Université de Montréal

In silico analysis of mitochondrial proteins

par
Yaoqing Shen

Département de Biochimie
Faculté de Médecine

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Doctorat
en Bio-informatique

Septembre, 2009

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée :

In silico analysis of mitochondrial proteins

présentée par :
Yaoqing Shen

a été évaluée par un jury composé des personnes suivantes :

Normand Brisson, président-rapporteur
Gertraud Burger, directeur de recherche
Mathieu Blanchette, membre du jury
John Parkinson, examinateur externe
Joelle Pelletier, représentant du doyen de la FES

Résumé

Le rôle important joué par la mitochondrie dans la cellule eucaryote est admis depuis longtemps. Cependant, la composition exacte des mitochondries, ainsi que les processus biologiques qui s'y déroulent restent encore largement inconnus. Deux facteurs principaux permettent d'expliquer pourquoi l'étude des mitochondries progresse si lentement : le manque d'efficacité des méthodes d'identification des protéines mitochondriales et le manque de précision dans l'annotation de ces protéines.

En conséquence, nous avons développé un nouvel outil informatique, YimLoc, qui permet de prédire avec succès les protéines mitochondriales à partir des séquences génomiques. Cet outil intègre plusieurs indicateurs existants, et sa performance est supérieure à celle des indicateurs considérés individuellement. Nous avons analysé environ 60 génomes fongiques avec YimLoc afin de lever la controverse concernant la localisation de la bêta-oxydation dans ces organismes. Contrairement à ce qui était généralement admis, nos résultats montrent que la plupart des groupes de Fungi possèdent une bêta-oxydation mitochondriale. Ce travail met également en évidence la diversité des processus de bêta-oxydation chez les champignons, en corrélation avec leur utilisation des acides gras comme source d'énergie et de carbone.

De plus, nous avons étudié le composant clef de la voie de bêta-oxydation mitochondriale, l'acyl-CoA déshydrogénase (ACAD), dans 250 espèces, couvrant les 3 domaines de la vie, en combinant la prédiction de la localisation subcellulaire avec la classification en sous-familles et l'inférence phylogénétique. Notre étude suggère que les gènes ACAD font partie d'une ancienne famille qui a adopté des stratégies évolutionnaires

innovatrices afin de générer un large ensemble d'enzymes susceptibles d'utiliser la plupart des acides gras et des acides aminés. Finalement, afin de permettre la prédiction de protéines mitochondriales à partir de données autres que les séquences génomiques, nous avons développé le logiciel TESTLoc qui utilise comme données des *Expressed Sequence Tags* (ESTs). La performance de TESTLoc est significativement supérieure à celle de tout autre outil de prédiction connu.

En plus de fournir deux nouveaux outils de prédiction de la localisation subcellulaire utilisant différents types de données, nos travaux démontrent comment l'association de la prédiction de la localisation subcellulaire à d'autres méthodes d'analyse *in silico* permet d'améliorer la connaissance des protéines mitochondriales. De plus, ces travaux proposent des hypothèses claires et faciles à vérifier par des expériences, ce qui présente un grand potentiel pour faire progresser nos connaissances des métabolismes mitochondriaux.

Mots clefs: mitochondrie, prédiction de la localisation subcellulaire, apprentissage par la machine, bêta-oxydation, dégradation des acides gras, dégradation des acides aminés, acyl-CoA déshydrogénase, évolution, marqueurs de séquence exprimés.

Abstract

The important role of mitochondria in the eukaryotic cell has long been appreciated, but their exact composition and the biological processes taking place in mitochondria are not yet fully understood. The two main factors that slow down the progress in this field are inefficient recognition and imprecise annotation of mitochondrial proteins.

Therefore, we developed a new computational tool, YimLoc, which effectively predicts mitochondrial proteins from genomic sequences. This tool integrates the strengths of existing predictors and yields higher performance than any individual predictor. We applied YimLoc to ~60 fungal genomes in order to address the controversy about the localization of beta oxidation in these organisms. Our results show that in contrast to previous studies, most fungal groups do possess mitochondrial beta oxidation. This work also revealed the diversity of beta oxidation in fungi, which correlates with their utilization of fatty acids as energy and carbon sources. Further, we conducted an investigation of the key component of the mitochondrial beta oxidation pathway, the acyl-CoA dehydrogenase (ACAD). We combined subcellular localization prediction with subfamily classification and phylogenetic inference of ACAD enzymes from 250 species covering all three domains of life. Our study suggests that ACAD genes are an ancient family with innovative evolutionary strategies to generate a large enzyme toolset for utilizing most diverse fatty acids and amino acids. Finally, to enable the prediction of mitochondrial proteins from data beyond genome

sequences, we designed the tool TESTLoc that uses expressed sequence tags (ESTs) as input. TESTLoc performs significantly better than known tools.

In addition to providing two new tools for subcellular localization designed for different data, our studies demonstrate the power of combining subcellular localization prediction with other *in silico* analyses to gain insights into the function of mitochondrial proteins. Most importantly, this work proposes clear hypotheses that are easily testable, with great potential for advancing our knowledge of mitochondrial metabolism.

Keywords: mitochondria, subcellular localization prediction, machine learning, beta oxidation, fatty acid degradation, amino acid degradation, acyl-CoA dehydrogenase, evolution, expressed sequence tags

Table of contents

Résumé iii
Abstract v
List of tables ix
List of figures x
Acknowledgements xiii
Introduction 1
1 Mitochondria and their importance in eukaryotic cells 1
1.1 Origin and morphology 1
1.2 Protein import 3
1.3 Metabolic pathways in mitochondria 10
1.4 Mitochondrial diseases 17
2. Identification of mitochondrial proteins 19
2.1 Experimental approaches to identify mitochondrial proteins 19
2.2 *In silico* identification of mitochondrial proteins 24
2.3 Limitations in predicting mitochondria-destined proteins 36
3 From protein inventory to biological processes 39
3.1 The beta oxidation puzzle 40
3.2 Acyl-CoA dehydrogenase 42
Objectives 45
Chapter 1 Mitochondrial protein prediction by integrating heterogeneous tools 47
Chapter 2 Plasticity of a key metabolic pathway in fungi 65
Chapter 3 Diversity and dispersal of acyl-CoA dehydrogenases 83
Chapter 4 *In silico* identification of mitochondrion-targeted proteins using EST data 123
Conclusions 171
1. Localization is an important aspect of protein function 171
2. From protein localization to pathway localization 172
3. The power of cross-taxon comparison 173

4 Factors that influence localization prediction accuracy.....175
4.1 The influence of training data.....176
4.2 The influence of sequence features.....176
4.3 The influence of computational methods.....177
Perspectives.....178
References.....I
Supplementary information.....IX

List of tables

Chapter 1

Table 1. Examples of conflicting results from individual prediction tools 51

Table 2. Decision trees built in this study and the individual tools employed to construct each tree 53

Table 3. Performance of the best predictors for the three different prediction schemes 54

Table 4. Example proteins used for decision tracing 56

Chapter 2

Table 1. Beta oxidation pathways in fungi 71

Chapter 3

Table 1. Pairwise sequence similarities between human ACAD subfamily members 86

Table 2. Seed sequences used for BLAST searches 87

Chapter 4

Table 1. Number of sequences (from Arabidopsis and all plants tested) used in this study 149

Table 2. The amino acids grouped according to their chemical properties or structures 150

Table 3. The independent evaluation results of different prediction schemes 151

List of figures

Introduction

Figure 1. Simplified structure of mitochondria 2
Figure 2. Two main protein import pathways of mitochondria 5
Figure 3. Mitochondrial intermembrane-space import and assembly machinery 7
Figure 4. Sorting and assembly machinery of the outer mitochondrial membrane 9
Figure 5. The TCA cycle. Figure from (Berg, Tymoczko et al. 2002) 11
Figure 6. The urea cycle in mitochondria and the cytosol 12
Figure 7. The electron transport chain and ATP synthesis 13
Figure 8. Heme biosynthesis 14
Figure 9. A model for the mechanism of Fe–S-protein biogenesis in eukaryote 15
Figure 10. One iteration of the beta oxidation spiral 17
Figure 11. Schema of a decision tree 26
Figure 12. The decision boundary for classification in SVM 27
Figure 13. Projection of linearly non-separable data into a higher dimension 30
Figure 14. The schema of an artificial neural network 31
Figure 15. Simplified scheme depicting the different roles of mitochondria and peroxisomes in the beta oxidation of fatty acids 41

Chapter 1

Figure 1. Prediction performance of individual and integrated tools on yeast mitochondrial proteins 52
Figure 2. Integration of heterogeneous prediction tools by decision trees 53
Figure 3. Prediction performance of individual and integrated tools on yeast mitochondrial membrane and matrix proteins 54
Figure 4. Decision tree topology for the prediction of mitochondrial proteins. 55

Chapter 2

Figure 1. Enzyme architecture of the various beta oxidation pathways 68
Figure 2. Beta oxidation enzymes mapped on the fungal phylogeny tree 70

Chapter 3

Figure 1. Optimal substrates of ACAD subfamilies 86
Figure 2. Flowchart of the procedure for assigning ACAD subfamilies 89
Figure 3. Distinction of ACD10 and ACD11 90
Figure 4. ACAD subfamily distribution mapped on the taxonomy hierarchy from NCBI 92
Figure 5. Alignment of residues that affect substrate specificity of human ACDSB 94
Figure 6. Schematic phylogenetic trees of ACAD subfamilies 95

Chapter 4

Figure 1. Prediction performance of top-ranked available tools and TESTLoc for mitochondrial proteins from plant EST-derived peptides 154
Figure 2. Selection of *Arabidopsis* ESTs. 155
Figure 3. Fragmentation of plant ESTs to expand the EST-peptide data. 156
Figure 4. Sequence identities within expanded data set, calculated from BLASTP alignment. 157
Figure 5. Training and evaluation of SVM. The procedure in each dash box was repeated ten times 158
Figure 6. The independent evaluation for SVMs based on different orders amino acid composition 159
Figure 7. Integration of predictions from SVM models based on individual features 161
Figure 8. Comparison of available tools and TESTLoc for the prediction power on recognizing mitochondrial proteins from jakobids ESTs 162

To my dear parents:

Shen, Rongken and Yin, Suxia

Acknowledgements

I would like to thank all the individuals who supported, encouraged, and inspired me to finish the work presented in this thesis.

First, I would like to express my heartiest gratitude to my supervisor, Prof. Gertraud Burger, for her devoted supervision, constant encouragement, and invaluable support during the whole study. Whenever I needed her help, she was always accessible and generously donated her time. Her insight has significantly facilitated the progress of my project. Outside the lab, her warm care made my life in a foreign country much easier.

I sincerely thank my thesis committee members, Prof. Normand Brisson, Prof. John Parkinson, and Prof. Mathieu Blanchette, who spent time evaluating my work. Special thanks also go to other committee members of my predoctoral exam, Prof. Michael Hallett and Dr. Sébastien Lemieux, whose constructive suggestions helped my project proceed.

My keen appreciation goes to Professor B. Franz Lang, who taught me phylogenetic analysis step by step. His ability to see through the data and get the history behind them is a true gift. I benefited so much from his wisdom and knowledge.

I would like to express my gratitude to Dr. Emmet O'Brien, expert on databases and elegant English sentences, for critically reading my manuscripts and this thesis. My warm thanks are extended to Eric Wang, for helping with the implementation of YimLoc, and Dr. Veronique Marie for the inspiring discussion and the excellent French translation.

My special thanks go to Dr. Sivakumar Kannan. Being the first student in the bioinformatics program, he generously shared with me his experiences with both academic and administrative issues, which helped me to find my way.

I am grateful to Dr. Henner Brinkmann. There is always something interesting to learn from him in discussion, within and beyond phylogeny. I also want to thank Prof. Nicolas Lartillot for his help with PhyloBayes.

My sincere thanks to everybody in the group, Claudia, Natacha, Cecile, Allan, Georgette, Lise, Shona, Yifei, Rachid, Pasha, who make the lab a convivial place to work in. Their friendship and company made my thesis journey not lonely.

Geneviève Galarneau and Jean-François Théroux are two lovely internship students. It was a great pleasure working with them. I really appreciate their contribution to my project.

Two beautiful ladies, Elaine Meunier and Marie Robichaud deserve my special thanks. Their kind help makes things easier for a student who barely speaks French in a francophone university.

I am deeply indebted to my parents, who are always supportive of my decisions, and encouraging me to pursue my dream.

I greatly appreciate the financial support from Canadian Institute for Health Research (CIHR) Strategic Training Grant in Bioinformatics.

Introduction

1 Mitochondria and their importance in eukaryotic cells

First descriptions of the organelle known as the mitochondrion can be dated back to 1850. But it has taken about 100 years of intense work to recognize this organelle as the powerhouse of the cell, which provides the energy currency ATP for various biological reactions. In the following half century, numerous studies revealed that the role of mitochondria in the cell is far more complex than merely being the ATP supplier (McBride, Neuspiel et al. 2006). Mitochondria host many life-maintenance processes of the eukaryotic cell, including carbohydrate metabolism, fatty acid metabolism, amino acid metabolism, heme biosynthesis, coenzyme Q biosynthesis, and Fe-S cluster biosynthesis. Mitochondria are also important regulators of the cell, involved in cell cycle regulation and initiation of apoptosis. Further, mitochondria are the main source of reactive oxygen species (ROS), which may cause oxidative damage of proteins, DNAs and lipids. Malfunction of mitochondria is implicated in a variety of diseases. The role of mitochondria in neurodegeneration, aging, and tumor formation is receiving increasing interests in biomedical studies.

1.1 Origin and morphology

It is widely accepted that mitochondria originated from an ancient α-Proteobacterium which established endosymbiosis with a host cell (an archaeal or a primitive eukaryotic cell) (Andersson, Zomorodipour et al. 1998; Gray 1998). Mitochondria of extant eukaryotes still retain many bacterial features, such as a double membrane (Figure 1), their own genome,

and transcription/translation machineries. However, during evolution the coding capacity of the mitochondrial genome has shrunk drastically. Most of the ancestral bacterial genes moved to the nucleus or were lost for good. The coding capacity of contemporary mitochondrial genomes varies from ~70 proteins in *Reclinomonas americana* to three proteins in *Plasmodium falciparum* (Lang, Burger et al. 1997; Conway, Fanello et al. 2000). These numbers are, by far, less than the number of proteins located in mitochondria. According to a recent estimation, animals possess ~1500 proteins in their mitochondria, and yeast ~1000 (Meisinger, Sickmann et al. 2008). Therefore, around 99% of mitochondrial proteins must be encoded in the nucleus, translated in the cytosol, and imported into mitochondria.

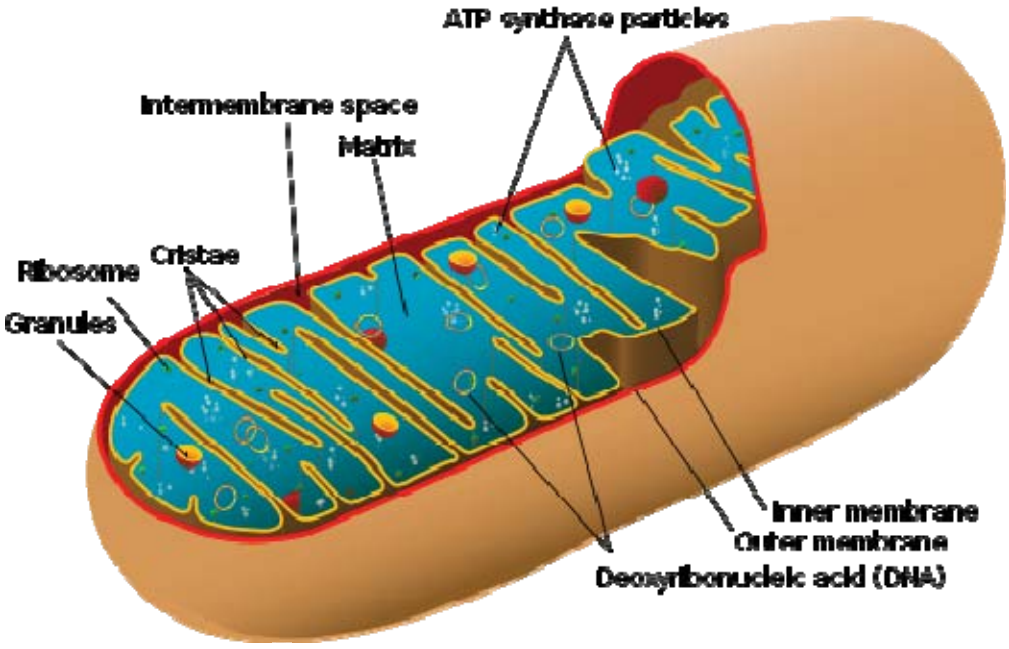


Figure 1. Simplified structure of mitochondria. (Figure from http://en.wikipedia.org/wiki/File:Animal_mitochondrion_diagram_en.svg)

1.2 Protein import

The machinery importing proteins into mitochondria is not fully understood. While co-translational import is only little characterized, we know most about post-translational import. The import machineries involved consist of two major protein complexes: the translocase of the outer membrane (TOM complex) and the translocase of the inner membrane (TIM complex), as well as numerous auxiliary chaperone complexes. Mitochondrial import machineries are sophisticated and versatile. Four different sorting pathways have been identified (reviewed in Bolender, Sickmann et al. 2008), but their exact mechanism remains unclear.

1.2.1 TOM complex

TOM is a complex residing in the outer membrane of mitochondria (Figure 2). It is composed of seven subunits: Tom5, Tom6, Tom7, Tom20, Tom22, Tom40, and Tom70, designated according to their molecular weights. These subunits have different roles in protein import: Tom20, Tom22, and Tom70 are the receptors; Tom40 forms the import channel; Tom5 mediates the insertion of proteins into the import channel; Tom6 and Tom7 stabilize the TOM complex (Bolender, Sickmann et al. 2008).

1.2.2 TIM complex

Two TIM complexes have been identified: The TIM23 complex transports proteins with an N-terminal targeting peptide, whereas TIM22 complex transports proteins with internal targeting signals (Figure 2). So far identified components of TIM23 complex include

Tim23 that forms the transmembrane channel, Tim17 that regulates the Tim23 channel, Tim21 that interacts with the TOM complex, and Tim50 that controls the opening/closing of Tim23 channel. The TIM22 complex consists of the channel-forming subunit Tim22, chaperon-interacting Tim12, and two subunits, Tim18 and Tim54, of unknown function (Bolender, Sickmann et al. 2008).

1.2.3 Import of matrix proteins

The import of most matrix proteins is guided by the mitochondrial targeting peptide (MTP, Figure 2), a poorly conserved N-terminal presequence that usually contains 20-80 residues forming a positively charged amphipathic α-helix. The targeting peptide is recognized by the receptors Tom20 and Tom22 of the TOM complex, and guides the whole protein passing through the Tom40 channel. By interaction of Tom22 and Tim50, the imported protein is transferred from the TOM complex to TIM23, and passes the inner membrane through the Tim23 channel. Then the matrix heat shock protein 70 drives the imported protein into the matrix, where the targeting peptide is removed (Bolender, Sickmann et al. 2008).

1.2.4 Import of inner membrane proteins

Some inner membrane proteins have in addition to the MTP a sorting signal that guides their integration into the lipid phase of the membrane (Figure 2). But many inner membrane proteins such as the ADP/ATP carrier and phosphate carrier do not possess a MTP. Instead, they carry internal targeting signals and form a loop topology when they are recognized by

Tom70 and imported through the Tom40 channel. In the intermembrane space, these looped proteins bind chaperone complexes such as Tim9-Tim10 and Tim8-Tim13, which transfer the proteins to the TIM22 complex in the inner membrane (Endres, Neupert et al. 1999; Rehling, Model et al. 2003). The imported proteins are then inserted into the inner membrane through the Tim22 pores, driven by the membrane potential.

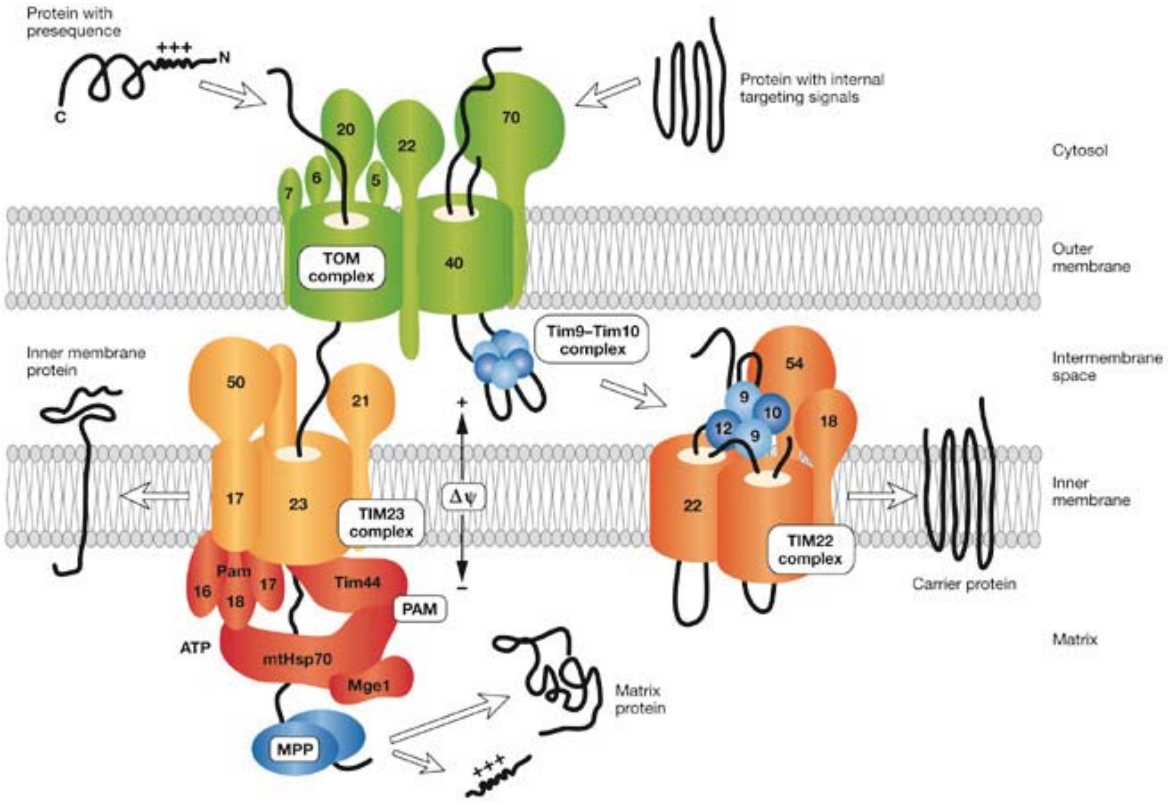


Figure 2. Two main protein import pathways of mitochondria. Presequences direct proteins through the TOM complex, TIM23 complex and motor PAM to the matrix; the mitochondrial processing peptidase (MPP) removes the presequences. Cleavable inner-membrane proteins are released laterally from the TIM23 complex. Carrier precursors with

internal targeting signals are recognized by the receptor Tom70 and translocated by the TOM complex and the Tim9–Tim10 chaperone of the intermembrane space. The TIM22 complex promotes insertion of carrier proteins into the inner membrane. MtHsp70, matrix heat shock protein 70; PAM, presequence translocase-associated motor; TIM, translocase of the inner membrane; TOM, translocase of the outer membrane. Figure and legend from (Bolender, Sickmann et al. 2008).

