation in Apicomplexa (version 3). 2011. http://apiloc.biochem.unimelb.edu.au/apiloc/apiloc. Updated 2011. Accessed June 11, 2016.

95. Yubuki N, Inagaki Y, Nakayama T, Inouye I. Ultrastructure and ribosomal RNA phylogeny of the free-living heterotrophic flagellate *Dysnectes brevis* n. gen., n. sp., a new member of the Fornicata. J Eukaryot Microbiol. 2007;54(2):191-200. doi:10.1111/j.1550-7408.2007.00252.x.

96. Yubuki N, Simpson AG, Leander BS. Comprehensive ultrastructure of *Kipferlia bialata* provides evidence for character evolution within the Fornicata (Excavata). Protist. 2013;164(3):423-39. doi:10.1016/j.protis.2013.02.002.

97. Zhang Q, Táborský P, Silberman JD, Pánek T, Čepička I, Simpson AG. Marine Isolates of *Trimastix marina* Form a Plesiomorphic Deep-branching Lineage within Preaxostyla, Separate from Other Known Trimastigids (Paratrimastix n. gen.). Protist. 2015;166(4):468-91. doi:10.1016/j.protis.2015.07.003.

98. Zubáčová Z, Novák L, Bublíková J, et al. The mitochondrion-like organelle of *Trimastix pyriformis* contains the complete glycine cleavage system. PLoS One. 2013;8(3):e55417. doi:10.1371/journal.pone.0055417.