1.2.5 Import of intermembrane space proteins

The intermembrane space is rich in small proteins carrying cysteine motifs, and the Tim chaperones are part of them. These proteins are synthesized in the cytosol without MTP, and their import requires the Mitochondrial Intermembrane Space Assembly machinery (MIA, Chacinska, Pfannschmidt et al. 2004; Wiedemann, Pfanner et al. 2006). It is believed that the central component of MIA, Mia40, binds to the incoming intermembrane-space proteins after they pass through the TOM complex, and promotes the assembly of these proteins into functional complexes (Figure 3).

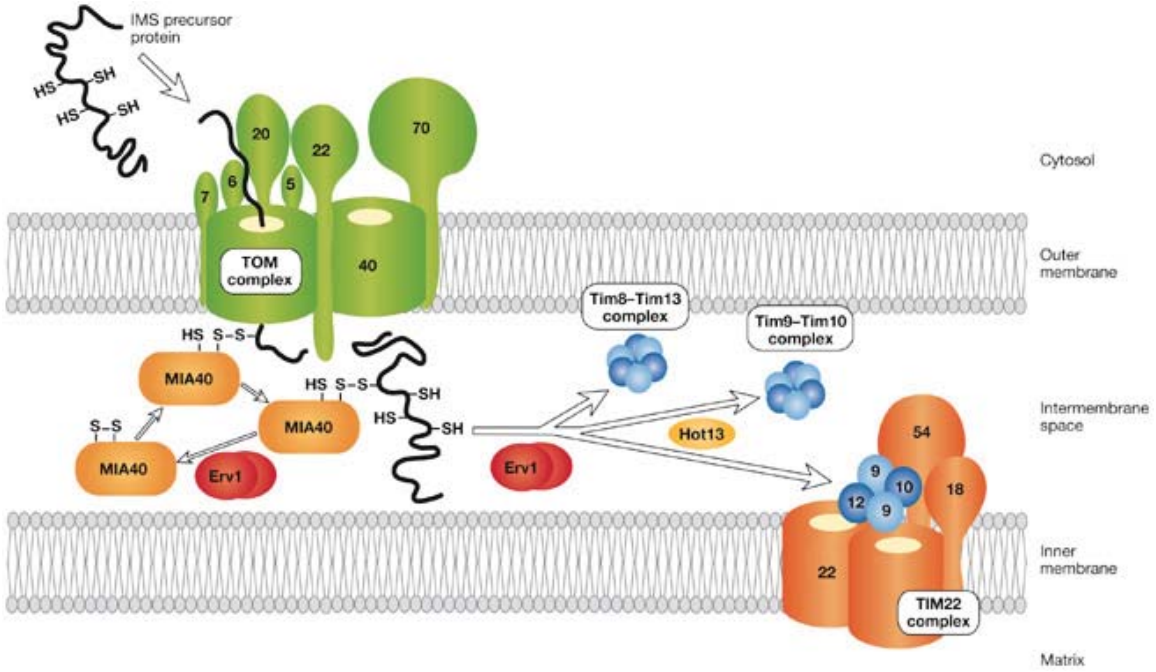


Figure 3. Mitochondrial intermembrane-space import and assembly machinery. Precursors of small intermembrane space (IMS) proteins are translocated through the TOM complex and bound by Mia40 through disulphide bonds. The sulphydryl oxidase Erv1 cooperates with Mia40 in the oxidation of precursor proteins and their assembly into oligomeric complexes. Further factors such as Hot13 support assembly of the protein complexes. Erv1, essential for respiration and viability 1; Hot13, helper of TIM13; Mia40, mitochondrial intermembrane space import and assembly; TIM, translocase of the inner membrane; TOM, translocase of the outer membrane. Figure and legend from (Bolender, Sickmann et al. 2008).

1.2.6 Import of outer membrane proteins

The mitochondrial outer membrane contains proteins with one or more transmembrane domains. For proteins with one or two alpha-helix transmembrane domains, it seems that their interaction with the TOM complex is sufficient for transport and insertion into the membrane. But proteins with beta-barrel need in addition the sorting and assembly complex (SAM) to assist in their import, folding, and membrane insertion (Figure 4, Wiedemann, Kozjak et al. 2003; Pfanner, Wiedemann et al. 2004). The exact process is yet unknown.

Overall, the mitochondrial protein import mechanism is only partially understood. Although MTP is widely used as a marker of mitochondrial proteins, it is estimated that more than 50% of mitochondrial proteins do not depend on the presence of a targeting signal for the import. So far no universal property of mitochondrial proteins has been identified, which makes their recognition challenging.

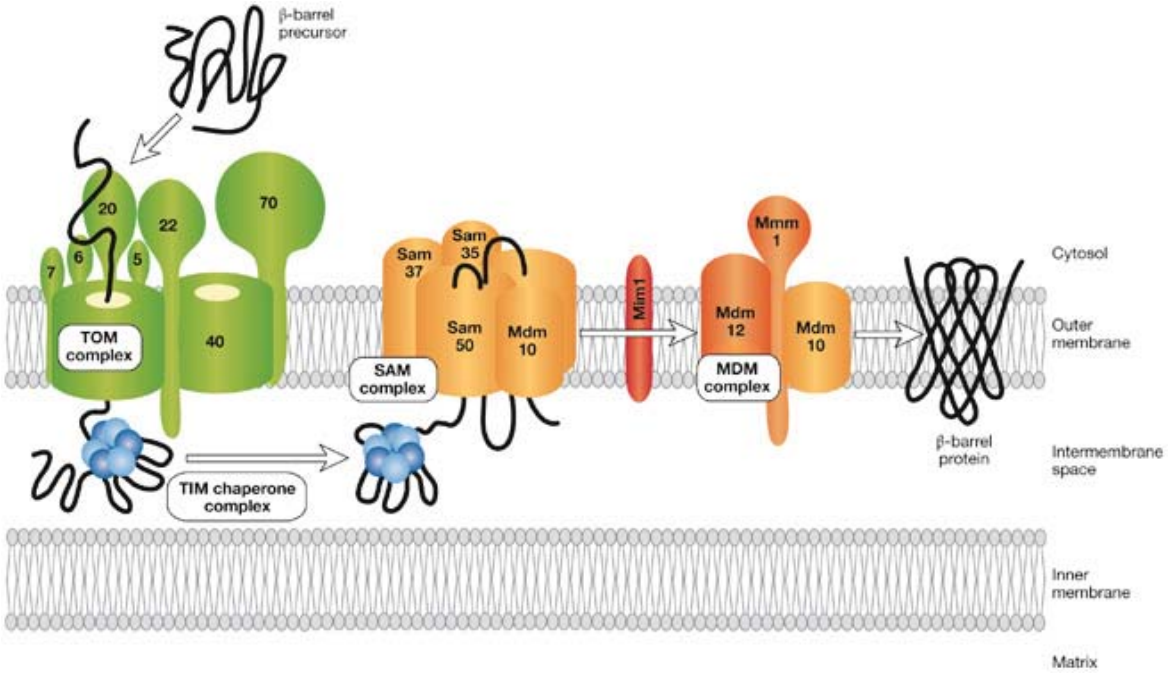


Figure 4. Sorting and assembly machinery of the outer mitochondrial membrane. The precursors of beta-barrel proteins are initially imported through the TOM complex, interact with small TIM chaperones (Tim9–Tim10 complex, Tim8–Tim13 complex) in the intermembrane space, and are inserted into the outer membrane by the SAM complex. Other outer membrane proteins—the MDM complex and Mim1—support assembly of beta-barrel proteins. Mdm, mitochondrial distribution and morphology; Mim1, mitchondrial import 1; Mmm1, maintenance of mitochondrial morphology 1; SAM, sorting and assembly machinery; TIM, translocase of the inner membrane; TOM, translocase of the outer membrane. Figure and legend from (Bolender, Sickmann et al. 2008).

1.3 Metabolic pathways in mitochondria

Mitochondria labor crucial pathways for eukaryotic cells, including metabolism of carbohydrates, fatty acids, and amino acids. The biosynthesis of heme, coenzyme Q, and Fe-S centers also take place within mitochondria. The following is a brief summary of the major mitochondrial metabolic pathways.

1.3.1 The tricarboxylic cycle (TCA) cycle

The TCA cycle converts the intermediate products of fatty acids, amino acids, and carbohydrates into CO_2 or other chemical compounds, and yields energy from this process (Figure 5, Berg, Tymoczko et al. 2002). The TCA cycle not only breaks down molecules, but also produces building blocks for other reactions, such as ketoglutarate for glutamate synthesis, citrate for cholesterol synthesis, oxaloacetate for glucose synthesis, and succinyl-CoA for heme synthesis. Therefore the TCA cycle fills a central position in the mitochondrial metabolism.

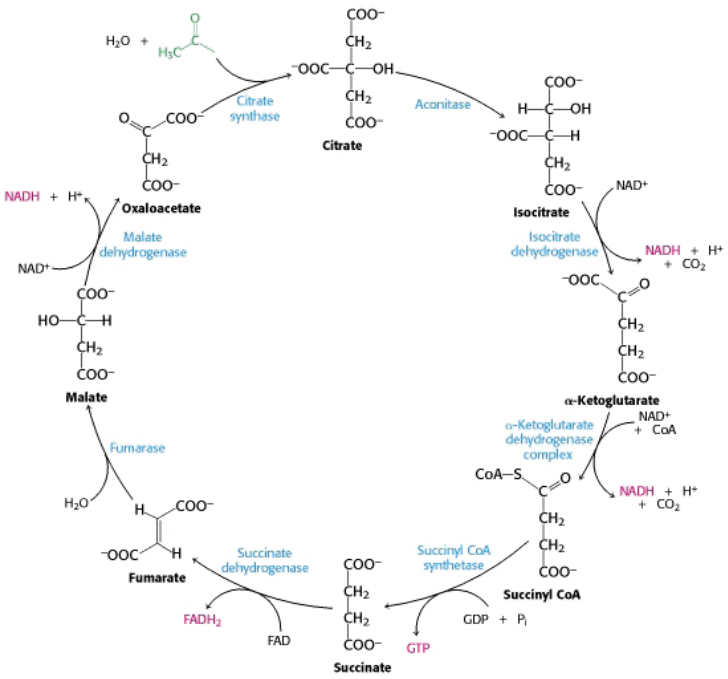


Figure 5. The TCA cycle. Figure from (Berg, Tymoczko et al. 2002).

1.3.2 Urea cycle

The urea cycle breaks down amino acids and converts nitrogen into urea. Two of the five reactions in the urea cycle occur in mitochondria, with carbamoyl phosphate synthetase converting ATP and bicarbonate to carbamoyl phosphate, and ornithine transcarbamoylase catalyzing the reaction between carbamoyl phosphate and ornithine to form citrulline. Citrulline is exported out of mitochondria and the remaining reactions of this cycle are carried out in cytosol (Figure 6). This pathway is crucial for removing excess amino acids and nitrogen from the cell.

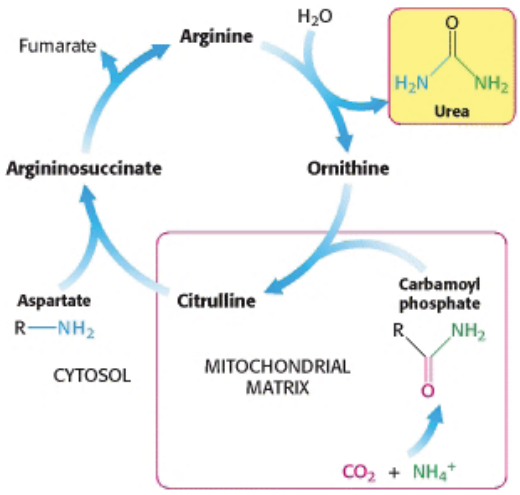


Figure 6. The urea cycle in mitochondria and the cytosol. Figure from (Berg, Tymoczko et al. 2002).

1.3.3 Oxidative phosphorylation

Oxidative phosphorylation is a mitochondrial process that generates ATP through oxidation of nutrients (Figure 7). Electrons from various metabolites are ultimately transferred to oxygen, via four complexes residing in the inner membrane of mitochondria: NADH-coenzyme Q oxidoreductase (complex I), Succinate-Q oxidoreductase (complex II), Q-cytochrome c oxidoreductase (complex III), Cytochrome c oxidase (complex IV). Protons are pumped into the intermembrane space via complex I, III, and IV, to build the proton potential which is used by ATP synthase (complex V) to generate ATP. The proton potential is also required for the protein import into mitochondria (reviewed in Scheffler 2008).

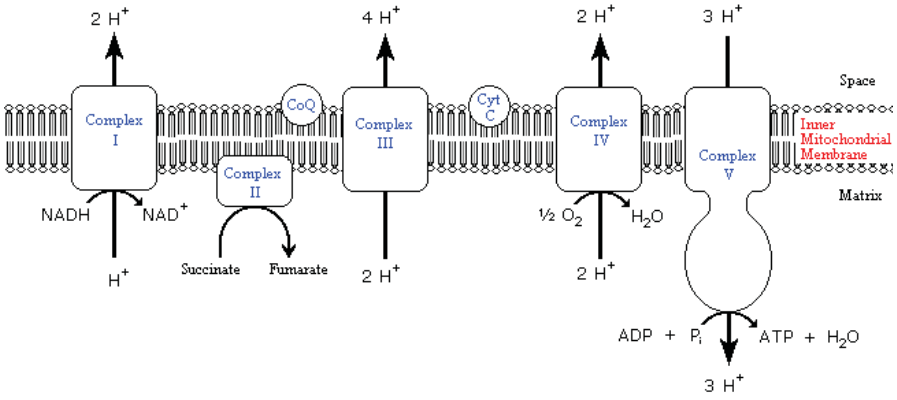


Figure 7. The electron transport chain and ATP synthesis. Figure from http://www.biomed.metu.edu.tr.

1.3.4 Heme biosynthesis

Heme consists of a porphyrin with an iron atom in its centre. It serves as a prosthetic group for hemoproteins and cytochromes. Several steps of heme biosynthesis occur in mitochondria (Figure 8). It starts with the conversion of succinyl-CoA, an intermediate from the TCA cycle, to δ-aminolevulinic acid (ALA). ALA is exported out of mitochondria and converted into coproporphyrinogen III through several reaction steps in the cytosol. The latter component is imported back into mitochondria, where it is converted into porphyrin and then, through the addition of the iron, into heme.

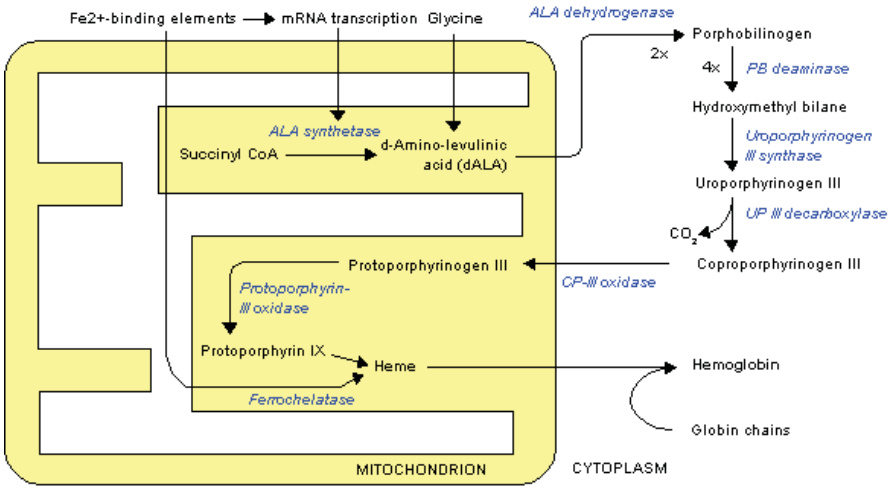


Figure 8. Heme biosynthesis. Figure from http://en.wikipedia.org/wiki/Heme.

1.3.5 Iron-sulfur cluster biosynthesis

Iron-sulfur clusters (ISCs) are ensembles of iron and sulfide that function as co-factors of iron-sulfur proteins such as complex I-III of the respiration chain, ferredoxin, glutamate dehydrogenase, and DNA glycosylase. ISCs are mainly involved in electron transfer and redox reactions, but they also contribute to substrate binding and structural stabilization of proteins. Iron-sulfur proteins have been found in mitochondria, cytosol, and nucleus, but the synthesis of ISC starts in mitochondria (Figure 9). Reduced iron Fe^{2+} and sulfur released from cysteine are bound to the scaffold proteins Isu1/2 and form the ISC. Then the ISCs are transferred to apoproteins in mitochondria, or exported to cytosol to synthesize extra-mitochondrial Fe-S proteins.

Interestingly, in organisms with degenerated mitochondria, such as *Encephalitozoon cuniculi* and *Giardia intestinalis*, many of the mitochondrial pathways were lost, while ISC

synthesis has remained (Henze and Martin 2003; Goldberg, Molik et al. 2008). This suggests that ISC synthesis is one of the most essential contributions of mitochondria to the cell and requires the particular environment of this organelle.

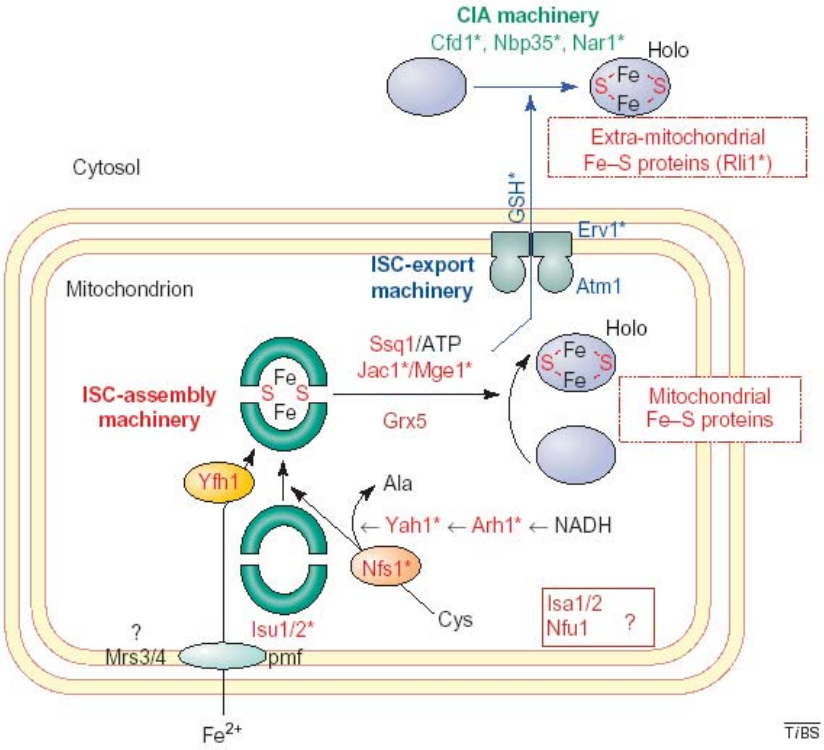


Figure 9. A model for the mechanism of Fe–S-protein biogenesis in eukaryotes. Pmf represent the proton motive force required for the import of reduced iron. Mrs3 or Mrs4 are carrier proteins. Isu1/Isu2 (green) serves as a scaffold for the synthesis of the ISC. Metal delivery to Isu1/Isu2 is assisted by frataxin (Yfh1; yellow). The cysteine desulfurase Nfs1 (orange) mediates the release of sulfur from cysteine, with the electron transferred from NADH, ferredoxin reductase (Arh1), and ferredoxin (Yah1). Ssq1–ATP, Jac1 and Mge1 (representing DnaK-, DnaJ- and GrpE-like chaperones, respectively), and the glutaredoxin

Grx5 are the chaperone systems required after ISC assembly on the Isu proteins. Isa1/Isa2 and Nfu1 proteins take part in the ISC biogenesis, but their function is still unclear. The maturation of extra-mitochondrial Fe–S proteins requires the cytosolic iron-sulfur protein assembly machinery, consisting Nar1, Cfd1 and Nbp35. The ABC transporter Atm1, the sulfhydryl oxidase Erv1, and the tripeptide glutathione (GSH) are part of the system that exports the yet uncharacterized product of ISC-assembly to the cytosol. '*' denotes components encoded by essential genes in yeast. Figure and modified legend from (Lill and Muhlenhoff 2005).

1.3.6 Beta oxidation

Beta oxidation, the major pathway for fatty acid degradation in the cell, breaks down fatty acids in a cyclic four-step process: dehydrogenation, hydration, dehydrogenation, and thiolysis (Figure 10). After each cycle, two carbons (α and β) are removed from the fatty acid, in the form of acetyl-CoA. This pathway feeds electrons into the respiratory chain for energy production, and acetyl-CoA into the TCA cycle, gluconeogenesis, and ketogenesis pathways. Beta oxidation disorders in human cause accumulation of acylcarnitine, which may lead to life-threatening liver dysfunction (Kompare and Rizzo 2008).

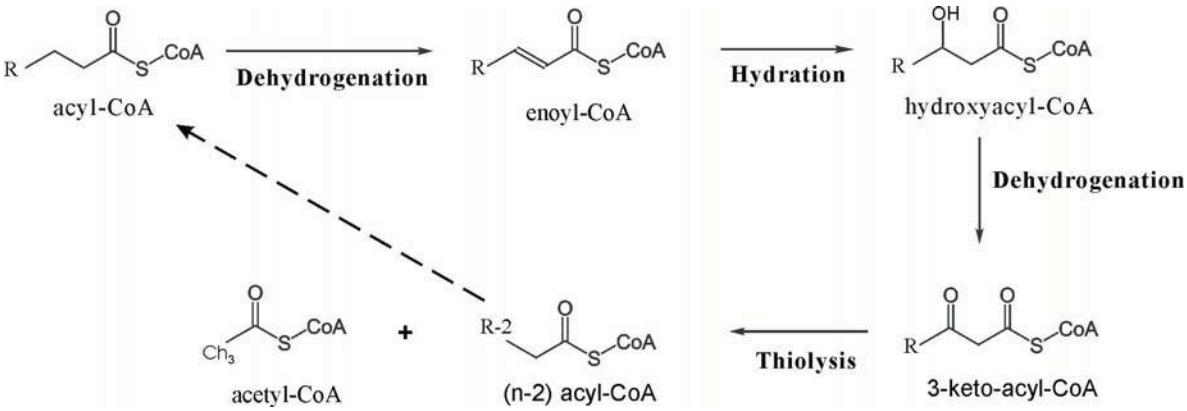


Figure 10. One iteration of the beta oxidation spiral.

Among the various pathways in mitochondria, beta oxidation is a special case because it can simultaneously reside in peroxisomes (Poirier, Antonenkov et al. 2006). The two pathway forms have very similar components that are difficult to distinguish by sequence similarity. Subcellular localization prediction of the components is required in order to identify the two forms. Intriguing questions are why the cell needs duplication, how prevalent the dual form is throughout the tree of life, and how the duplication arose during evolution. The dual localization of beta oxidation will be presented in more details in Section 3.1.

1.4 Mitochondrial diseases

In human, malfunction of mitochondria can cause disease or even be lethal, and mostly entails neuromuscular disorder, due to the high energy demand of brain and muscle cells (Smeitink, van den Heuvel et al. 2001). Hereditary mitochondrial diseases are relatively frequent (Schaefer, Taylor et al. 2004; Uusimaa, Moilanen et al. 2007), with only a few

therapies available. Abnormal mitochondrial function can be caused by both mitochondrion-encoded and nucleus-encoded genes. Mutations in mitochondrial DNA have been identified in approximately 50% of the known mitochondrial diseases, including mitochondrial encephalomyopathy with lactic acidosis and stroke-like episodes (Fan, Civalier et al. 2006), and Leber's hereditary opticus neuropathy (Martin-Kleiner, Gabrilovac et al. 2006). Malfunction of nucleus-encoded mitochondrial proteins such as frataxin, whose mutation leads to Friedreich's ataxia (Priller, Scherzer et al. 1997), is likely the cause of the numerous mitochondrial disorders for which no mitochondrial DNA mutation has been found. Mitochondrial malfunction can also be the consequence of disease. For example, impaired electron transport chain and increased ROS are observed in diabetes and obesity (Ritov, Menshikova et al. 2005; Amaral, Oliveira et al. 2008), and reduced ATP levels are detected in Alzheimer's and Parkinson's disease (Crouch, Cimdins et al. 2007; Schapira 2008).

With the role of mitochondrial dysfunction in various diseases being increasingly appreciated, studies on mitochondria as drug target are emerging (Koene and Smeitink 2009). Mitochondrial gene therapy, as well as metabolic manipulation of perturbed biological processes, have been applied in attempts to correct the dysfunction. Mitochondria are also the target of antioxidant drugs, to prevent the damage of proteins, DNA, or lipids caused by ROS. In addition, mitochondria are regarded as promising targets for cancer therapies, which aim at selectively disabling mitochondria to shut down ATP

supply, or alternatively, at stimulating mitochondria to induce apoptosis (Ralph and Neuzil 2009).

2. Identification of mitochondrial proteins

Great efforts have been made and are still ongoing to reveal the composition of the mitochondrial proteome, and to unravel the biological processes in which the proteome takes part. It is the complex nature of mitochondria and the diversity of their proteins among and within species that makes this undertaking highly challenging.

2.1 Experimental approaches to identify mitochondrial proteins

A variety of molecular biology approaches are available today to identify the subcellular localization of proteins. These approaches are briefly reviewed below.

2.1.1 Immonofluorescence analysis of epitope-tagged proteins

In this approach, antibodies or other affinity reagents that specifically bind to the target proteins are labeled with fluorescence, and used to probe where proteins are localized in the cell. However, since it requires fixed and permeabilized cells, it cannot capture the temporal patterns of protein expression. This approach is usually used for individual proteins, but recently a large scale study was reported in the context of the Swedish Human Protein Atlas program (Barbe, Lundberg et al. 2008), which aims to label all human

proteins with antibodies in order to visualize their localization in different cells and tissues. The current release of the Human Protein Atlas contains more than 8,800 antibodies and over 7,300,000 images. However, application of immonofluorescence analysis in such large-scale studies requires prior knowledge of the target proteins and availability of their antibodies, which makes it unsuitable for *de novo* identification of unknown proteins of a given cellular compartment.

2.1.2 Co-expression of fluorescent proteins

Localization of proteins can also be visualized via co-expression of a target protein with a fluorescent reporter protein, for example the green fluorescent protein (GFP). GFP genes are fused with cDNA clones of target proteins to transfect cells, in which GFP is expressed together with the target protein and its localization is captured by fluorescence microscopy. An advantage of this approach is that it shows the protein's location in the living cell. A proteome-scale co-expression study has been conducted in yeast (Kumar, Agarwal et al. 2002; Huh, Falvo et al. 2003). Although it works well in many cases, some studies showed that changing the natural amino acid sequence of a protein by tagging with GFP may cause mis-localization of the fused protein. Furthermore, the tagged proteins have to be expressed in high levels to guarantee sufficient signal intensity, which may also lead to artifacts (Sickmann, Reinders et al. 2003; Seibel, Eljouni et al. 2007).

2.1.3 Transcriptomics approach

Transcriptomics tracks the expression level of mRNAs under different conditions. Genes showing a similar expression pattern often encode functionally or physically interacting proteins, which in turn suggests the common subcellular localization (Lascaris, Bussemaker et al. 2003). In particular, genes whose expression level changes when cells switch from aerobic to anaerobic growth conditions likely encode mitochondrial proteins (DeRisi, Iyer et al. 1997). However, as respiration is only one of several mitochondrial functions, a large number of mitochondrial proteins remain undetected. Also, it is difficult to distinguish mitochondria-located proteins from those that are essential to mitochondrial function, but located elsewhere, such as the nucleus-localized ribonuclotide-diposphate reductase complex which affects the stability of mitochondrial DNA by modulating the mitochondrial deoxynucleoside triphosphate pool (O'Rourke, Doudican et al. 2005).

2.1.4 Knockout/knockdown phenotype

Mitochondrial proteins can also be recognized by gene silencing studies. When mitochondrial metabolism is turned down or mitochondrial morphology is affected after a gene is knocked out or knocked down, it is likely that this gene encodes a mitochondrial protein. This method has been applied to mitochondrial gene screens in yeast, fruit fly, and *Caenorhabditis elegans* (Dimmer, Fritz et al. 2002; Chen, Shi et al. 2008; Ichishita, Tanaka et al. 2008). But proteins encoded by redundant genes are difficult to find by this approach, unless all the copies are knocked out. In addition, proteins encoded by essential genes will not be detected, because deletion of these genes leads to cell death.

2.1.5 Large-scale proteomics

The proteomics approach is the most direct way of identifying mitochondrial proteins. Mitochondria are first isolated by density gradient centrifugation or antibody-coated beads. Proteins are extracted and then separated by methods like two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) or sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), and identified by mass spectrometry techniques such as liquid-chromatography-tandem MS (LC-MS/MS). Proteomics identification of mitochondrial proteins has been applied to rice (Heazlewood, Howell et al. 2003), *Arabidopsis* (Heazlewood, Tonti-Filippini et al. 2004), yeast (Sickmann, Reinders et al. 2003; Reinders and Sickmann 2007), a few mammals (Taylor, Fahy et al. 2003; Forner, Foster et al. 2006; Johnson, Harris et al. 2007), *C. elegans* (Li, Cai et al. 2009), and the protist *Tetrahymena thermophila* (Smith, Gawryluk et al. 2007). The proteomics approach is particularly suited for detecting new mitochondrial proteins, but it has two main drawbacks: (i) contamination by proteins from other organelles, which leads to a false-positive rate up to 40%; and (ii) the low coverage (estimated to be 23~40%) of the proteome, because proteins of low abundance or high hydrophobicity often escape detection (Pagliarini, Calvo et al. 2008).

2.1.6 Contribution of experimental studies on mitochondria

The above mentioned studies have provided valuable information about the composition of the mitochondrial proteome and the function of this organelle. Protein sequences and their annotation have been compiled in several databases, such as MitoP2 for yeast, *Neurospora*

crassa, human, mouse, and *Arabidopsis* (Elstner, Andreoli et al. 2009), MitoCarta for human and mouse (Pagliarini, Calvo et al. 2008), MitoProteome (Cotter, Guda et al. 2004) and HMPDb (http://bioinfo.nist.gov/) for human, MitoDrome (Sardiello, Licciulli et al. 2003) for fruit fly, TRIPLES (Kumar, Agarwal et al. 2002), YMPD (http://bmerc-www.bu.edu/projects/mito/), and YDPM (Steinmetz, Scharfe et al. 2002) for yeast, and AMPDB for *Arabidopsis* (Heazlewood and Millar 2005). In addition, SWISSPROT stores a large number of sequences with experimentally confirmed localization.

The experimental efforts have identified new pathways to complete our understanding of mitochondrial biology, and provided insights into the mechanisms of mitochondrial bio-processes. For example, the screening of yeast mitochondrial proteome for essential proteins leads to the identification of Mia40, an important component of the machinery for import and assembly of mitochondrial intermembrane space proteins (Sickmann, Reinders et al. 2003; Chacinska, Pfannschmidt et al. 2004). A comprehensive analysis of protein localization in yeast allowed to build the global interaction map of proteins in the cell (Huh, Falvo et al. 2003). A systematic screen of a yeast deletion mutant library of ~5,000 nonessential yeast genes identified a set of genes involved in mitochondrial structure and function, including known genes that were never previously related to mitochondrial morphogenesis, and new genes as well (Dimmer, Fritz et al. 2002). A genome-wide RNA interference screen in *Drosophila* revealed novel modulators of mitochondrial biogenesis and function, such as *klumpfuss* in apoptosis, *smt3* in protein stability and proteolysis, and *barren* in cell cycle regulation (Chen, Shi et al. 2008).

2.2 *In silico* identification of mitochondrial proteins

Experimental approaches for identifying mitochondrial proteins can be expensive in time and costs, in particular for non-model systems. This sets the stage for computational methods, which infer the localization of a protein from its sequence. Design and implementation of *in silico* localization prediction approaches is an active research area, and a number of tools have been developed. Computational methods, mostly based on machine learning techniques, learn from proteins of known localization the rules that allow prediction of unknown ones. Machine learning approaches are especially suitable for localization prediction, since the mechanisms of protein sorting to different compartments are not well understood, and expert knowledge to guide the prediction is limited. The machine learning scheme is able to learn from data the distinctive features of proteins targeted to various subcellular compartments, features which are otherwise difficult to detect. Several machine learning algorithms have been applied in this context, including Naïve Bayes classifiers (Duda, Hart et al. 2001), K-nearest neighbors (Shakhnarovich, Darrell et al. 2005), neural networks (Priddy and Keller 2005), and support vector machines (Boser, Guyon et al. 1992). In the world of machine learning, the localization prediction is a classification problem. The features used to describe the sequence are called attributes, and the subcellular compartment of each sequence is called a class. The following is a brief review of some widely used machine learning approaches for classification.

2.2.1 Commonly used machine learning techniques for classification

2.2.1.1 Decision tree

Decision tree is a machine learning algorithm which learns from training data to generate IF-THEN rules for classification. The graphical representation of the decision-making process resembles a tree (Figure 11), in which the internal nodes specify the attributes, and the terminal nodes indicate the classes.

A decision tree is constructed by selecting attributes to split the data set for classifying them as accurately as possible. The selection of attributes is via the calculation of information gain (IG), which in turn is evaluated by the reduction of entropy (MacKay 2003). For an information source S which emits symbols from an alphabet (Berg, Tymoczko et al. 2002) with probabilities $\{p_1, p_2,\ldots p_k\}$, given that the emission of each symbol is independent of the others, the entropy of S is defined as:

$$H(S)=\sum_{i=1}^{k} - p_i \log p_i \quad (1)$$

If each element in S is described by the attribute x, which has the value v with probability of p(x=v), then the conditional entropy H(S|x) is the entropy of S when the value of x is given.

$$H(S|x)=\sum_{v\in V} p(x = v)H(S \mid x = v) \quad (2)$$

The information gain by knowing x is the reduction of entropy when the value of x is given.

$$IG(S|x)=H(S)-H(S|x) \quad (3)$$

A number of decision tree algorithms have been developed. Most of them are variations of a core algorithm that employs a top-down, greedy search through the space of possible decision trees, by selecting at each step the attributes yielding the largest information gain. By this principle, decision tree programs construct a decision tree T from the given training data. One of the most successful and widely used decision tree algorithms is C4.5 (Quinlan 1993), and it has been applied to subcellular localization prediction by the tool LOCtree (Nair and Rost 2005).

One advantage of decision trees is that it generates understandable rules that illustrate how the decisions are made. It also indicates which of the attributes are most important for the prediction, which can be regarded as feature selection procedure. Another advantage of decision trees is the fast computation with large datasets. However, decision trees are prone to errors due to class imbalance. It favours the correct classification for data belonging to classes of larger size, at the cost of wrong predictions for classes with fewer instances.

.

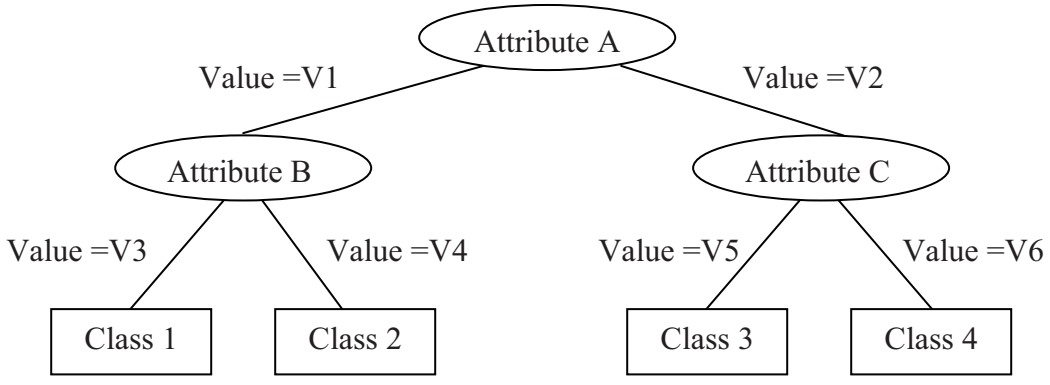


Figure 11. Schema of a decision tree.

2.2.1.2 Support Vector Machine

Support Vector Machine (SVM) is a widely used machine learning method. It is designed for the binary classification and separates the two classes by a hyperplane (Boser, Guyon et al. 1992). For data $\{x_i, y_i\} \in S$, where x_i is the attribute and $y_i \in \{1,-1\}$ is the class, SVM finds a hyperplane

$$w \cdot x + b = 0 \quad (4)$$

so that all instances of class=1 are on one side of the hyperplane while those of class=-1 are on the other side (Figure 12). The instances that satisfy $w \cdot x_i + b = 1$ (H_1 in Figure 12) or $w \cdot x_i + b = -1$ (H_2 in Figure 12) are called support vectors, and for all instances in S,

$$y_i(w \cdot x_i + b) \geq 1, \ \forall i. \quad (5)$$

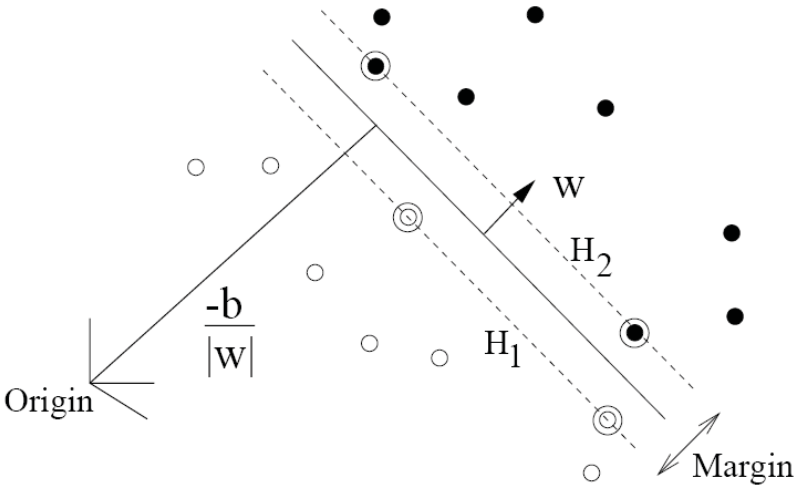


Figure 12. The decision boundary for classification in SVM. Open symbols and filled symbols are two classes. Symbols in circles are support vectors. Figure from (Burges 1998).

The hyperplane (4) is chosen by maximizing the margin $\frac{2}{\| w \|}$ between the two hyperplanes H_1 and H_2. This is equal to minimize ||w||, satisfying

$y_i(w \cdot x_i+b) \geq 1$ (6)

To find the hyperplane that minimizes ||w||, the question is first transformed to minimize $\frac{1}{2}||w||^2$, subject to $y_i(w \cdot x_i+b) \geq 1$. However, in reality the data may not be linearly separable, a few instances may be in the margin area or misclassified while the majority of data is still well separated. To allow such error, a slack factor ξ is introduced in (6) so that

$y_i(w \cdot x_i+b) \geq 1-\xi_i$ where $\xi_i \geq 0$ (7)

To control the error, the term $C\sum_i \xi_i$ is added to the optimization problem $\min \frac{1}{2}||w||^2$, which is changed to:

$$\min \frac{1}{2}||w||^2 C\sum_i \xi_i, \quad \text{subject to } y_i(w \cdot x_i+b) \geq 1-\xi_i, \tag{8}$$

where C is a parameter chosen by the user. The higher the value of C, the higher is the penalty assigned to errors.

Using the method of Lagrange multipliers, the form (8) is converted to

$$\max \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j (x_i \bullet x_j), \text{ subject to } \sum_i \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C \tag{9}$$

$w = \sum_i y_i \alpha_i x_i$ with $\alpha_i \geq 0$

The discriminant function for classification is

$$\sum_i y_i \alpha_i x_i \cdot x + b \quad (10)$$

Some data are not separable even with the slack factor being introduced. To make them linearly separable, one way is to project them into a higher dimension Φ: x-> Φ(x) (Figure 13). After the projection, the discriminant function (10) is transformed to

$$\sum_i y_i \alpha_i \langle \phi(x_i), \phi(x) \rangle + b, \quad (11)$$

where α_i is the solution to the optimization problem (12)

$$\max \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j \langle \phi(x_i) \cdot \phi(x_j) \rangle \text{ with constrains } \sum_i \alpha_i y_i = 0,\ 0 \le \alpha_i \le C. \quad (12)$$

It is not necessary to know exactly the function Φ(x), since solving the problem just requires the product of Φ(x) and $\Phi(x_i)$. If we define a kernel k (x, x`) = $\Phi(x) \cdot \Phi(x_i)$, then the product can be effectively computed with a proper kernel k without specifying Φ. Commonly used kernels include the polynomial (13) and the RBF kernel (or Gaussian kernel, 14)

$$k(x, x') = (x \cdot x' + k)^d \quad (13)$$

$$k(x, x') = \exp(-\gamma ||x - x'||^2) \quad (14)$$

Several subcellular localization prediction tools are built with SVM, such as LOCSVMPSI (Xie, Li et al. 2005), and ESLpred (Bhasin and Raghava 2004). SVM is fast to compute, and finding the optimal hyperplane is guaranteed. Unlike decision tree, SVM is less prone to class imbalance, as the classification boundary is only determined by the

support vectors. However, it is difficult to interpret how the classification is made and extract the underlining rules.

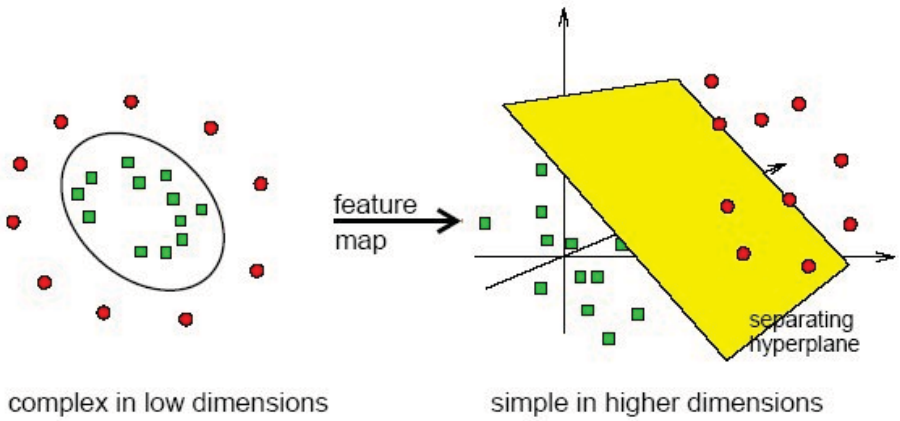


Figure 13. Projection of linearly non-separable data into a higher dimension. Figure from http://www.dtreg.com/svm.htm.

2.2.1.3 Artificial neural network

Artificial neural network (ANN) is a machine learning method that borrows the principles of biological neural network. The basic component of an ANN is the neuron, which is a computation node that accepts inputs and calculates the output according to a pre-defined function. There are multiple types of neural networks as well as the training algorithms, for example, the widely used feed-forward neural network and back-propagation algorithm. A feed-forward neural network is composed of three parts, an input layer, one or more hidden layers, and an output layer (Figure 14). The connection of nodes from different layers has a

certain strength, which is called weight. Training of a neural network starts with random weights, then calculates the output and the error, which is used to adjust the weights so that the error decreases. The procedure is repeated until an error minimum is found (reviewed in Krogh 2008). The tool TargetP for MTP recognition is based on ANN (Reinhardt and Hubbard 1998).

One advantage of ANN is its ability to capture long range dependency within input data. However, the adjustment of weights is often trapped in a local error minimum instead of reaching the global error minimum. As in the case of SVM, the result of ANN does not provide straightforward insights into the underlying rules of classification, while the computation time is usually longer than SVM.

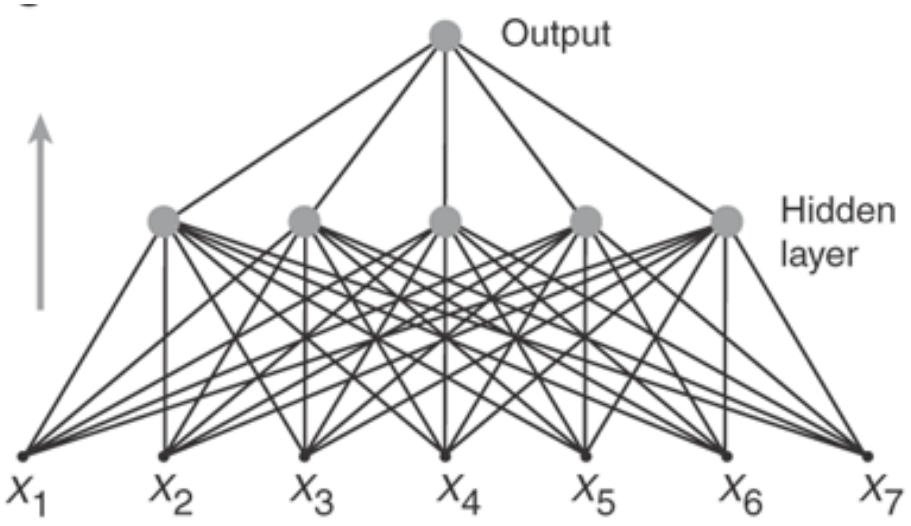


Figure 14. The schema of an artificial neural network. X_1 to X_7 indicates nodes of the input layer. The arrow indicates the direction of data flow. Figure from (Krogh 2008).

2.2.1.4 K-nearest neighbors

K-nearest neighbours (KNN) is a non-parametric learning method which, instead of formulating a generalized model, makes predictions by comparing the distance between unknown data and known data (Shakhnarovich, Darrell et al. 2005). Given a set of training data $\{(x_i,y_i)\}^m$ in which x are the attributes and y are the classes, and a distance matrix d which measures the distance between x, KNN just remembers the data. When a query is given, KNN will calculate its distance to each x_i, sort the distances, and assign the query to the class to which the majority of k nearest data points belongs, where k is predefined by the user. KNN is a straightforward method for classification, with easily interpretable results. But it is influenced by the class imbalance problem and the definition of distance matrix. KNN has been exploited in the localization predictor PSORT (Horton and Nakai 1997).

2.2.2 Commonly used sequences features for classification

Compared with the selection of the particular computational algorithm, the choice of sequence features that represent the proteins has a higher impact on the accuracy of prediction. According to the sequence features used, localization prediction tools can be divided into two categories: predictions based on annotation and predictions based on sequence, which are reviewed below.

2.2.2.1 Annotation-based localization prediction

Textual descriptions of protein function may bear clues to the localization. For example, if the annotation contains terms like “succinate dehydrogenase” and “Krebs cycle”, it is very likely that the protein is located in mitochondria. Several predictors use the annotation to predict where a protein is destined in the cell. The web-server PA-SUB (Proteome Analyst Specialized Subcellular Localization Server) first searches for presumptive homologous sequences of the query protein in the SWISSPROT database, and then makes the prediction based on keywords in the SWISSPROT entries of matched proteins (Lu, Szafron et al. 2004). Similarly, ProLoc-GO (Huang, Tung et al. 2008) and LocKey (Lu and Hunter 2005) exploit Gene Ontology (GO) terms in the annotation, and SherLoc looks for keywords in PubMed abstracts (Shatkay, Hoglund et al. 2007).

Another approach relying on sequence annotation makes predictions based on co-occurrence of functional motifs/domains in proteins from different subcellular compartments. The tool PSLT (Sarda, Chua et al. 2005) predicts the localization by searching for common InterPro motifs and specific membrane domains.

Despite being straightforward and biologically interpretable, annotation-based predictions have several limitations. First, homology is inferred from sequence similarity, but this alone is not always stringent enough. Second, some biological processes occur in multiple subcellular compartments, such as beta oxidation found in mitochondria and peroxisomes, and ATP synthesis found in mitochondria, chloroplasts, and vacuoles. Obviously, predictions based on imprecise annotation terms are prone to error. Moreover,

for many species, particularly the poorly studied, a large portion of proteins have no similar sequences in public databases, making the prediction based on annotation unfeasible.

2.2.2.2 Sequence-based localization prediction

Sequence-based prediction methods do not rely on homology or annotation of proteins. Instead, they search for intrinsic sequence features that distinguish proteins in different subcellular locations. The most commonly used feature is the targeting signal, i.e., a short peptide motif typically at the N-terminus that guides proteins to their destination. Recognized signals include the signal peptide (SP) for proteins exported out of the cell, the nuclear localization signal (NLS), the mitochondrial targeting peptide (MTP), the chloroplast targeting peptide (CTP), and the peroxisomal targeting signal (PTS). Some signals have a clear consensus pattern, for example the NLS and PTS, while others are poorly conserved, in particular MTP, and therefore difficult to detect. Several tools identify targeting signals from protein sequence, and predict the localization accordingly. Examples are MitoProt (Claros and Vincens 1996), TargetP (Emanuelsson, Nielsen et al. 2000), iPSORT (Bannai, Tamada et al. 2002), Protein Prowler (Boden and Hawkins 2005), and Predotar (Small, Peeters et al. 2004). But these tools only work effectively when: (i) the accurate N-terminal sequence of a protein is known, which in itself is a challenge for proteins inferred from genomic data; and (ii) the protein carries a known targeting signal, which does not apply to all proteins. These two factors limit the sensitivity of localization prediction based on targeting signals.

Since the protein sorting mechanism in the cell is not fully understood, *ab initio* prediction methods have been developed, which do not rely on any prior knowledge. *Ab initio* methods learn from the sequence the most relevant features that disclose a protein's localization. Clues may be obtained from the amino acid composition or from physicochemical properties such as molecular weight, net charge, and hydrophobicity (Andrade, O'Donoghue et al. 1998; Guda and Subramaniam 2005). The tools NNPSL and Subloc classify proteins according to the frequency of each amino acid (Reinhardt and Hubbard 1998; Hua and Sun 2001), while PLOC exploits the dipeptide and gapped amino acid pair composition (Park and Kanehisa 2003). Physicochemical properties of amino acids are calculated in pSLIP to distinguish proteins from different compartments (Sarda, Chua et al. 2005).

2.2.2.3 Localization prediction based on the integration of multiple protein features

Several tools combine multiple features to improve prediction accuracy. Targeting motifs plus amino acid hydrophobicity are used in PSORTII (Horton and Nakai 1997). Mitopred evaluates PI value, amino acid composition, and the presence of functional domains (Guda, Fahy et al. 2004). Amino acid composition is combined with physicochemical properties in ESLpred (Bhasin and Raghava 2004). Sequence-based and annotation-based prediction are exploited together in SherLoc (Shatkay, Hoglund et al. 2007), and the upgraded version of MultiLoc2 adds GO terms and phylogenetic profile (Hoglund, Donnes et al. 2006). A more detailed list of the various prediction tools is compiled in the Supplementary Table 1 of Chapter 4.

For well studied model organisms such as yeast, integration of localization information from various experimental and computational approaches yields comprehensive and accurate prediction. For example, a Bayesian system was used to integrate over 30 features from yeast sequences, including signal peptide, mRNA expression level, knockout mutation, various functional motifs, and amino acid composition (Drawid and Gerstein 2000). Another study, also on yeast, combined 22 datasets of reference mitochondrial proteins from previous experiments, such as GFP localization, deletion phenotype, orthologs, protein abundance, and MS data, as well as computational predictions from PSORT and Predotar (Prokisch, Scharfe et al. 2004). Not surprisingly, the integration outperforms simple predictions, but such rich information is restricted to only a few well-studied species where abundant experimental data are available.

2.3 Limitations in predicting mitochondria-destined proteins

2.3.1 Contradictory predictions by available tools

Many of the prediction tools have been applied to identify the mitochondria proteome in species whose whole genome sequence is available. Depending on the approach used, the predicted proteomes vary considerably in size, and the sets of proteins predicted as being part of the proteome overlap only partially. For example, three predictors iPSORT, PSORT II and TargetP have been applied to all hypothetical proteins from ten eukaryotic species, including two vertebrates, two arthropods, one nematode, two yeasts, one plant, and two

protists (Richly, Chinnery et al. 2003). The same trend was observed in all ten species: the number of mitochondrial proteins identified by all three predictors is less than 20% of that predicted by at least one predictor.

When the predictions of diverse tools disagree, deciding which one to trust is a conundrum for users. In fact, as the tools are based on different sequence features and training sets, they could have complementary prediction power. For example, targeting peptide recognition for mitochondrial matrix proteins combined with transmembrane domain recognition of mitochondrial membrane proteins may yield a more complete repertoire of mitochondrial proteome than any scheme used alone. To fully exploit the potential complementary strength of each method, it appears promising to integrate not only features, but also the tools exploiting different features. We will address this issue in Chapter 1.

2.3.2 Limited type of data that current tools are applicable to

Another limitation of the current tools is that they are designed for full-length proteins. Many of them require that input data must contain the starting methionine. Therefore, large-scale predictions have only been reported in species for which well annotated whole genome data are available. But for many organisms, such comprehensive information is lacking, which largely precludes *in silico* identification of their mitochondrial proteome.

2.3.3 Can Expressed Sequence Tag data serve for subcellular localization prediction?

The most abundant available data for non-model organisms are expressed sequence tags (ESTs, Parkinson and Blaxter 2009). Unlike genomic data, ESTs correspond to fragments of a gene's coding region, sometimes including 3' untranslated regions (3'-UTRs). To generate ESTs, mRNAs are extracted and reverse transcribed to cDNAs, which are sequenced to get single-pass reads (ESTs) of typically 50-500 nt in length. Since ESTs exclude non-coding regions, which can make up over 90% of a genome, they can be obtained at a lower cost and in a shorter time-frame compared to genomic sequences.

EST data are available for a large number of species. In September, 2009, the dbEST database, a repository of ESTs at National Center for Biotechnology Information (NCBI), contained ~63 million entries from over 1,800 species. Among them, ~500 species have over 10,000 ESTs, while around 1,000 species have more than 1,000 ESTs. In addition, ~370,000 clustered EST sequences from 49 organisms, mostly unicellular eukaryotes, are compiled in TBestDB, a taxonomically broad database of ESTs (O'Brien, Koski et al. 2007). Many of these species are poorly studied and cover a much wider taxonomic range than all the eukaryotic species with nuclear genomes completely sequenced. Therefore, these data should provide valuable insights into the diversity of the eukaryotic proteome. In particular for subcellular localization study, ESTs have an additional advantage over genome sequences. Distinct products of the same gene, caused by alternative splicing, may be destined for different subcellular compartments (Ashibe, Hirai et al. 2007; Hunt, Greene

et al. 2007; Ueyama, Lekstrom et al. 2007). This may be recognized by predictions based on ESTs.

So far, no study has been reported using ESTs for subcellular localization predictions. Indeed, the available tools work only poorly on EST-derived protein fragments (referred to as EST-peptides hereafter) due to the way ESTs are generated. EST-peptides usually lack the N-terminal part of the proteins, where otherwise many targeting signals are located. Furthermore, as partial sequences, EST-peptides may have a different amino acid composition than full length proteins, which may also cause inferior performance. In order to exploit EST data for protein localization prediction, new tools specifically tailored for this type of sequence data are needed. We will address this issue in Chapter 4.

3 From protein inventory to biological processes

Knowing the mitochondrial proteome is only the first step toward elucidating the biology of mitochondria. Proteins identified by localization prediction must undergo further analysis to reveal their function, the pathways in which they participate, and ultimately how the pathways interact to form the metabolic network. Localization prediction not only serves as a starting point for subsequent functional analyses. It sometimes provides clues in itself for unraveling biological processes, especially those that occur in multiple subcellular compartments, where each carries out a distinct role. Beta oxidation is one such process.

3.1 The beta oxidation puzzle

Initially discovered in rat, beta oxidation is one of the first recognized mitochondrial metabolic pathways (Kennedy and Lehninger 1949). However, its localization in most other species remains controversial. In mammals, two forms of beta oxidation have been identified, the mitochondrial form and the peroxisomal form. Although resembling each another in several aspects, the two forms require different enzyme sets (Figure 1 of Chapter 2) and serve distinct physiological roles (Figure 15). Mitochondria host the beta oxidation of all short- and medium-chain, and most long-chain fatty acids. But fatty acids with extremely long chains, for example hexacosanoic acid (26 carbons), and several types of branched-chain fatty acids, such as pristanic acid (2,6,10,14-tetramethylpentadecanoic acid), di- and tri-hydroxycholestanoic acid, are exclusively degraded in peroxisomes (Figure 15). In mitochondria, fatty acids are completely oxidized to CO_2 and H_2O, while peroxisomal beta oxidation only shortens fatty acid chains to a certain length, which are then imported into mitochondria for further degradation. In contrast, fatty acids can be completely degraded by the peroxisomal pathway if it is the only form of beta oxidation in the species, as is the case in some plants and *Saccharomyces*. Therefore, knowing the form of beta oxidation in a given species helps to track the degradation path of fatty acids in the cell.

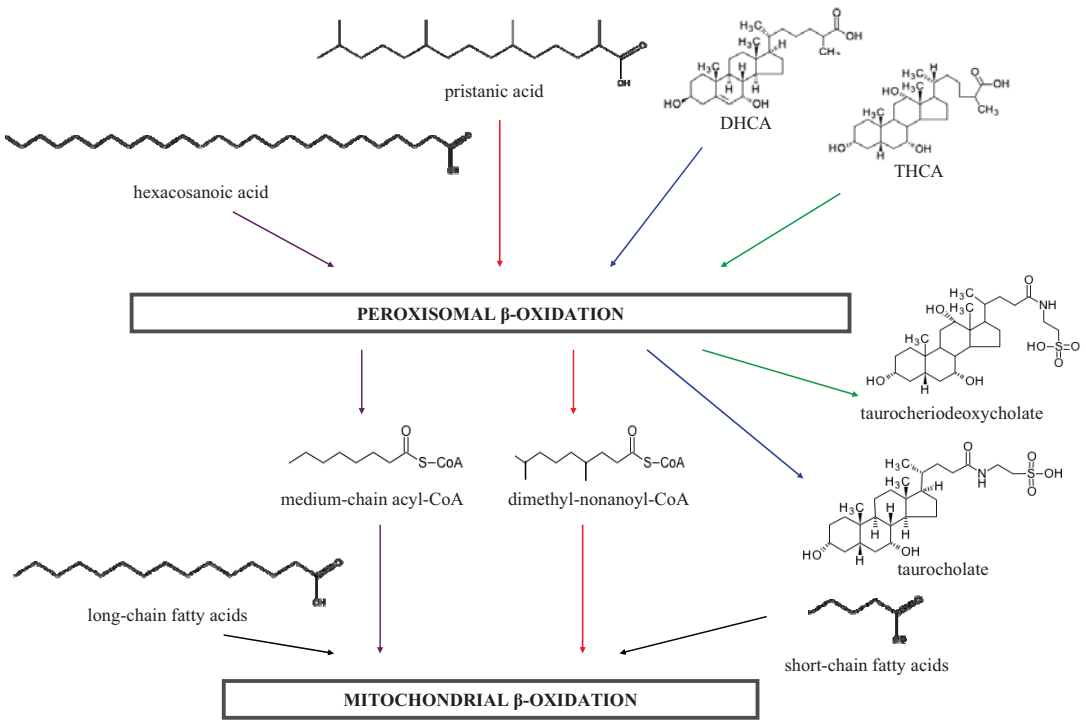


Figure 15. Simplified scheme depicting the different roles of mitochondria and peroxisomes in the beta oxidation of fatty acids. DHCA: dihydroxycholestanoic acid; THCA: trihydroxycholestanoic acid (THCA). Colored arrows show the degradation of different types of fatty acids. Modified from (Wanders, Vreken et al. 2001).

Does mitochondrial beta oxidation exist beyond animals? Not long ago, the answer was thought to be no. Plants and fungi seemed to possess only peroxisomal beta oxidation. While the absence of mitochondrial beta oxidation has been confirmed in several yeast species, a recent study suggests the presence of this pathway in another ascomycete fungus, *Aspergillus nidulans* (Maggio-Hall and Keller 2004). The question arises whether other

fungi have mitochondrial beta oxidation, and which role it plays in their energy metabolism. We will address this question in Chapter 2.

3.2 Acyl-CoA dehydrogenase

Investigation of metabolic pathways relies primarily on the study of participating enzymes. Acyl-CoA dehydrogenase (ACAD) is the enzyme catalyzing the first reaction in each cycle of the mitochondrial beta oxidation spiral (Figure 10). ACAD enzymes are found in all three domains of life, with a single protein in *Escherichia coli*, but a large family of 11 members in human. The 11 members are not functionally redundant. Each has its own substrate preference for fatty acids with different length.

The majority of beta oxidation disorders in human are due to ACAD deficiency. Several clinic phenotypes caused by ACAD subfamily deficiency have been identified (Gregersen, Bross et al. 2004). Their clinical presentation is heterogeneous, but most of the cases have onset in infancy or early age. Symptoms range from muscle soreness/weakness, hypoglycemia, and coma, to sudden death. For example, deficiency of medium-chain acyl-CoA dehydrogenase, if untreated, leads to a death rate of 1:5 to 1:4.

What is known about ACAD enzymes, as well as mitochondrial beta oxidation as a whole, derives mainly from mammals. In other species, the ACAD family has not been well characterized, and the substrate preference remains largely unclear. Revealing different profiles of ACAD subfamilies in various species will shed light on the diversity of mitochondrial fatty acid catabolism across taxa.

Starting from the well-defined ACAD subfamilies in a few model species, identification of subfamily members in other taxa requires homology detection. Sequences are homologs if they share common ancestors. More precisely, two homologous genes are orthologs if they originate from speciation event; they are paralogs if they originate from a gene duplication event. Usually these two scenarios can be distinguished by phylogenetic analysis. Orthology detection is an important topic in bioinformatics, as orthologs have often similar function so that orthology can be used for function annotation. A caveat of function inference from homology is that the lack of homologs does not mean the corresponding gene function is absent in the species. The function can be carried out by a remote homolog with fast evolving sequence obscuring similarity, or by a non-homologous gene with the same function.

A number of methods exist for orthology inference, from basic sequence similarity comparison by BLAST (Altschul, Gish et al. 1990) and FASTA (Pearson 1990), more complicated bidirectional best-hit (Remm, Storm et al. 2001; Altenhoff and Dessimoz 2009) and reciprocal smallest distance calculation (Deluca, Wu et al. 2006), to sophisticated clustering (Remm, Storm et al. 2001; Li, Stoeckert et al. 2003) and tree construction algorithms (Wicker, Perrin et al. 2001; Storm and Sonnhammer 2002; Arvestad, Berglund et al. 2003). Sequence similarity search for orthologs is fast and straightforward, but it often misses the evolutionary context and mixes orthologs with paralogs. Strictly speaking, phylogenetic inference is indispensable for ortholog detection. But since evolution scenarios can be very complicated, phylogenetic construction, in particular when automated,

may lead to incorrect grouping of candidates as orthologs (Altenhoff and Dessimoz 2009). In Chapter 3, we described phylogenetic analyses with manual identification and removal of paralogs for detecting ACAD subfamilies across eukaryotes.

Objectives

As exemplified in the introduction, our current knowledge concerning the mitochondrial proteome is still limited. To advance the understanding of this important organelle, my thesis research focuses on two main issues:

1. the development of tools that effectively recognize mitochondrial proteins from genomic data, or alternatively, from EST data
2. application of the prediction tools for identification and analysis of mitochondrial proteins and the processes in which these proteins take part.

Chapter 1 Mitochondrial protein prediction by integrating heterogeneous tools

Available subcellular localization predictors are built with different training data, computational methods, and sequence features. This diversity leads to a bias towards certain types of sequences, as well as contradictory predictions from different tools. Both empirical and theoretical studies suggest that ensembles of individual predictors are often more accurate than the individual ones (Valentini and Masulli 2002; Assfalg, Gong et al. 2009). We applied the decision tree method to select and integrate available predictors. Prior knowledge about protein targeting is incorporated in the decision tree to enhance the performance.

BMC Bioinformatics

BioMed Central

Methodology article

Open Access

'Unite and conquer': enhanced prediction of protein subcellular localization by integrating multiple specialized tools

Yao Qing Shen* and Gertraud Burger

Address: Robert Cedergren Center for Bioinformatics and Genomics, Biochemistry Department, Université de Montréal, 2900 Edouard-Montpetit, Montreal, QC, H3T 1J4, Canada

* Corresponding author

Published: 29 October 2007

BMC Bioinformatics 2007, **8**:420 doi:10.1186/1471-2105-8-420

Received: 11 June 2007

Accepted: 29 October 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/420

Abstract

Background: Knowing the subcellular location of proteins provides clues to their function as well as the interconnectivity of biological processes. Dozens of tools are available for predicting protein location in the eukaryotic cell. Each tool performs well on certain data sets, but their predictions often disagree for a given protein. Since the individual tools each have particular strengths, we set out to integrate them in a way that optimally exploits their potential. The method we present here is applicable to various subcellular locations, but tailored for predicting whether or not a protein is localized in mitochondria. Knowledge of the mitochondrial proteome is relevant to understanding the role of this organelle in global cellular processes.

Results: In order to develop a method for enhanced prediction of subcellular localization, we integrated the outputs of available localization prediction tools by several strategies, and tested the performance of each strategy with known mitochondrial proteins. The accuracy obtained (up to 92%) surpasses by far the individual tools. The method of integration proved crucial to the performance. For the prediction of mitochondrion-located proteins, integration via a two-layer decision tree clearly outperforms simpler methods, as it allows emphasis of biologically relevant features such as the mitochondrial targeting peptide and transmembrane domains.

Conclusion: We developed an approach that enhances the prediction accuracy of mitochondrial proteins by uniting the strength of specialized tools. The combination of machine-learning based integration with biological expert knowledge leads to improved performance. This approach also alleviates the conundrum of how to choose between conflicting predictions. Our approach is easy to implement, and applicable to predicting subcellular locations other than mitochondria, as well as other biological features. For a trial of our approach, we provide a webservice for mitochondrial protein prediction (named YimLOC), which can be accessed through the AnaBench suite at http://anabench.bcm.umontreal.ca/anabench/. The source code is provided in the Additional File 2.

Background

The eukaryotic cell is highly organized: various biological processes are associated with specialized subcellular structures (such as protein export across the cell membrane), or confined to particular compartments (e.g., respiration in mitochondria). Subcellular location provides important clues about a protein's function and this knowledge is therefore used to assist in the annotation of newly dis-

covered or sequence-inferred proteins. On the other hand, the location of proteins with known function unravels where the corresponding biological processes take place and how they are connected amongst each other. Proteomics and microscopic detection of tagged or labelled proteins are powerful experimental approaches for determining protein localization. However, for most species, these approaches are costly in time and expense, and so there is a need for *in silico* prediction. A plethora of bioinformatic prediction methods have been developed in the past [1-21], and a dozen or so computational tools are publicly available (for a review see [22]). Most of these tools employ machine learning methods, i.e., they learn location-specific sequence features from known examples, and then extrapolate the learned rules to make predictions for proteins of unknown locations.

The targeting peptide, a conserved sequence motif usually located at the N-terminus of proteins, is a widely used sequence feature to identify a protein's location within the cell. This signal interacts with the import machineries of organelles such as mitochondria, chloroplasts and the endoplasmic reticulum. A number of tools use this signal for identifying proteins imported into organelles, notably **MitoProt** [23], **TargetP** [24], **iPSORT** [25], **Protein Prowler** [26], **Signal-CF** [27], and **Predotar** [28]. However, some organelle-imported proteins lack a N-terminal targeting peptide (e.g., the ADP/ATP carrier that is embedded in the inner mitochondrial membrane [29]) and therefore remain undetected by the tools above. In addition, application of these tools for genome-sequence-inferred proteins is limited, because the N-terminus of hypothetical proteins is often uncertain.

Another approach to identifying protein localization is based on sequence similarity with proteins of known location. For instance, a protein which shares a high similarity with a mitochondrial NADH:ubiquinone oxidoreductase subunit is very likely located in mitochondria. Sequence similarity combined with text annotation is used, for example, by the web-server 'Proteome Analyst Specialized Subcellular Localization Server' (**PASUB**) [30]. **PSLT** [31] predicts protein localization by searching for particular protein motifs and membrane domains. The underlying assumption is that proteins belonging to the same compartment share common domains. Both sequence-similarity-based and domain-based predictions have the limitation of depending on the existence of known homologs or known domains.

Several prediction tools do not rely on sequence similarity to known proteins or domains, but instead exploit a protein's amino acid composition and biochemical properties. **Subloc** [32], for instance, classifies proteins according to amino acid frequency, while **CELLO** [33] uses ungapped and gapped amino acid pair composition.

Certain tools combine several inherent sequence features and some also include textual information. For example, **ESLpred** [34] uses n-peptide composition and physicochemical properties, together with PSI-BLAST results. **pTARGET** [35] calculates scores based on the occurrence pattern of Pfam domains [36] and amino acid composition. **SherLoc** [37] exploits amino acid composition, targeting peptides, and motifs, as well as annotation and text description drawn from the literature or SwissProt entries.

It has been shown before that combining various prediction methods often yields better accuracy than the individual methods [38]. In fact, several of the above mentioned tools integrate different classifiers. **CELLO** [33], for instance, employs a two-level support vector machine (SVM) classification system. The first level builds individual SVM classifiers, one each for n-peptide composition, gapped-dipeptide composition, and so on. Each of these classifiers generates a probability distribution, which is then processed by a second-level SVM to calculate the final probability for a protein to belong in a certain subcellular location. The second-level SVM achieves a notably higher accuracy than the individual first-level classifiers. Similarly, **SherLoc** [37] uses the output vectors of different sequence-based classifiers and a text-based classifier as input for the final SVM classifier. An alternative approach builds Bayesian classifiers based on Markov chains, and constructs a hierarchical ensemble of these classifiers [39].

Each of the available localization prediction tools (subsequently referred to as LOC-tools) has different strength, and no tool is clearly and globally optimal. Any given LOC-tool performs well on certain data but poorly on others, and often predictions by different tools disagree (see examples in Table 1). This is not surprising, because LOC-tools employ different machine learning algorithms, sequence features, and training data.

This report introduces a comprehensive and simple system for protein location prediction. Following the maxim '*unite* and conquer', our approach combines the complementary strengths of existing prediction methods. Using the example of mitochondrial location, we integrated heterogeneous localization predictors by different strategies, tested performance with known data and selected the most efficient way of integration. The presented methodology is readily applicable to proteins from subcellular locations other than mitochondria, and even to the prediction of other biological features for which multiple, heterogeneous tools exist.

Table 1: Examples of conflicting results from individual prediction tools

Sequence ID[1]	Experimentally verified location	Predictions of mitochondrial location by individual LOC-tools[2,3]								
		TargetP	Subloc	pTARGET	SherLoc	Predotar	MitoProt	CELLO	PProwler	PASUB
YOR297C	Mitochondria	mit	mit	mit	non	non	mit	non	non	mit
YDR378C	Nucleus	mit	mit	non	non	non	mit	mit	non	non

[1] The example sequences are retrieved from the yeast genome database [52]
[2] For references see text
[3] "mit", predicted as mitochondrial protein; "non", predicted as non-mitochondrial protein

Results

As described in the Method section, we collected ~1,000 yeast proteins, ~1,000 *Arabidopsis* proteins, and ~3,000 human proteins of known subcellular location. Figure 1 shows the performance of nine individual LOC-tools on these data sets: TargetP, Subloc, SherLoc, pTARGET, Predotar, PProwler, PASUB, MitoProt, and CELLO. In the subsequent step, the predictions of these heterogeneous tools were integrated by different strategies. We employed the same procedure for all three datasets. Here, we show the results for yeast; those for *Arabidopsis* and human are given in Additional File 1.

Integration of LOC-tool predictions by grouping and majority-win voting

We formed 502 different groups ("voting groups") from nine individual LOC-tools. The predictions of the tools within each group are integrated by majority-win voting (see Methods section). Figure 1 (dots) shows that the performance on mitochondrial proteins varies greatly among the groups (see also Additional File 1: Figures S1 – S2). While the False Positive Rate (FPR) is generally low (< 0.05), the True Positive Rate (TPR) varies from 0.26 to 0.75. The best result is produced by the voting group pTARGET+PASUB+CELLO (TPR: 0.75, FPR: 0.02), but PASUB alone performs nearly as well (TPR: 0.74, FPR: 0.05). Thus, the gain of integration by majority-win voting is only moderate.

Integration of LOC-tool predictions by decision tree

For integration by decision trees, we took the predictions of the LOC-tools as input to construct classifiers by the C4.5 algorithm [40]. A total of six different decision trees were built as summarized in Table 2. First, outputs of all LOC-tools were employed as equivalent attributes. The resulting decision tree (referred to as LOC-DT, Figure 2a) recognizes mitochondrial proteins with an average TPR of 0.86 and FPR of 0.07, as evaluated by the ten-fold cross validation test (Figure 1, open symbols; Additional File 1: Figures S1 – S2). Note that the decision tree classifiers did not retain all the LOC-tools provided in the training process. The elimination of a given tool is due either to redundancy or to low accuracy such that its inclusion would cause performance to deteriorate.

Second, we introduced biological expert knowledge into the construction of decision trees. The mitochondrial targeting peptide (MTP) is a feature exclusive to mitochondrial proteins, and four LOC-tools rely on it to make predictions. In order to better exploit this feature, we implemented a decision tree integrating four MTP-based tools used in this study, notably TargetP, MitoProt, Predotar and PProwler. The output of this decision tree (referred to as MTP-DT) was then combined with the other five tools by constructing a stacked decision tree (STACK-DT; Figure 2b). As expected, stacking results in a major performance increase with a TPR of 0.9 and FPR of 0.04.

Effect of including transmembrane domain prediction tools

We realized that LOC-tools recognize membrane proteins less efficiently than matrix proteins (Figure 3). To alleviate this shortcoming, we integrated the LOC-tools with four additional tools that predict transmembrane domains (MEM-tools), i.e., **Phobius** [41], **TMHMM** [42], **HMMTOP** [43], and **SOSUI** [44]. The decision trees incorporating MEM-tools and LOC-tools are termed LOC-mem-DT, MTP-mem-DT and STACK-mem-DT (see Table 2).

Figure 3 shows that the integration of MEM-tools with LOC-tools clearly improves the recognition of mitochondrial membrane proteins. It should be noted that such improvement is not directly reflected in the overall performance, because mitochondrial membrane proteins account for only ~10% of our dataset.

Out of the six decision trees described above, STACK-mem-DT displays by far the best performance. Compared with the best individual LOC-tool and the best voting group (see above), STACK-mem-DT excels particularly in its high TPR (Table 3). This result was obtained from a dataset clustered at a cutoff of 80% sequence identity (data_C80). We repeated these experiments with datasets clustered more stringently at a 25% sequence identity cut-

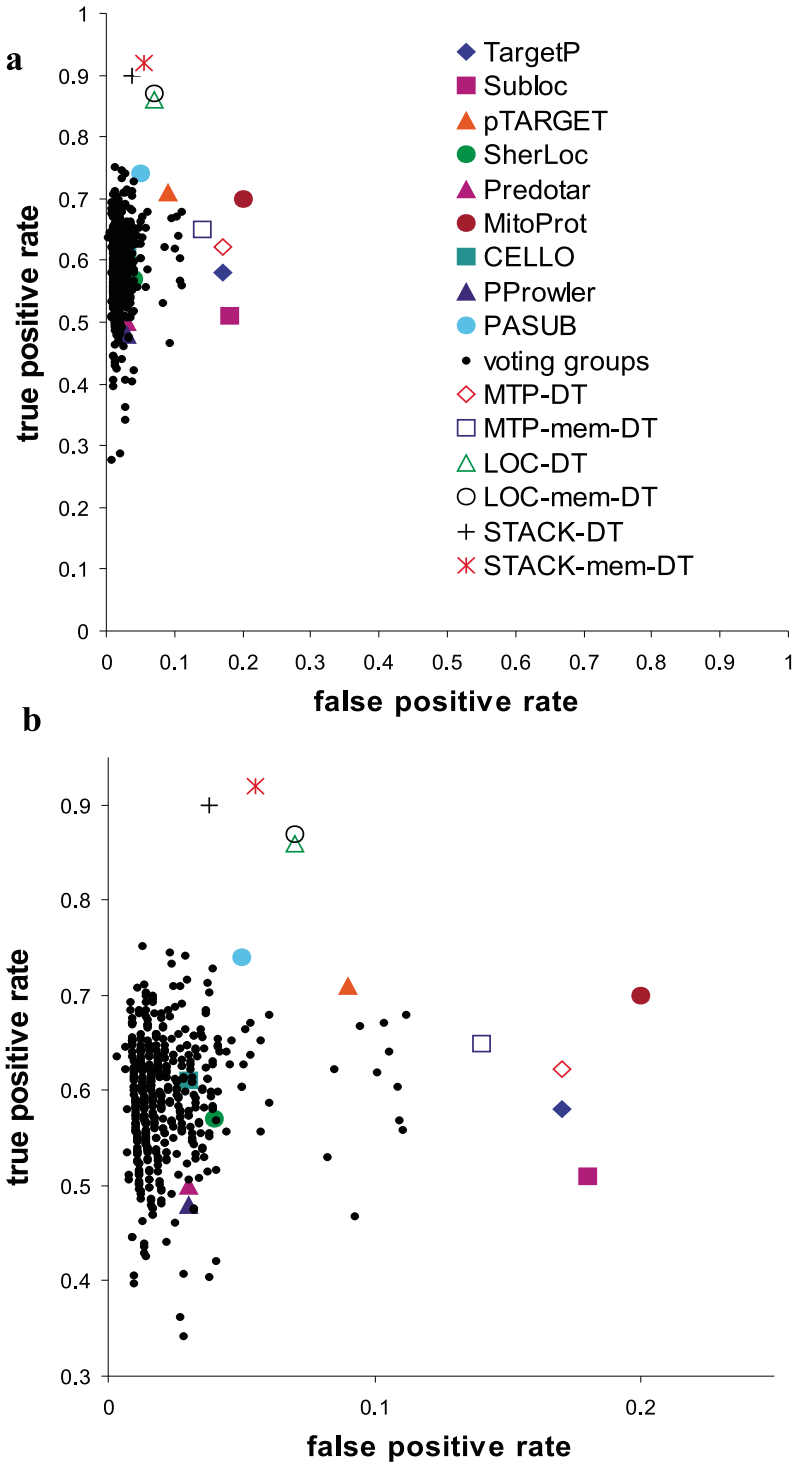


Figure 1
Prediction performance of individual and integrated tools on yeast mitochondrial proteins. **Filled symbols**: individual LOC-tools; **Dots**: voting groups (tools integrated by majority-win voting); **Open symbols**: decision trees. The desired results are located in the top left of the plot area, representing high true positive rate and low false positive rate. **a**, the result shown at full scale. **b**, the zoom-in of the region with false positive rate 0~0.25, and true positive rate 0.3~0.95.

Table 2: Decision trees built in this study and the individual tools employed to construct each tree[a]

Decision trees	LOC-tools									MEM-tools			
	TargetP	Predotar	MitoProt	PProwler	CELLO	Subloc	pTARGET	SherLoc	PASUB	Phobius	TMHMM	HMMTOP	SOSUI
LOC-DT	X	X	X	X	X	X	X	X	X				
MTP-DT	X	X	X	X									
STACK-DT	MTP-DT				X	X	X	X	X				
LOC-mem-DT	X	X	X	X	X	X	X	X	X	X	X	X	X
MTP-mem-DT	X	X	X	X						X	X	X	X
STACK-mem-DT	MTP-DT				X	X	X	X	X	X	X	X	X

[a] "X", if the tool is included in the decision tree listed in the leftmost column (for the references see text)

off (data_C25, Additional File 1: Table S2). The outcome was essentically the same as with data_C80 (Additional File 1: Table S3), which means that the good performance of STACK-mem-DT is not a result of data redundancy.

We were concerned that this superior performance was caused by a 20~50% overlap of our yeast data and the training data of individual LOC-tools. Therefore, we constructed a data subset, excluding proteins present in, or similar to, the training data of any LOC-tool, to build new decision trees. The result shows that the superior performance of STACK-mem-DT over both individual LOC-tools and majority-win voting is retained with this subset (Additional File 1: Figure S3).

To dissect how STACK-mem-DT makes its predictions, we followed the specific decision paths of the mitochondrial and nuclear proteins listed in Table 1, proteins that individual tools predict conflictingly. The mitochondrial protein follows a path down to SherLoc with all three predictions being wrong (Figure 4a). But in the end, the decision tree recognizes the mitochondrial location due to the two correct predictions made by pTARGET and PASUB. Similarly, the nuclear protein is first wrongly classified by CELLO, but the subsequent steps of the path identify its true location.

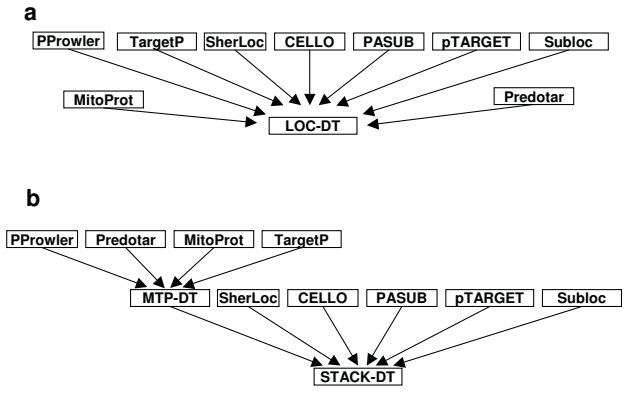


Figure 2
Integration of heterogeneous prediction tools by decision trees. **a**, The LOC-DT was built with outputs from nine LOC-tools. **b**, The MTP-DT was built with outputs from four tools whose prediction is based on the mitochondrial targeting peptide. The output of MTP-DT, together with the outputs of five other LOC-tools, was used to construct the STACK-DT.

Finally, we inspected the paths of three other proteins, constituents of the mitochondrial outer membrane, the plasma membrane and the nucleus, respectively. All of these proteins cannot be distinguished by the individual LOC-tools (Table 4), nor by trees without MEM-tools. STACK-mem-DT correctly classifies all three proteins due to the final two steps in the tree that employ MEM-tools (Figure 4a, coloured line).

Implementation

STACK-mem-DT was implemented as a webservice, YimLOC, accessible via the public bioinformatics workbench AnaBench [45]. The current version takes the prediction results from individual tools as input, and outputs the prediction for a protein to be mitochondrion-localized or not. For thorough analyses, we recommend that users build the decision tree on their local computer, with their own training data and choice of individual LOC-tools. The source code is available under the GNU licence.

Discussion

The purpose of this study was to enhance prediction accuracy by integrating the available subcellular localization prediction tools. Successful integration of specialized tools takes advantage of their complementary strengths, which are drawn from three sources: the different sequence features the tools exploit, the different computational algorithms they employ, and the different training sets they are built from.

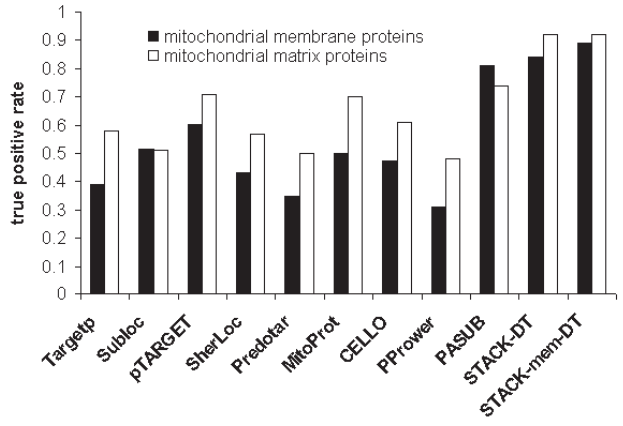


Figure 3
Prediction performance of individual and integrated tools on yeast mitochondrial membrane and matrix proteins. Loc-tools recognize mitochondrial membrane proteins less efficiently than matrix proteins. The effectiveness of PASUB is due to the fact that it exploits annotations and that the portion of *annotated* mitochondrial membrane proteins is higher compared to matrix proteins.

Integration by decision tree outperforms group voting

The best performance obtained from majority-win voting of LOC-tool groups shows almost the same TPR as the best individual LOC-tool (PASUB in this case), with a slightly lower FPR. Some of the voting groups yield even lower TPRs than individual LOC-tools. In contrast, decision tree classifiers built from the ensemble of LOC-tools all outperform the individual tools as well as any of the majority-win voting combinations (see Figure 1. Note that MTP-DT and MTP-mem-DT are special cases as they were given only a subset of LOC-tools for training.). The most effective of the presented integrative predictors is STACK-mem-DT, which exceeds by far the performance of the best LOC-tool (TPR of 0.92 compared to 0.75, with the same FPR of 0.05; Table 3). Yet, for fairness, it should be stressed that many of the tools have been developed with the aim of predicting multiple locations, while we optimize here mitochondrial location.

A fair and rigorous comparison of YimLOC with all other prediction methods should use the same test data, as we did for the comparison of YimLOC with nine LOC-tools shown in Figure 1, and in Additional File 1: Figures S1 – S2. Unfortunately, this is not feasible for some prediction methods because of several reasons: the training data are not provided; there are no webservices or software distributions available; the webservices are available but not tuned for large-scale predictions.

Among the various machine leaning methods, we chose here decision trees for integration because they have the advantage that they allow tracing back how the predictions are made, and thus may provide a biological meaningful interpretation of the predictions. Note that for the more complex problem of predicting proteins targeted to multiple subcellular locations [4-6], neural network or Naïve Bayes would be more appropriate than decision trees, because they allow handling of prediction probabilities in a flexible manner.

Trade-off between sensitivity and specificity

For any given prediction method, an increase of the TPR is usually accompanied by an increase of the FPR. How to balance the two rates depends on the purpose of the prediction. If biologists wish to identify *all* mitochondrial proteins from a whole genome sequence, they should choose a prediction method with highest TPR (in this study the STACK-mem-DT). On the other hand, if the purpose is to determine the subcellular localization of a few candidate proteins of interest, a prediction method with lowest FPR should be favoured (in this study the combination of pTARGET+PASUB+CELLO).

Table 3: Performance[1] of the best predictors for the three different prediction schemes

Classes[2]		Individual tool (PASUB)			Combination of tools by voting[3]			Decision tree classifier (STACK-mem-DT)		
		TPR	FPR	ACC	TPR	FPR	ACC	TPR	FPR	ACC
Yeast	Mit	0.74	0.05	0.69	0.75	0.02	0.84	0.92	0.05	0.95
	Non	0.65	0.06		0.99	0.20		0.97	0.05	
Arabidopsis	Mit	0.75	0.09	0.81	0.67	0.07	0.88	0.87	0.12	0.94
	Non	0.83	0.05		0.95	0.09		0.96	0.04	
Human	Mit	0.87	0.09	0.68	0.88	0.01	0.97	0.90	0.02	0.99
	Non	0.65	0.02		0.98	0.02		0.99	0.01	

[1] TPR: true positive rate; FPR: false positive rate; ACC: accuracy (all correctly predicted instances/all instances)
[2] Mit: mitochondrial proteins; Non: proteins of other subcellular locations
[3] The best combination of tools is pTARGET+PASUB+CELLO for yeast data, PASUB+MitoPort+CELLO for *Arabidopsis* data, and pTARGET +SherLoc+ PASUB for human data

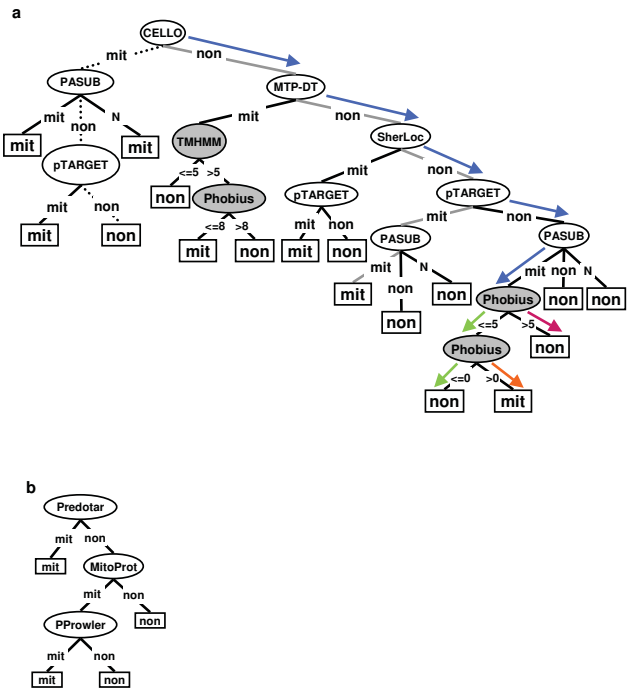


Figure 4

Decision tree topology for the prediction of mitochondrial proteins. **a**, STACK-mem-DT; **b**, MTP-DT. The trees were built by C4.5 (see Methods). Each oval represents a prediction tool. **Filled ovals** represent transmembrane domain predictors. **Rectangle** represents a decision: "mit" for mitochondrial proteins and "non" for proteins of other subcellular locations. If a tool predicts the query protein as a mitochondrial protein, the branch (edge) is labeled "mit"; otherwise "non". If PASUB makes no prediction, the branch is labeled "N". Several decision-making paths are highlighted, as follows: **Dotted line**: for non-mitochondrial protein YDR378C. **Grey line**: for mitochondrial protein YOR297C. **Blue arrow**: the common path for three differently localized proteins: mitochondrial (YIL065C), plasma membrane (YBR069C) and nuclear (YLL022C). **Orange arrow**: for mitochondrial protein YIL065C. **Red arrow**: for non-mitochondrial protein YBR069C. **Green arrow**: for non-mitochondrial protein YLL022C.

Making use of prior biological knowledge

During decision tree construction, LOC-tools are retained if they have a good overall performance on the training data. In this process, all tools (and therefore the sequence features exploited) are considered of equal importance. To further enhance performance, we put more emphasis on certain tools based on domain-specific knowledge. In particular, the mitochondrial targeting peptide (MTP) is specific to proteins imported into mitochondria, but not all mitochondrial proteins possess one. Therefore, a tool that recognizes mitochondrial proteins based on the presence of MTP has high specificity (a protein with MTP is reliably targeted to mitochondria), but low sensitivity (mitochondrial proteins without MTP cannot be recognized). We employed four MTP-based tools in this study. Yet, LOC-DT retained only one of them, although the other three tools may be complementary in recognizing the various instances.

Since the targeting peptide is known to be an important determinant of protein localization, but not necessarily rewarded by decision trees, we modified the training process to make use of this external knowledge. This was achieved by a two-layer decision tree (STACK-DT, see Figure 2b). Indeed, STACK-DT performes significantly better than LOC-DT (see Figure 1, "+"), testifying to the value of incorporating expert knowledge in decision tree construction.

Inclusion of transmembrane domain prediction

We observed that LOC-tools often misclassified mitochondrial membrane proteins (Figure 3). This may be due to several reasons: (i) the training sets of some tools do not include mitochondrial membrane proteins (e.g., Subloc); (ii) mitochondrial membrane proteins typically lack a targeting peptide, while MTP-based tools rely on the presence of this signal [46]; and (iii) tools based on amino acid composition and physicochemical properties may confuse mitochondrial membrane proteins with membrane proteins from other subcellular compartments. We have addressed these limitations by building decision tree classifiers that integrate predictions of both subcellular localization and transmembrane domains. In fact, information on the number of such domains boosts recognition of mitochondrial membrane proteins from 81% to 89% (Figure 3).

Conclusion

This study devises a simple, practical and highly effective approach to exploiting complementary bioinformatics tools by integration through machine learning. Using mitochondrial location as a test case, we observe that tool integration with decision trees significantly improves prediction accuracy compared to individual tools or their simple combination. Inclusion of biological expert knowledge in machine learning further enhances the performance. Particularly improved is prediction of membrane proteins, which is notoriously difficult. Further, our approach alleviates the conundrum of how to choose between conflicting predictions from different LOC-tools. The methodology is easy to implement and applicable to the prediction of other biological feature for which multiple, heterogeneous tools exist.

Table 4: Example proteins used for decision tracing

Sequence ID[1]	Experimentally verified location	Predictions of mitochondrial location by individual LOC-tools[2,3]								Predicted number of transmembrane domains[2,3]			
		TargetP	Subloc	pTARGET	SherLoc	Predotar	MitoProt	CELLO	PProwler	PASUB	Phobius	TMHMM	HMMTOP	SOSUI
YIL065C	Mitochondrial outer membrane	non	non	non	non	non	non	non	non	mit	1	1	1	1
YBR069C	Plasma membrane	non	non	non	non	non	non	non	non	mit	12	12	12	12
YLL022C	Nucleus	non	non	non	non	non	non	non	non	mit	0	0	0	0

[1] The sequences are retrieved from the yeast genome database [52]

[2] For references see text

[3] "mit", predicted as mitochondrial protein; "non", predicted as non-mitochondrial protein

Methods

Data set

Protein sequences from yeast in Swiss-Prot release 50.3 were selected by the following criteria: 1) they are encoded in the nucleus; 2) their subcellular location is experimentally verified; and 3) the localization annotation is not ambiguous (i.e., terms like "probable" or "possible" are absent from their annotation of subcellular localization). In addition, we retrieved 522 yeast mitochondrial protein sequences from MITOP2 [47], a manually curated database of nucleus-encoded mitochondrial proteins with experimental evidence. Sequences having identities over 80% were clustered by Cd-hit [48] to reduce data redundancy. The final yeast dataset contains 503 mitochondrial and 872 non-mitochondrial proteins.

In a similar way, *Arabidopsis* and human protein sequences from Swiss-Prot were collected. The *Arabidopsis* dataset was enriched by sequences from AMPDB [49], a database for *Arabidopsis* mitochondrial proteins. After being clustered with 80% sequence identity, 193 mitochondrial and 608 non-mitochondrial proteins constitute the *Arabidopsis* dataset. The human dataset contains 353 mitochondrial and 2,679 non-mitochondrial proteins.

In addition, we further clustered the three datasets (yeast, *Arabidopsis*, and human) with the threshold of 25% sequence identity to build more stringent datasets (Additional File 1: Table S2).

To compile a dataset which does not overlap with the training data of the LOC-tools employed (see Table 2), we searched our yeast dataset against the training data of the nine LOC-tools with BLAST. A protein was removed from the yeast data if it had >80% identity to a protein in the training set of any LOC-tool. The remaining proteins constitute a non-overlapping subset of yeast data, which contains 190 mitochondrial and 344 non-mitochondrial proteins.

Integration of heterogeneous tools

a Prediction by individual tools

We selected nine prediction tools for subcellular localization: TargetP [24], Subloc [32], SherLoc [37], pTARGET [35], Predotar [28], Protein Prowler (PProwler) [26], PASUB [30], MitoProt [23], and CELLO [33]. The selection was based on the diversity of the algorithms and the sequence features they employ. These tools were used as base-level classifiers, whose prediction results were combined to build new classifiers. Prediction results from most tools were obtained via web services. The only exception is MitoProt, which has been installed and run locally.

b Consistent representation of the output from heterogeneous LOC-tools

LOC-tools output a categorical prediction (mitochondria, cytoplasm, nucleus, etc.) for each query sequence. Predictions were converted to "mit" for mitochondrial location and "non" otherwise. A special case is PASUB, which makes no predictions for proteins that lack significant similarity to known sequences. In these cases, we issued "N".

Together with the categorical prediction, LOC-tools also output a positive numerical value indicating the confidence of prediction. The range of numerical values differs among LOC-tools. Intuitively, numerical encoding seems advantageous, since it reflects the confidence that LOC-tools have in their predictions. However, it also may introduce a hidden bias in the integration, because the various tools evaluate and measure confidence differently (Additional File 1: Table S1). For example, CELLO outputs a score (for example 2.064) to show the reliability that a protein is affiliated with each of 12 subcellular locations. In contrast, pTARGET distinguishes nine locations, and outputs the confidence value in the form of percentage (for example 98%). Since it is not straightforward to consolidate the particular confidence factors of the various LOC-tools, we decided to use categorical encoding.

c Integration of LOC-tools by grouping and voting

For nine LOC-tools, with group size from two to nine, there were a total of 502 different groups. Within each group, predictions of individual LOC-tools were combined with a majority-win voting scheme. A given sequence was regarded as a mitochondrial protein, if more than half of the combined tools assigned it to mitochondria. No prediction was made if there was a tie.

d Integration of LOC-tools by decision tree

For building decision trees, we used J4.8, a program based on the C4.5 algorithm [40], available in the Weka package [50]. Default parameters were employed. The individual LOC-tools and MEM-tools were used as attributes of input data, and the prediction results of each tool as attribute values.

The decision trees were evaluated by a ten-fold cross validation test, where the data set was equally divided into ten parts. Nine parts were combined to form the training set for building the decision tree, which was then evaluated by the remaining part. The process was repeated ten times. Alternatively, jackknife test can be employed for examining the power of a prediction method [1-3]. Although jackknife test is deemed the most rigorous and objective [51], it is time consuming, particularly for large datasets. Therefore, 10-fold cross validation is a good and wildly adopted alternative.

The performance of each prediction method was measured as true positive rate and false positive rate, where

true positive rate (**TPR**) = true positives/(true positives + false negatives), and

false positive rate (**FPR**) = false positives/(true positives + false positives).

Authors' contributions

GB conceived the study. YQS designed, developed and implemented the methods. GB participated in the design and supervised the process. YQS drafted the manuscript. Both authors approved the final manuscript.

Additional material

Additional file 2

This file contains scripts for the online server YimLOC. Please note that there scripts only codes for the ready-to-use STACK-mem-DT described in the main text. The scripts do not provide the training process.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-420-S2.pdf]

Additional file 1

This file contains figures and tables depicting the performance of different integration methods on Arabidopsis *data, human data, a non-overlapping subset of yeast data, and three more stringent datasets. The results were obtained in the same way as for the yeast data. This file also contains a table showing the range of numerical predictions from individual LOC-tools.* ***Figure S1 – Prediction performance of individual and integrated tools on* Arabidopsis *mitochondrial proteins****. Filled symbols****: individual LOC-tools;* ***Dots****: voting groups (tools integrated by majority-win voting);* ***Open symbols****: decision trees. The desired results are located in the top left of the plot area, representing high true positive rate and low false positive rate.* ***Figure S2 – Prediction performance of individual and integrated tools on human mitochondrial proteins****. Filled symbols: individual LOC-tools;* ***Dots****: voting groups (tools integrated by majority-win voting);* ***Open symbols****: decision trees. The desired results are located in the top left of the plot area, representing high true positive rate and low false positive rate.* ***Figure S3 – Prediction performance of individual and integrated tools on yeast data which does not overlap with the training data of any individual LOC-tool****. Filled symbols: individual LOC-tools;* ***Dots****: voting groups (tools integrated by majority-win voting);* ***Open symbols****: decision trees. The desired results are located in the top left of the plot area, representing high true positive rate and low false positive rate.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-420-S1.doc]

Acknowledgements

This work was supported by Genome-Canada and Genome-Quebec in the context of the Protist EST program (PEP). We would like to thank Sébastien Lemieux, Sivakumar Kannan and Amy Hauth for their helpful suggestions to this work. We also thank Sivakumar Kannan, Emmet O'Brien and Henner Brinkmann for improving the manuscript. YQS is a Canadian Institute for Health Research (CIHR) Strategic Training Fellow in Bioinformatics. GB is a member of the Canadian Institute for Advanced Research (CIAR), program in Evolutionary Biology, whom we thank for interaction support.

References

1. Pan YX, Zhang ZZ, Guo ZM, Feng GY, Huang ZD, He L: **Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach.** *J Protein Chem* 2003, **22:**395-402.
2. Chou KC, Shen HB: **Predicting protein subcellular location by fusing multiple classifiers.** *J Cell Biochem* 2006, **99:**517-527.
3. Shi JY, Zhang SW, Pan Q, Cheng YM, Xie J: **Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition.** *Amino Acids* 2007, **33:**69-74.
4. Shen HB, Chou KC: **Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites.** *Biochem Biophys Res Commun* 2007, **355:**1006-1011.
5. Chou KC, Shen HB: **Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites.** *J Proteome Res* 2007, **6:**1728-1734.
6. Chou KC, Cai YD: **Predicting protein localization in budding yeast.** *Bioinformatics* 2005, **21:**944-950.
7. Chen YL, Li QZ: **Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition.** *J Theor Biol* 2007, **248(2):**377-381.
8. Shen HB, Yang J, Chou KC: **Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction.** *Amino Acids* 2007, **33:**57-67.
9. Chou KC, Shen HB: **Large-scale plant protein subcellular location prediction.** *J Cell Biochem* 2007, **100:**665-678.
10. Zhang T, Ding Y, Chou KC: **Prediction of protein subcellular location using hydrophobic patterns of amino acid sequence.** *Comput Biol Chem* 2006, **30:**367-371.
11. Matsuda S, Vert JP, Saigo H, Ueda N, Toh H, Akutsu T: **A novel representation of protein sequences for prediction of subcellular location using support vector machines.** *Protein Sci* 2005, **14:**2804-2813.
12. Gao QB, Wang ZZ, Yan C, Du YH: **Prediction of protein subcellular location using a combined feature of sequence.** *FEBS Lett* 2005, **579:**3444-3448.
13. Chou KC, Cai YD: **Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition.** *J Cell Biochem* 2003, **90:**1250-1260.
14. Chou KC, Cai YD: **Using functional domain composition and support vector machines for prediction of protein subcellular location.** *J Biol Chem* 2002, **277:**45765-45769.
15. Cai YD, Liu XJ, Xu XB, Chou KC: **Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect.** *J Cell Biochem* 2002, **84:**343-348.
16. Cai YD, Liu XJ, Xu XB, Chou KC: **Support vector machines for prediction of protein subcellular location.** *Mol Cell Biol Res Commun* 2000, **4:**230-233.
17. Chou KC, Elrod DW: **Protein subcellular location prediction.** *Protein Eng* 1999, **12:**107-118.
18. Park KJ, Kanehisa M: **Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs.** *Bioinformatics* 2003, **19:**1656-1663.
19. Huang Y, Li Y: **Prediction of protein subcellular locations using fuzzy k-NN method.** *Bioinformatics* 2004, **20:**21-28.
20. Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC: **Using complexity measure factor to predict protein subcellular location.** *Amino Acids* 2005, **28:**57-61.
21. Chou KC SHB: **Recent progresses in protein subcellular location prediction.** *Analytical Biochemistry* 2007, **370:**1-16.
22. Donnes P, Hoglund A: **Predicting protein subcellular localization: past, present, and future.** *Genomics Proteomics Bioinformatics* 2004, **2:**209-215.
23. Claros MG, Vincens P: **Computational method to predict mitochondrially imported proteins and their targeting sequences.** *Eur J Biochem* 1996, **241:**779-786.

24. Emanuelsson O, Nielsen H, Brunak S, von Heijne G: **Predicting subcellular localization of proteins based on their N-terminal amino acid sequence.** *J Mol Biol* 2000, **300:**1005-1016.
25. Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S: **Extensive feature detection of N-terminal protein sorting signals.** *Bioinformatics* 2002, **18:**298-305.
26. Boden M, Hawkins J: **Prediction of subcellular localization using sequence-biased recurrent networks.** *Bioinformatics* 2005, **21:**2279-2286.
27. Chou KC, Shen HB: **Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides.** *Biochem Biophys Res Commun* 2007, **357:**633-640.
28. Small I, Peeters N, Legeai F, Lurin C: **Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences.** *Proteomics* 2004, **4:**1581-1590.
29. Wiedemann N, Pfanner N, Ryan MT: **The three modules of ADP/ATP carrier cooperate in receptor recruitment and translocation into mitochondria.** *EMBO J* 2001, **20:**951-960.
30. Lu Z, Szafron D, Greiner R, Lu P, Wishart DS, Poulin B, Anvik J, Macdonell C, Eisner R: **Predicting subcellular localization of proteins using machine-learned classifiers.** *Bioinformatics* 2004, **20:**547-556.
31. Scott MS, Thomas DY, Hallett MT: **Predicting subcellular localization via protein motif co-occurrence.** *Genome Res* 2004, **14:**1957-1966.
32. Hua S, Sun Z: **Support vector machine approach for protein subcellular localization prediction.** *Bioinformatics* 2001, **17:**721-728.
33. Yu CS, Chen YC, Lu CH, Hwang JK: **Prediction of protein subcellular localization.** *Proteins* 2006, **64(3):**643-651.
34. Bhasin M, Raghava GP: **ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST.** *Nucleic Acids Res* 2004, **32:**W414-9.
35. Guda C, Subramaniam S: **pTARGET [corrected] a new method for predicting protein subcellular localization in eukaryotes.** *Bioinformatics* 2005, **21:**3963-3969.
36. Guda C, Fahy E, Subramaniam S: **MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins.** *Bioinformatics* 2004, **20:**1785-1794.
37. Shatkay H, Hoglund A, Brady S, Blum T, Donnes P, Kohlbacher O: **SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data.** *Bioinformatics* 2007.
38. Džeroski S, Ženko B: **Is combining classifiers with stacking better than selecting the best one?** *Machine Learning* 2004, **54:**255-273.
39. Bulashevska A, Eils R: **Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains.** *BMC Bioinformatics* 2006, **7:**298.
40. Quinlan JR: **C4.5: programs for machine learning.** San Mateo, California, Morgan Kaufmann Publishers; 1993.
41. Kall L, Krogh A, Sonnhammer EL: **A combined transmembrane topology and signal peptide prediction method.** *J Mol Biol* 2004, **338:**1027-1036.
42. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.** *J Mol Biol* 2001, **305:**567-580.
43. Tusnady GE, Simon I: **The HMMTOP transmembrane topology prediction server.** *Bioinformatics* 2001, **17:**849-850.
44. Hirokawa T, Boon-Chieng S, Mitaku S: **SOSUI: classification and secondary structure prediction system for membrane proteins.** *Bioinformatics* 1998, **14:**378-379.
45. Badidi E, De Sousa C, Lang BF, Burger G: **AnaBench: a Web/CORBA-based workbench for biomolecular sequence analysis.** *BMC Bioinformatics* 2003, **4:**63.
46. Pfanner N, Wiedemann N, Meisinger C, Lithgow T: **Assembling the mitochondrial outer membrane.** *Nat Struct Mol Biol* 2004, **11:**1044-1048.
47. Andreoli C, Prokisch H, Hortnagel K, Mueller JC, Munsterkotter M, Scharfe C, Meitinger T: **MitoP2, an integrated database on mitochondrial proteins in yeast and man.** *Nucleic Acids Res* 2004, **32:**D459-62.
48. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics* 2006, **22:**1658-1659.
49. Heazlewood JL, Millar AH: **AMPDB: the Arabidopsis mitochondrial protein database.** *Nucleic Acids Res* 2005, **33:**D605-10.
50. Frank E, Hall M, Trigg L, Holmes G, Witten IH: **Data mining in bioinformatics using Weka.** *Bioinformatics* 2004, **20:**2479-2481.
51. Chou KC, Zhang CT: **Prediction of protein structural classes.** *Crit Rev Biochem Mol Biol* 1995, **30:**275-349.
52. **Saccharomyces Genome Database** [http://www.yeastgenome.org]

Supplementary figures and tables

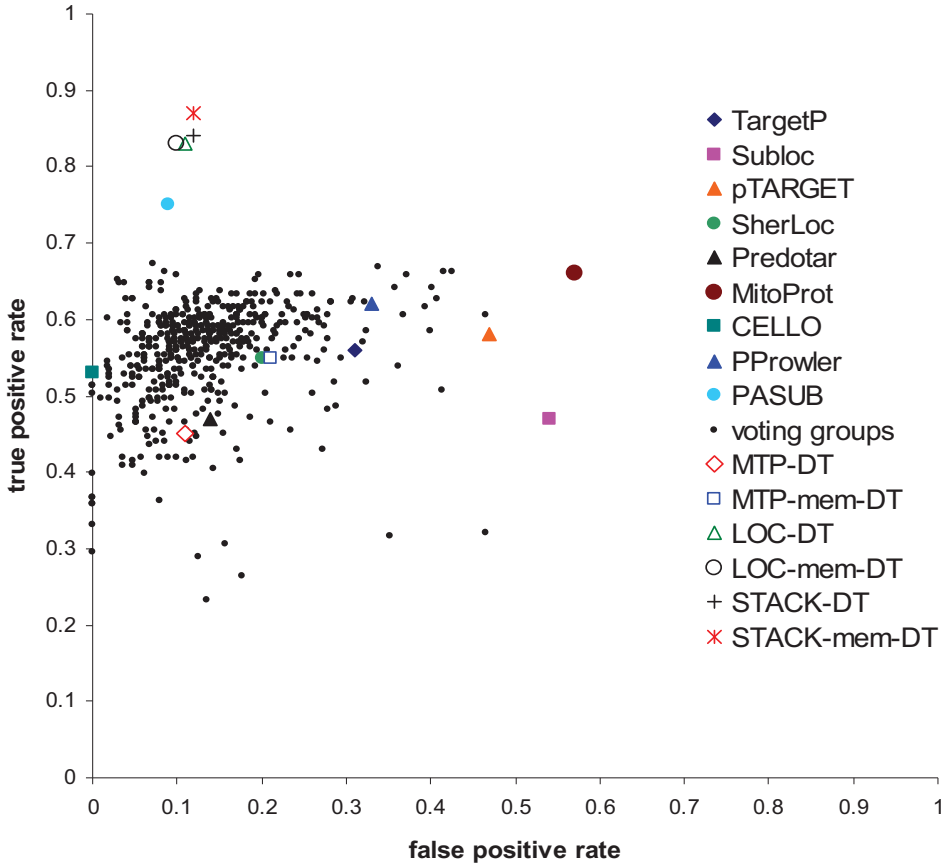


Figure S1 - Prediction performance of individual and integrated tools on *Arabidopsis* mitochondrial proteins.

Filled symbols: individual LOC-tools; **Dots**: voting groups (tools integrated by majority-win voting); **Open symbols**: decision trees. The desired results are located in the top left of the plot area, representing high true positive rate and low false positive rate.

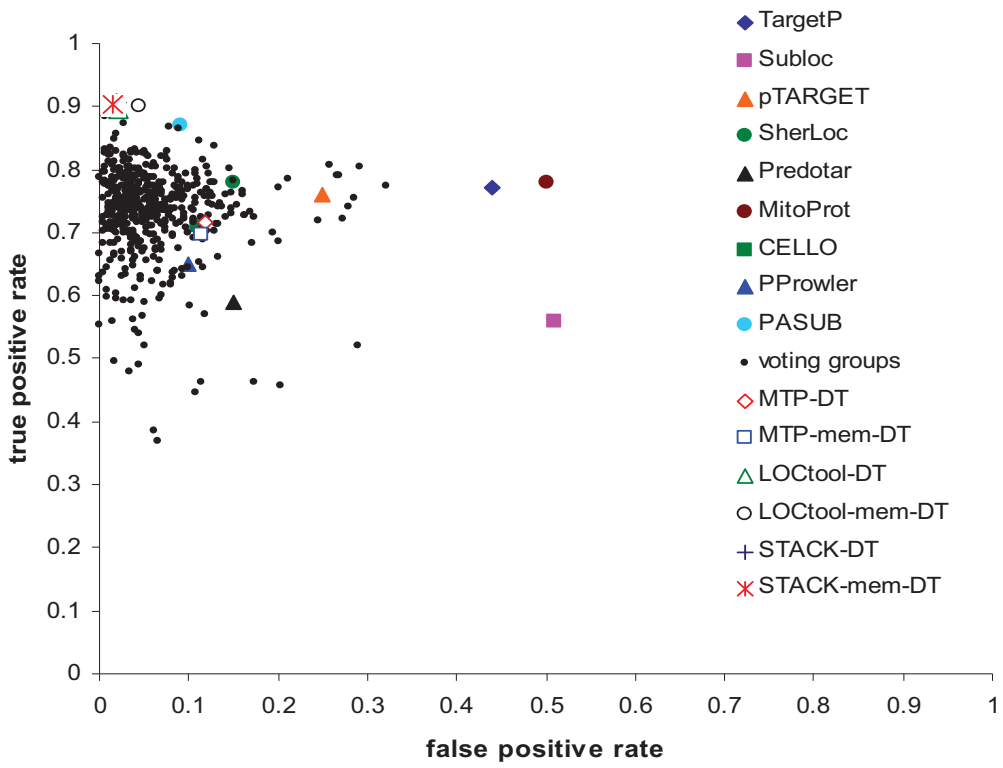


Figure S2 - Prediction performance of individual and integrated tools on human mitochondrial proteins.

Filled symbols: individual LOC-tools; **Dots**: voting groups (tools integrated by majority-win voting); **Open symbols**: decision trees. The desired results are located in the top left of the plot area, representing high true positive rate and low false positive rate.

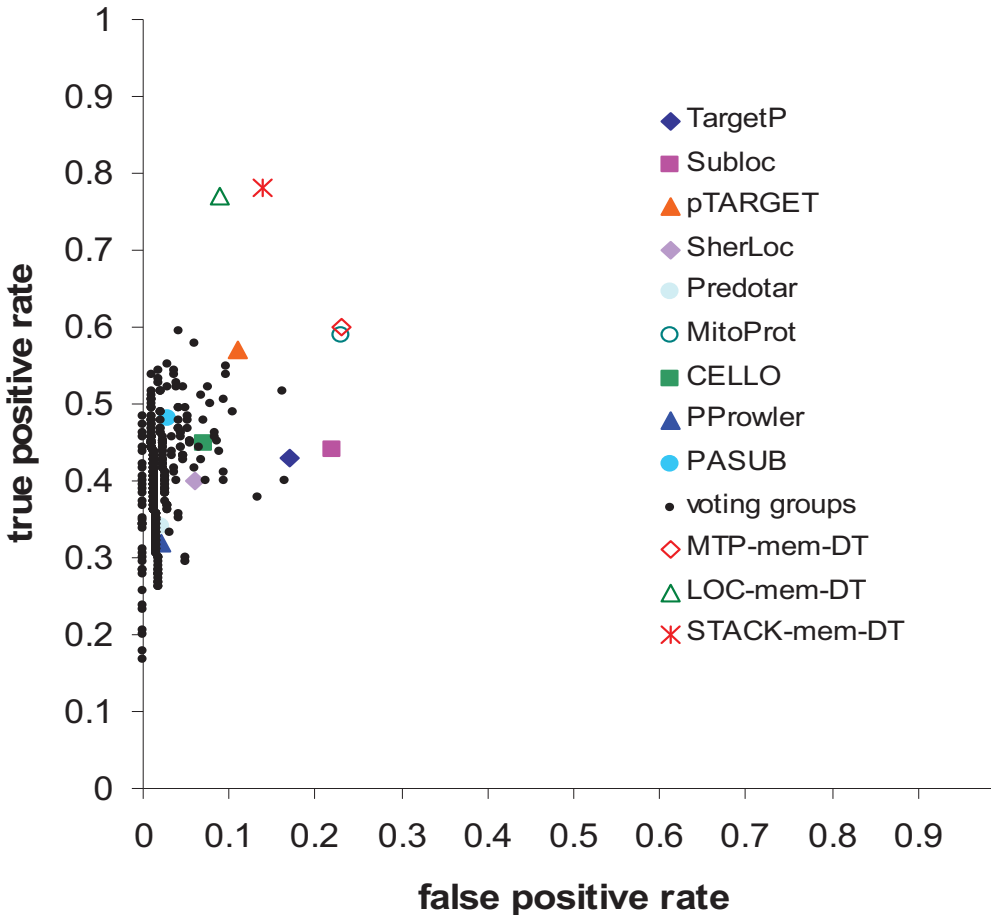


Figure S3 - Prediction performance of individual and integrated tools on yeast data without overlap with the training data of any individual LOC-tool.

Filled symbols: individual LOC-tools; **Dots**: voting groups (tools integrated by majority-win voting); **Open symbols**: decision trees. The desired results are located in the top left of the plot area, representing high true positive rate and low false positive rate.

Table S1 The number of predicted classes and the range of numerical prediction values from individual LOC-tools

	TargetP	Subloc	pTARGET	SherLoc	Predotar	MitoProt	CELLO	PProwler	PASUB
Number of classes	4	10	9	9	4	2	12	4	10
Value range	1-5	1-10	1%-100%	0-1	0-1	0-1	>0	0-1	0-1
Value of most reliable prediction	1	10	100%	1	1	1	The higher the better	1	1

Table S2 Number of instances in each dataset, after being clustered at threshold of 80% sequence identity and 25% sequence identity

	Yeast		*Arabidopsis*		Human	
Threshold at clustering	80%	25%	80%	25%	80%	25%
Mitochondrial proteins	503	446	193	158	353	290
Non-mitochondrial proteins	872	781	608	383	2679	1505
Total	1375	1227	802	541	3032	1795

Table S3 Performance[1] of the best predictors for the three different prediction schemes (for dataset clustered at threshold of 25% identity)

Classes[2]		**Individual tool (PASUB)**			**Combination of tools by voting[3]**			**Decision tree classifier (STACK-mem-DT)**		
		TPR	FPR	ACC	TPR	FPR	ACC	TPR	FPR	ACC
Yeast	Mit	0.72	0.05	0.69	0.73	0.04	0.89	0.85	0.06	0.93
	Non	0.67	0.06		0.98	0.13		0.97	0.08	
Arabidopsis	Mit	0.73	0.10	0.82	0.66	0.04	0.89	0.84	0.11	0.92
	Non	0.85	0.06		0.98	0.12		0.96	0.07	
Human	Mit	0.86	0.08	0.74	0.86	0.03	0.97	0.91	0.03	0.98
	Non	0.72	0.02		0.99	0.03		0.99	0.02	

[1] TPR: true positive rate; FPR: false positive rate; ACC: accuracy (all correctly predicted instances / all instances) [2] Mit: mitochondrial proteins; Non: proteins of other subcellular locations[3] The best combination of tools is pTARGET+Mitoprot+CELLO for yeast data, PASUB+Sherloc+CELLO for *Arabidopsis* data, and pTARGET +SherLoc+ PASUB+subloc+Mitoprot for human data

Chapter 2 Plasticity of a key metabolic pathway in fungi

The presence of mitochondrial beta oxidation in fungi has been elusive. Certain experiments demonstrated its absence, while others proved its presence in individual species. Further, observations from the few species examined could not be generalized to the whole kingdom. To address this problem, we searched computationally for enzymes of beta oxidation in ~60 fungi, and we used YimLOC to predict if these enzymes are localized in mitochondria.

Funct Integr Genomics
DOI 10.1007/s10142-008-0095-6

REVIEW

Plasticity of a key metabolic pathway in fungi

Yao-Qing Shen · Gertraud Burger

Received: 3 July 2008 / Revised: 12 August 2008 / Accepted: 17 August 2008

Abstract Beta oxidation is the principal metabolic pathway for fatty acid degradation. The pathway is virtually universally present throughout eukaryotes yet displays different forms in enzyme architecture, substrate specificity, and subcellular location. In this review, we examine beta oxidation across the fungal kingdom by conducting a large-scale *in silico* screen and localization prediction for all relevant enzymes in >50 species. The survey reveals that fungi exhibit an astounding diversity of beta oxidation pathways and shows that the combined presence of distinct mitochondrial and peroxisomal pathways is the prevailing and likely ancestral type of beta oxidation in fungi. In addition, the available information indicates that the mitochondrial pathway was lost in the common ancestor of *Saccharomycetes*. Finally, we infer the existence of a hybrid peroxisomal pathway in several *Sordariomycetes*, including *Neurospora crassa*. In these cases, a typically mitochondrion-located enzyme compensates for the lack of a peroxisomal one.

Keywords Beta oxidation · Mitochondrion · Peroxisome · Metabolic compartmentalization

Electronic supplementary material The online version of this article (doi:10.1007/s10142-008-0095-6) contains supplementary material, which is available to authorized users.

Y.-Q. Shen (✉) · G. Burger
Robert Cedergren Center for Bioinformatics and Genomics,
Biochemistry Department, Université de Montréal,
2900 Edouard-Montpetit,
Montreal, QC H3T 1J4, Canada

Introduction

Fatty acids play multiple important roles in the living cell. They are the building blocks of the cell membrane, regulate enzymes and membrane channels, and serve as signaling molecules and precursors for hormones. Most importantly, they store and provide energy (Poirier et al. 2006). Fatty acids can be utilized as a sole carbon source by numerous species. The principal pathway for fatty acid degradation is beta oxidation, by which molecules are broken down in a repeating spiral of four steps. Each spiral removes two carbons, in the form of acetyl-CoA, from the fatty acid chain (Fig. 1). While taxonomically virtually ubiquitous throughout eukaryotes, beta oxidation displays a startling diversity of substrate specificity, enzyme architecture, and subcellular localization across various taxonomic groups (Wanders and Waterham 2006).

In mammals, fatty acids are degraded in two subcellular locations: mitochondria and peroxisomes. The reaction steps in the two organelles are similar, but are catalyzed by different sets of enzymes (Fig. 1). Mammalian mitochondria host two beta oxidation pathways. The type I pathway degrades medium- and short-chain fatty acids through four basic reactions involving the enzymes acyl-CoA dehydrogenase (acyl-CoA-DH), enoyl-CoA hydratase (enoyl-CoA-HT), 3-OH-acyl-CoA dehydrogenase (3-OH-acyl-CoA-DH), and 3-keto-acyl-CoA thiolase (3-keto-acyl-CoA-TH); we will designate these reactions as step 1, 2, 3, and 4 of beta oxidation. The type II pathway, which breaks down long-chain substrates, includes only two enzymes because the final three steps of the reaction are catalyzed by a single enzyme, the so-called trifunctional enzyme (TFE; Uchida et al. 1992). In mammalian peroxisomes, it is acyl-CoA oxidase (acyl-CoA-OX) rather than a dehydrogenase that performs the first step of beta oxidation, and the

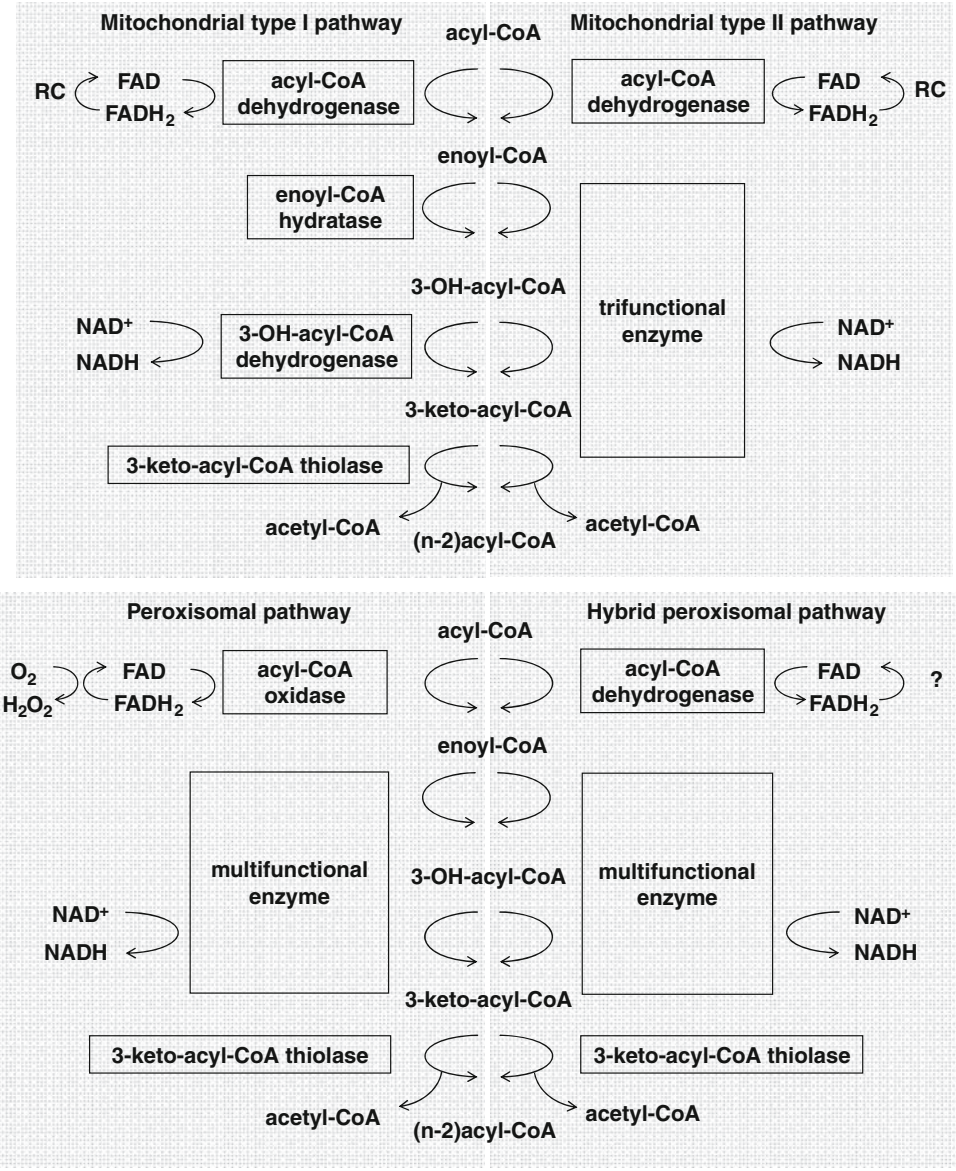


Fig. 1 Enzyme architecture of the various beta oxidation pathways. The multifunctional enzyme displays both enoyl-CoA hydratase and 3-OH-acyl-CoA dehydrogenase activities and shares sequence similarity to the two corresponding mitochondrial enzymes. Trifunctional enzyme combines enoyl-CoA hydratase, 3-OH-acyl-CoA dehydrogenase, and 3-keto-acyl-CoA thiolase activities. The major differences between the four beta oxidation forms are the participating enzymes and the hydrogen acceptors. *RC* Respiratory chain

multifunctional enzyme (MFE) combines the activity of enoyl-CoA-HT and 3-OH-acyl-CoA-DH to catalyze reaction steps 2 and 3.

In non-animal eukaryotes, fatty acid degradation is less well studied. For example, in plants, it is disputed whether beta oxidation operates only in peroxisomes or in mitochondria as well (Masterson and Wood 2000). In the fungal kingdom, the pathway is traditionally thought to be located exclusively in peroxisomes. Specifically, the yeasts *Saccharomyces cerevisiae* and *Candida tropicalis* reportedly lack a mitochondrion-located pathway (Hiltunen et al. 1992; Kurihara et al. 1992). However, this view is challenged by recent studies, which show that *Emericella* (previously designated *Aspergillus*) *nidulans* possesses both peroxisomal and mitochondrial beta oxidation (Maggio-Hall and Keller 2004).

Fungal fatty acid degradation has been investigated in only a small number of species, and the picture emerging is incoherent. It is unknown which type of beta oxidation is more representative of fungi, that of *Emericella* with two distinct pathways (one each located in mitochondria and peroxisomes) or that of yeast with a single peroxisomal pathway. This question can now be readily addressed because of the recent availability of numerous complete genome sequences from a taxonomically broad spectrum of species. For example, a recent in silico study searched for hallmark beta oxidation enzymes in 34 fungal genomes published at that time (Cornell et al. 2007). Yet, this study leaves several central issues unresolved. (a) Since only a single protein was used to represent a whole pathway, it is not known whether the enzyme set for a particular pathway is complete. (b) The enzyme localization was inferred from

sequence similarity alone, although the homologs from diverse taxa may well be targeted to different subcellular locations. Both issues are addressed here, together with a perspective on global pathway evolution. Specifically, we conducted a large-scale and detailed in silico analysis of beta oxidation enzymes from 57 fungal species whose complete nuclear genome sequences have been released to date. Genomes were screened for all key enzymes involved in beta oxidation, subcellular localization of proteins was predicted independently, and enzymes were mapped on a phylogenetic tree. This analysis provides new insights into the diversity of beta oxidation across fungi as detailed below.

In silico identification of beta oxidation pathways

Enzymes constituting beta oxidation pathways have been well characterized in several model organisms (Hiltunen and Qin 2000; Poirier et al. 2006). Although such studies provide valuable references for in silico identification of beta oxidation enzymes, we realized that the localization of this pathway cannot be easily inferred from sequence-similarity-based function annotation of the participating enzymes, especially for 3-keto-acyl-CoA-TH. In rat, mitochondrial and peroxisomal 3-keto-acyl-CoA-TH share ~37% sequence identity, so certain remote homologues match enzymes from both compartments with similar scores. Thus, unambiguous distinction of beta oxidation enzymes requires also *ab initio* prediction of their subcellular localization. Together with this latter approach, ~94% of potential 3-keto-acyl-CoA-TH sequences could be reliably assigned to either mitochondria or peroxisomes (Supplementary Table 1). The remaining 6% (32 proteins) have predicted affiliations to both organelles. (Enzymes for which all protein-family members were assigned to two locations are labeled gray in Fig. 2; individual proteins are cross-referenced in Supplementary Tables 2 and 3 and compiled in Supplementary Table 4) Dually assigned proteins may represent false positive in silico predictions, but it is quite likely that they indeed reside in both organelles as shown experimentally for two enzymes involved in lipid metabolism, dienoyl-CoA-isomerase and carnitine acetyl-transferase (Filppula et al. 1998; Kawachi et al. 1996).

To identify the presence, type, and location of beta oxidation across fungi, we collected the proteins deduced from completely sequenced fungal nuclear genomes (Table 1). We searched by BLAST among these sequences for possible homologues of well-characterized mitochondrial and peroxisomal beta oxidation enzymes. For each putative homologue found, we separately predicted its subcellular localization (methods are described in the Supplementary Methods). The combined evidence of both analyses was used to infer where beta oxidation occurs in a given organism (Table 1).

Mitochondrial beta oxidation in fungi

We searched in fungal sequences for type I and II of mitochondrial beta oxidation (see "Introduction"), but did not detect any homolog of TFE, indicating that type II beta oxidation is absent from fungi. In the following, "mitochondrial beta oxidation" refers only to type I.

Sequence similarity plus subcellular localization prediction identified at least one homologue for each of the four mitochondrial beta oxidation enzymes in over 50% of the investigated species, including members of all major fungal groups (Fig. 2; sequence IDs are listed in Supplementary Table 2). We propose that these species possess an intact mitochondrial pathway. The corresponding genes generally occur in sizable families, the largest being acyl-CoA-DH with up to 10 members with an average family size of five (Supplementary Table 2).

We encountered incomplete sets of mitochondrial beta oxidation enzymes in 12 species (Fig. 2, Supplementary Table 2). Recognizable homologs of 3-OH-acyl-CoA-DH, which catalyzes step 3 of the reaction, are lacking in two basidiomycetes (*Postia* and *Phanerochaete*) and one ascomycete (*Ajellomyces*). Interestingly, these species possess tentative homologs of 3-OH-butyryl-CoA dehydrogenase. Therefore, the three species most probably possess a functional mitochondrial beta oxidation, but one which is limited to butyric acid as substrate. Homologues of enzymes catalyzing both step 3 and 4 are missing in *Candida*, *Lodderomyces*, *Eremothecium*, and *Yarrowia*. Three out of four enzymes seem absent in *Pichia stipitis*, *Kluyveromyces*, and *Debaryomyces*. Finally none of the four mitochondrial beta oxidation enzymes could be detected in the *Encephalitozoon*, *Pichia guilliermondii*, *Clavispora*, *Saccharomyces*, *Lachancea*, and the two *Schizosaccharomyces* species (Fig. 2).

The findings described above extend as well as question conclusions drawn in an earlier investigation of fungal genomes (Cornell et al. 2007). This previous study used acyl-CoA-DH as a hallmark of 'non-peroxisomal' beta oxidation. However, acyl-CoA-DH participates also in the degradation of valine, leucine, and isoleucine. Therefore, the occurrence of this enzyme is not sufficient to infer functional beta oxidation outside peroxisomes. In fact, we found a total of ten fungi possessing acyl-CoA-DH, but lacking up to three other enzymes of the mitochondrial pathway (Fig. 2), strongly suggesting the absence of mitochondrial beta oxidation in these species. Confirmation of our in silico results comes from experimental studies in *C. tropicalis* and *S. cerevisiae*, showing absence of beta

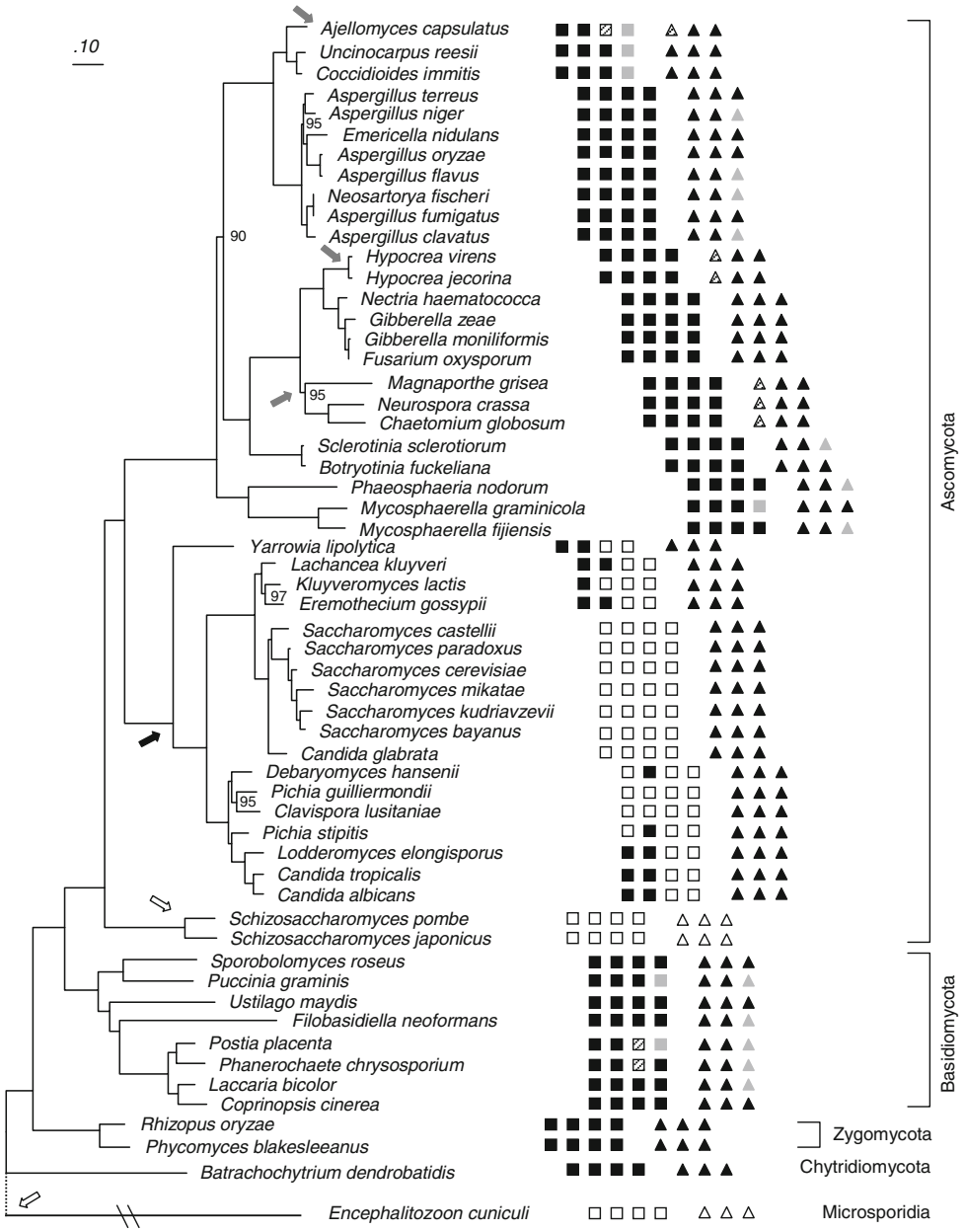


Fig. 2 Beta oxidation enzymes mapped on the fungal phylogeny tree. The tree was built with concatenated homologs of eight proteins: the two largest subunits of RNA polymerase II (RPB1 and RPB2) and III (RPC1 and RPC2), mitochondrial and cytoplasmic heat shock proteins 70 (HSP70_mit and HSP70_cyt), elongation factor II (EF2), and 60 kDa chaperonin (CPN60). Multiple alignment was performed with MUSCLE (Edgar 2004) and tree construction with the Bayesian program PhyloBayes (Lartillot and Philippe 2004), using the CAT+Γ model with eight discrete gamma rate categories. *Homo sapiens* and *Monosiga brevicollis* were used as outgroup (not shown). *Numbers at branches* indicate posterior probabilities if <100%. Maximum likelihood trees (not shown) display the same topology except for slight differences in the three branches leading to *Clavispora*, *Eremothecium*, and *Magnaporthe*, which all have insignificant bootstrap support. The position of *Encephalitozoon* could not be resolved, and its branch has been manually added to the tree. Each *square* represents a mitochondrial beta oxidation enzyme: (from *left* to *right*) acyl-CoA dehydrogenase, enoyl-CoA hydratase, 3-OH-acyl-CoA dehydrogenase, 3-keto-acyl-CoA thiolase. Each *triangle* stands for a peroxisomal beta oxidation enzyme: (from *left* to *right*) acyl-CoA oxidase, multifunctional enzyme, 3-keto-acyl-CoA thiolase. *Open symbols* The corresponding enzyme was not found. *Closed symbols* The corresponding enzyme was found. *Gray symbols* All homologs of the corresponding enzymes are predicted to reside in both mitochondria and peroxisomes. ▨ 3-OH-butyryl-CoA dehydrogenase instead of 3-OH-acyl-CoA dehydrogenase was found. ◬ Acyl-CoA dehydrogenase substitutes for acyl-CoA oxidase. *Arrows* indicate suggested evolutionary events. ➡ Loss of mitochondrial beta oxidation; ⇨ loss of both mitochondrial and peroxisomal beta oxidation; ➡ hybrid peroxisomal beta oxidation

Table 1 Beta oxidation pathways in fungi

Species name	Data source	Mitochondrial pathway	Peroxisomal pathway	Hybrid peroxisomal pathway
Ajellomyces capsulatus	Broad Institute[a]	?		Y
Aspergillus clavatus	Broad Institute	Y	Y	
Aspergillus flavus	Broad Institute	Y	Y	
Aspergillus fumigatus	Broad Institute	Y	Y	
Aspergillus niger	DOE-JGI[b]	Y	Y	
Aspergillus oryzae	Broad Institute	Y	Y	
Aspergillus terreus	Broad Institute	Y	Y	
Batrachochytrium dendrobatidis	Broad Institute	Y	Y	
Botryotinia fuckeliana	Broad Institute	Y	Y	
Candida albicans	Broad Institute		Y	
Candida glabrata	Génolevures[c]		Y	
Candida tropicalis	Broad Institute		Y	
Chaetomium globosum	Broad Institute	Y		Y
Clavispora lusitaniae	Broad Institute		Y	
Coccidioides immitis	Broad Institute	Y	Y	
Coprinopsis cinerea	Broad Institute	Y	Y	
Debaryomyce hansenii	Génolevures		Y	
Emericella nidulans	Broad Institute	Y	Y	
Encephalitozoon cuniculi	NCBI[d]			
Eremothecium gossypii	NCBI		Y	
Filobasidiella neoformans	Broad Institute	Y	Y	
Fusarium oxysporum	Broad Institute	Y	Y	
Gibberella moniliformis	Broad Institute	Y	Y	
Gibberella zeae	Broad Institute	Y	Y	
Hypocrea jecorina	DOE-JGI	Y		Y
Hypocrea virens	DOE-JGI	Y		Y
Kluyveromyces lactis	Génolevures		Y	
Laccaria bicolor	DOE-JGI	Y	Y	
Lachancea kluyveri	SGD[e]		Y	
Lodderomyces elongisporus	Broad Institute		Y	
Magnaporthe grisea	Broad Institute	Y		Y
Mycosphaerella fijiensis	DOE-JGI	Y	Y	
Mycosphaerella graminicola	DOE-JGI	Y	Y	
Nectria haematococca	DOE-JGI	Y	Y	
Neosartorya fischeri	Broad Institute	Y	Y	
Neurospora crassa	Broad Institute	Y		Y
Phaeosphaeria nodorum	Broad Institute	Y	Y	
Phanerochaete chrysosporium	Broad Institute	?	Y	
Phycomyces blakesleeanus	DOE-JGI	Y	Y	
Pichia guilliermondii	Broad Institute		Y	
Pichia stipitis	Broad Institute		Y	
Postia placenta	Broad Institute	?	Y	
Puccinia graminis	Broad Institute	Y	Y	
Rhizopus oryzae	Broad Institute	Y	Y	
Saccharomyces bayanus	SGD		Y	
Saccharomyces castellii	SGD		Y	
Saccharomyces cerevisiae	Broad Institute		Y	
Saccharomyces kudriavzevii	SGD		Y	
Saccharomyces mikatae	SGD		Y	
Saccharomyces paradoxus	SGD		Y	
Schizosaccharomyces japonicus	Broad Institute			
Schizosaccharomyces pombe	Broad Institute			
Sclerotinia sclerotiorum	Broad Institute	Y	Y	
Sporobolomyces roseus	DOE-JGI	Y	Y	

Table 1 (continued)

Species name	Data source	Mitochondrial pathway	Peroxisomal pathway	Hybrid peroxisomal pathway
Uncinocarpus reesii	Broad Institute	Y	Y	
Ustilago maydis	Broad Institute	Y	Y	
Yarrowia lipolytica	Génolevures		Y	

Y The pathway exists, *?* the type of pathway could not be determined
[a] Broad Institute of MIT and Harvard (http://www.broad.mit.edu/annotation/fgi/)
[b] DOE Joint Genome Institute (http://www.jgi.doe.gov/)
[c] Génolevures Consortium (http://cbi.labri.fr/Genolevures/)
[d] http://www.ncbi.nlm.nih.gov/Genomes/
[e] Saccharomyces Genome Database (http://www.yeastgenome.org/)

oxidation activity in isolated mitochondria (Hiltunen et al. 1992; Kurihara et al. 1992).

Some of our in silico results challenge the experimental data. As stated above, our analyses identified the homologues of all four mitochondrial beta oxidation enzymes in *Neurospora*. In contrast, enzyme tests with isolated mitochondria failed to detect activities of 3-OH-acyl-CoA-DH and 3-keto-acyl-CoA-TH, which led to the assumption that mitochondrial beta oxidation is absent from this fungus (Thieringer and Kunau 1991). We explain the conflicting bioinformatic and biochemical evidence as follows. The cells used in the experiment were cultivated in medium containing oleate (C_{18}, long-chain fatty acid) as sole carbon source, while the two mitochondrial enzymes are specialized to degrade short-chain fatty acids (Hiltunen and Qin 2000). Since beta oxidation activity in *Neurospora* requires substrate induction (Kionka and Kunau 1985), the absence of enzyme activity is most likely due to lack of expression rather than lack of genes.

Peroxisomal beta oxidation in fungi

We searched fungal sequences for potential homologues of acyl-CoA-OX, MFE, and 3-keto-acyl-CoA-TH and then predicted whether or not these proteins are localized in peroxisomes. Homologues of all three peroxisomal enzymes are present in ~80% of investigated fungal species (across all groups, Fig. 2, Supplementary Table 3), which indicates that these taxa host a peroxisomal beta oxidation pathway. Again, most genes encoding components of beta oxidation exist in families, the largest being acyl-CoA-OX with up to 11 members with an average size of two.

Incomplete enzyme sets occur in six ascomycetes: *Ajellomyces*, *Chaetomium*, the two *Hypocrea* species, *Magnaporthe*, and *Neurospora*. In each instance, it is the enzyme acyl-CoA-OX (catalyzing step 1 in peroxisomal beta oxidation) that appears absent (Fig. 2). Previous experimental studies in *Neurospora* and *Magnaporthe* (Kionka and Kunau 1985; Thieringer and Kunau 1991; Wang et al. 2007) report beta oxidation activity in isolated peroxisomes, with acyl-CoA-DH substituting for acyl-CoA-OX, the former enzyme being typical for the mitochondrial pathway. Combining the experimental evidence and our analysis, we conclude that *Neurospora* and *Magnaporthe* possess beta oxidation in both organelles: a canonical mitochondrial pathway and a hybrid peroxisomal pathway. This pattern is apparently shared with two other *Sordariomycetes*, *Chaetomium*, and *Hypocrea* (Fig. 2), in which we observe the same predicted enzyme profile. A hybrid peroxisomal pathway seems also present in *Ajellomyces*. However, its mitochondrial pathway remains obscure because 3-OH-acyl-CoA-DH could not be unambiguously identified in this species.

Finally, all peroxisomal (and mitochondrial) beta oxidation enzymes appear absent in *Encephalitozoon* and *Schizosaccharomyces*, as proposed by others before (Cornell et al. 2007).

Evolution of beta oxidation in fungi

Our study demonstrates convincingly that contrary to common belief, most fungi host both peroxisomal and mitochondrial beta oxidation (Table 1). However, the two pathways show different trends in evolution. To retrace the evolutionary history of beta oxidation in fungi, we mapped the presence/absence of individual enzymes on a phylogenetic species tree (Fig. 2). This shows that certain losses of enzymes and pathways are ancient events, whereas others occurred relatively recently in evolutionary time.

Consistent with previous views (Cornell et al. 2007), the mitochondrial pathway has been lost early on in the common ancestor of all *Saccharomycetes*. Further, our data suggest that *Sordariomycetes* have "invented" a hybrid peroxisomal beta oxidation. In *Encephalitozoon* and *Schizosaccharomyces*, we and others (Cornell et al. 2007) did not detect any enzyme involved in beta oxidation. Absence of fatty acid degradation in *Encephalitozoon*, an obligate intracellular parasite, is not surprising, given the loss of canonical mitochondria and

peroxisomes all together. In turn, the free-living *Schizosaccharomyces* do retain fully functional organelles, but populate sugar-rich environmental niches. This allows the fission yeast to specialize on fermentable substrates and lose the capacity to degrade other carbon sources, including glycerol and fatty acids.

Conclusion

Our large-scale in silico study reveals a most diverse make-up of beta oxidation in fungi. The majority, and in particular early diverging taxa such as *Rhizopus* and *Phycomyces*, possess both mitochondrial and peroxisome-located pathways. This feature is shared by the sister clade of fungi, the animals, and thus most likely represents the primitive state in Opisthokonts. Another noteworthy finding is that certain fungi have lost a functional mitochondrial pathway, while the peroxisomal pathway remains intact (e.g., *Saccharomyces*), yet the inverse case was not detected. When the peroxisomal enzyme set is incomplete, it is in all instances acyl-CoA-OX that is missing and which can be complemented by acyl-CoA-DH to form a functional hybrid pathway in peroxisomes.

In summary, beta oxidation is an illustrative example of the catabolic versatility of the fungal kingdom, with a spectrum covering from hyper-specialists to most broad generalists, the latter capable of degrading virtually all forms of organic matter encountered on Earth.

Acknowledgements This work was supported by the Genome-Canada and Canadian Institutes of Health Research (CIHR, Institute of Genetics, MOP-79309). We would like to thank B. Franz Lang, Henner Brinkmann, Pierre Rioux, and Nicolas Lartillot (Université de Montréal) for their help with phylogenetic analyses. We also thank Emmet O'Brien (Université de Montréal) for improving the manuscript. YQS is a Canadian Institute for Health Research (CIHR) Strategic Training Fellow in Bioinformatics.

References

Cornell MJ, Alam I, Soanes DM, Wong HM, Hedeler C, Paton NW, Rattray M, Hubbard SJ, Talbot NJ, Oliver SG (2007) Comparative genome analysis across a kingdom of eukaryotic organisms: specialization and diversification in the fungi. Genome Res 17:1809–1822

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

Filppula SA, Yagi AI, Kilpelainen SH, Novikov D, FitzPatrick DR, Vihinen M, Valle D, Hiltunen JK (1998) Delta3,5-delta2,4-dienoyl-CoA isomerase from rat liver. Molecular characterization. J Biol Chem 273:349–355

Hiltunen JK, Qin Y (2000) Beta-oxidation—strategies for the metabolism of a wide variety of acyl-CoA esters. Biochim Biophys Acta 1484:117–128

Hiltunen JK, Wenzel B, Beyer A, Erdmann R, Fossa A, Kunau WH (1992) Peroxisomal multifunctional beta-oxidation protein of *Saccharomyces cerevisiae*. Molecular analysis of the fox2 gene and gene product. J Biol Chem 267:6646–6653

Kawachi H, Atomi H, Ueda M, Tanaka A (1996) Peroxisomal and mitochondrial carnitine acetyltransferases of the *n*-alkane-assimilating yeast *Candida tropicalis*. Analysis of gene structure and translation products. Eur J Biochem 238:845–852

Kionka C, Kunau WH (1985) Inducible beta-oxidation pathway in *Neurospora crassa*. J Bacteriol 161:153–157

Kurihara T, Ueda M, Okada H, Kamasawa N, Naito N, Osumi M, Tanaka A (1992) Beta-oxidation of butyrate, the short-chain-length fatty acid, occurs in peroxisomes in the yeast *Candida tropicalis*. J Biochem 111:783–787

Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095–1109

Maggio-Hall LA, Keller NP (2004) Mitochondrial beta-oxidation in *Aspergillus nidulans*. Mol Microbiol 54:1173–1185

Masterson C, Wood C (2000) Mitochondrial beta-oxidation of fatty acids in higher plants. Physiol Plant 109:217–224

Poirier Y, Antonenkov VD, Glumoff T, Hiltunen JK (2006) Peroxisomal beta-oxidation—a metabolic pathway with multiple functions. Biochim Biophys Acta 1763:1413–1426

Thieringer R, Kunau WH (1991) The beta-oxidation system in catalase-free microbodies of the filamentous fungus *Neurospora crassa*. Purification of a multifunctional protein possessing 2-enoyl-CoA hydratase, L-3-hydroxyacyl-CoA dehydrogenase, and 3-hydroxyacyl-CoA epimerase activities. J Biol Chem 266:13110–13117

Uchida Y, Izai K, Orii T, Hashimoto T (1992) Novel fatty acid beta-oxidation enzymes in rat liver mitochondria. II. Purification and properties of enoyl-coenzyme A (CoA) hydratase/3-hydroxyacyl-CoA dehydrogenase/3-ketoacyl-CoA thiolase trifunctional protein. J Biol Chem 267:1034–1041

Wanders RJ, Waterham HR (2006) Biochemistry of mammalian peroxisomes revisited. Annu Rev Biochem 75:295–332

Wang ZY, Soanes DM, Kershaw MJ, Talbot NJ (2007) Functional analysis of lipid metabolism in *Magnaporthe grisea* reveals a requirement for peroxisomal fatty acid beta-oxidation during appressorium-mediated plant infection. Mol Plant Microbe Interact 20:475–491

Supplementary files

Materials and Methods

Data collection

Protein sequences deduced from 57 completed fungi genomes was retrieved from Broad Institute, DOE Joint Genome Institute, *Saccharomyces* Genome Database, Génolevures Consortium, and NCBI (Table 1). The Uniref90 database was retrieved from UniProt.

Selection of beta oxidation enzymes via similarity search

We conducted a BLASTP search between the fungal proteins and sequences from UniRef90 database, using the threshold $e=10^{-10}$ for significant matches. If the top hit against UniRef90 database was a beta oxidation enzyme, the query protein was annotated as such. To infer remote homologs, the enzymes identified were used for a second 'transitive' screen to annotate sequences without any matches in the first run, a procedures employed with success by the annotation tool AutoFACT (Koski, Gray et al. 2005). For species in which beta oxidation enzymes were not found, we also searched by TBLASTN in their genome sequences to avoid false negatives caused by gene prediction error. However, it cannot be completely ruled out that some enzymes appear to be missing only because of incomplete genome data. For species in which beta oxidation enzymes were not found in the genome sequence, we scrutinized other available data sources such as ESTs.

Mitochondrial protein prediction

The subcellular localization prediction of the proteins was based on nine predictors: TargetP (Emanuelsson, Nielsen et al. 2000), Subloc (Hua and Sun 2001), SherLoc (Shatkay, Hoglund et al. 2007), pTARGET (Guda and Subramaniam 2005), Predotar (Small, Peeters et al. 2004), Protein Prowler (PProwler) (Boden and Hawkins 2005), PASUB (Lu, Szafron et al. 2004), MitoProt (Claros and Vincens 1996), and CELLO (Yu, Chen et al. 2006). All the results were obtained from online servers of the predictors, except for MitoProt, which was installed and run locally.

For most proteins, the predictors gave contradictory results. We integrated these predictions by employing the tool YimLOC (Shen and Burger 2007), which has been shown to be more accurate than any of the individual predictors. We used the YimLOC result as the final localization prediction.

Peroxisomal protein prediction

Six of the nine predictors above include the class "peroxisomes" (SherLoc, pTARGET, PProwler, PASUB, PST1 (Neuberger, Maurer-Stroh et al. 2003), and CELLO). Since the predictors have low to medium sensitivity for peroxisomal proteins (22%-77%), we employed the following scheme to maximize the sensitivity by combing function annotation with localization prediction: for a query protein, if 1) by similarity search it is

annotated as a peroxisomal beta oxidation enzyme; 2) any of these predictors classifies it into peroxisomes; 3) it is not targeted to mitochondria according to YimLOC, this protein will be predicted as peroxisome-destined. Dual-localization will be assigned if this protein is also recognized as a mitochondrial one by YimLOC.

Phylogenetic inference

We chose eight proteins that have been successfully used in previous fungal phylogenetic analyses. These are the two largest subunits of RNA polymerase II (RPB1 and RPB2), and III (RPC1 and RPC2), mitochondrial and cytoplasmic heat shock proteins 70 (HSP70_mit and HSP70_cyt), elongation factor II (EF2), and 60kDa chaperonin (CPN60). We used *Homo sapiens* and *Monosiga brevicollis* as outgroup (not shown). Homologs of these proteins were initially searched in each species with a BLASTP cutoff of $e=10^{-100}$). Only unambiguous orthologs were retained, identified by phylogenetic analyses of individual proteins (Supplementary Table 5). Protein sequence alignments were generated using MUSCLE (Edgar 2004). Aligned sequences were concatenated, and Gblocks (Castresana 2000) was employed to remove ambiguously aligned regions and highly divergent parts of the alignment. Three independent phylogenetic runs were performed with the Bayesian program PhyloBayes (Lartillot and Philippe 2004), using the CAT+Γ model with eight discrete gamma rate categories. After the convergence of likelihood values across generations, the final consensus tree was built based on the combined runs. Bayesian posterior probabilities were obtained from the majority rule consensus of the tree sampled

after 1,000 generations. We also constructed maximum likelihood trees using RAxML (Stamatakis 2006) using the WAG+ Γ model with four discrete gamma rate categories, and we evaluated the statistical support by 100 bootstrap replicates.

Reference:

Boden, M., and Hawkins, J. (2005) Prediction of subcellular localization using sequence-biased recurrent networks. Bioinformatics **21**: 2279-2286.

Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol **17**: 540-552.

Claros, M.G., and Vincens, P. (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur J Biochem **241**: 779-786.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32**: 1792-1797.

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol **300**: 1005-1016.

Guda, C., and Subramaniam, S. (2005) pTARGET [corrected] a new method for predicting protein subcellular localization in eukaryotes. Bioinformatics **21**: 3963-3969.

Hua, S., and Sun, Z. (2001) Support vector machine approach for protein subcellular localization prediction. Bioinformatics **17**: 721-728.

Koski, L.B., Gray, M.W., Lang, B.F., and Burger, G. (2005) AutoFACT: an automatic functional annotation and classification tool. BMC Bioinformatics **6**: 151.

Lartillot, N., and Philippe, H. (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol **21**: 1095-1109.

Lu, Z., Szafron, D., Greiner, R., Lu, P., Wishart, D.S., Poulin, B., Anvik, J., Macdonell, C., and Eisner, R. (2004) Predicting subcellular localization of proteins using machine-learned classifiers. Bioinformatics **20**: 547-556.

Neuberger, G., Maurer-Stroh, S., Eisenhaber, B., Hartig, A., and Eisenhaber, F. (2003) Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. J Mol Biol **328**: 581-592.

Shatkay, H., Hoglund, A., Brady, S., Blum, T., Donnes, P., and Kohlbacher, O. (2007) SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. Bioinformatics.

Shen, Y.Q., and Burger, G. (2007) 'Unite and conquer': enhanced prediction of protein subcellular localization by integrating multiple specialized tools. BMC Bioinformatics **8**: 420.

Small, I., Peeters, N., Legeai, F., and Lurin, C. (2004) Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics **4**: 1581-1590.

Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**: 2688-2690.

Yu, C.S., Chen, Y.C., Lu, C.H., and Hwang, J.K. (2006) Prediction of protein subcellular localization. Proteins.

Supplementary Table 1 Genome-deduced proteins with significant sequence similarity to both mitochondrial and peroxisomal keto-acyl-CoA thiolase

Sequence ID	Species name	BLASTP score to mitochondrial enzyme[1]	BLASTP score to peroxisomal enzyme[2]	Localization prediction[3]
HCAG_07123.1	*Ajellomyces capsulatus*	91.3	125	per
ACLA_070310	*Aspergillus clavatus*	196	350	mit
AFL2G_01917	*Aspergillus flavus*	211	387	per
AFL2G_06769	*Aspergillus flavus*	205	358	mit
Afu2g11350	*Aspergillus fumigatus*	202	367	mit
Afu7g04080	*Aspergillus fumigatus*	203	385	mit
fge1_pm_C_2000057	*Aspergillus niger*	189	338	mit
AO090003001121	*Aspergillus oryzae*	212	387	per
AO090026000515	*Aspergillus oryzae*	205	358	mit
ATEG_00456	*Aspergillus terreus*	196	357	mit
ATEG_01444	*Aspergillus terreus*	199	365	mit
ATEG_03795	*Aspergillus terreus*	197	331	per
ATEG_03980	*Aspergillus terreus*	201	367	mit
BDEG_06810	*Batrachochytrium dendrobatidis*	202	399	mit
BC1G_00632	*Botrytis cinerea*	210	355	per
BC1G_03480	*Botrytis cinerea*	185	332	mit
BC1G_13290	*Botrytis cinerea*	192	323	per
CAWG_01376	*Candida albicans*	192	358	per
CAWG_02477	*Candida albicans*	210	338	per
CTRG_01068.3	*Candida tropicalis*	202	357	per
CTRG_02168.3	*Candida tropicalis*	211	326	per
CHG05249.1	*Chaetomium globosum*	187	334	mit
CHG08784.1	*Chaetomium globosum*	199	342	per
CLUG_04882	*Clavispora lusitaniae*	196	355	per
CIMG_01533	*Coccidioides immitis*	155	310	per
CC1G_06261.1	*Coprinus cinereus*	152	281	mit
AN1050	*Emericella nidulans*	192	359	per
AN5646	*Emericella nidulans*	201	387	per
AN5878	*Emericella nidulans*	192	360	mit
CNAG_00524.1	*Filobasidiella neoformans*	403	215	mit
FOXG_02954	*Fusarium oxysporum*	205	347	per
FOXG_04827	*Fusarium oxysporum*	181	270	mit
FOXG_11385	*Fusarium oxysporum*	200	345	mit
FOXG_13516	*Fusarium oxysporum*	52	80.9	per
FVEG_01806	*Gibberella moniliformis*	205	350	per
FVEG_03315	*Gibberella moniliformis*	218	305	mit
FVEG_10433	*Gibberella moniliformis*	199	345	mit

FVEG_13890	*Gibberella moniliformis*	160	286	mit
FGSG_04243	*Gibberella zeae*	197	348	mit
FGSG_13398	*Gibberella zeae*	187	329	per
jgi_Trire2_123720	*Hypocrea jecorina*	203	343	mit
jgi_Trire2_75368	*Hypocrea jecorina*	189	345	per
jgi_Trive1_73014	*Hypocrea virens*	199	337	mit
jgi_Trive1_80821	*Hypocrea virens*	194	349	per
jgi_Lacbi1_301021	*Laccaria bicolor*	160	310	mit
LELG_02195	*Lodderomyces elongisporus*	184	326	per
LELG_05753	*Lodderomyces elongisporus*	186	325	per
MGG_09512	*Magnaporthe grisea*	189	332	mit
MGG_10700	*Magnaporthe grisea*	196	349	per
MGG_13647	*Magnaporthe grisea*	65.1	123	mit
jgi_Mycfi1_79474	*Mycosphaerella fijiensis*	173	308	mit
jgi_Mycgr1_83653	*Mycosphaerella graminicola*	202	354	per
jgi_Necha2	*Nectria haematococca*	206	342	mit
jgi_Necha2	*Nectria haematococca*	209	338	mit
jgi_Necha2	*Nectria haematococca*	196	337	per
NFIA_086620	*Neosartorya fischeri*	201	364	mit
NCU04796	*Neurospora crassa*	204	369	per
NCU05558	*Neurospora crassa*	193	336	per
NCU09646	*Neurospora crassa*	198	337	mit
NCU10691	*Neurospora crassa*	204	369	per
SNU13439.1	*Phaeosphaeria nodorum*	216	354	mit
jgi_Phchr1_125276	*Phanerochaete chrysosporium*	195	298	mit
jgi_Phybl_33195	*Phycomyces blakesleeanus*	436	240	mit
jgi_Phybl_68683	*Phycomyces blakesleeanus*	230	183	mit
jgi_Phybl_69337	*Phycomyces blakesleeanus*	412	238	mit
jgi_Phybl_26171	*Phycomyces blakesleeanus*	192	370	per
jgi_Phybl_75034	*Phycomyces blakesleeanus*	201	412	per
PGUG_01256.1	*Pichia guilliermondii*	209	367	per
SCRG_05333	*Saccharomyces cerevisiae*	196	382	per
SS1G_04381.1	*Sclerotinia sclerotiorum*	184	311	mit
SS1G_08207.1	*Sclerotinia sclerotiorum*	191	336	mit
jgi_Sporo1_9693	*Sporobolomyces roseus*	194	327	mit
jgi_Sporo1_12251	*Sporobolomyces roseus*	198	380	per
UREG_05293.1	*Uncinocarpus reesii*	202	377	per
UM01090.1	*Ustilago maydis*	209	320	mit
UM02715.1	*Ustilago maydis*	188	364	per

[1] BLASTP search against mitochondrial 3-keto-acyl-CoA thiolase from *Rattus norvegicus* (P13437)
[2] BLASTP search against two enzymes, peroxisomal 3-keto-acyl-CoA thiolase A (P21775) and B (P07871) from *Rattus norvegicus*. The higher score of the two alignments is reported
[3] mit: mitochondria; per: peroxisomes

Supplementary Table 2-5 mainly compile sequence IDs, and each table is more than two pages long. Therefore instead of the tables, their links are provided below:

Supplementary Table 2 Complete set of mitochondrial beta oxidation enzymes assigned in this study (sequences in parentheses are also predicted as peroxisomal proteins)

http://www.springerlink.com/content/46422l55622434v0/MediaObjects/10142_2008_95_MOESM3_ESM.doc

Supplementary Table 3 Complete sets of peroxisomal beta oxidation enzymes assigned in this study (sequences in parentheses are also predicted as mitochondrial proteins)

http://www.springerlink.com/content/46422l55622434v0/MediaObjects/10142_2008_95_MOESM4_ESM.doc

Supplementary Table 4 Genome-deduced proteins that show significant similarity to both mitochondrial and peroxisomal keto-acyl-CoA thiolase, and predicted as targeted to both mitochondria and peroxisomes

http://www.springerlink.com/content/46422l55622434v0/MediaObjects/10142_2008_95_MOESM5_ESM.doc

Supplementary Table 5 Protein sequences used for constructing the phylogenetic tree

http://www.springerlink.com/content/46422l55622434v0/MediaObjects/10142_2008_95_MOESM6_ESM.doc

Chapter 3 Diversity and dispersal of acyl-CoA dehydrogenases

Acyl-CoA dehydrogenases form a large protein family with at least 13 subfamilies. Each subfamily has a distinct substrate preference, and accordingly participates in the degradation of specific fatty acids or amino acids. Except for a few species, ACAD subfamilies have not been well characterized. We screened for ACAD enzymes in 250 species and integrated the analysis of subcellular localization, taxomonic mapping, and subfamily phylogenies in order to survey the substrate specificity, distribution, and evolution of ACAD homologs.

Nucleic Acids Research Advance Access published July 22, 2009

Nucleic Acids Research, 2009, 1–13
doi:10.1093/nar/gkp566

Diversity and dispersal of a ubiquitous protein family: acyl-CoA dehydrogenases

Yao-Qing Shen*, B. Franz Lang and Gertraud Burger

Robert Cedergren Center for Bioinformatics and Genomics, Biochemistry Department, Université de Montréal, 2900 Edouard-Montpetit, Montreal, QC, H3T 1J4, Canada

Received May 18, 2009; Revised June 17, 2009; Accepted June 18, 2009

ABSTRACT

Acyl-CoA dehydrogenases (ACADs), which are key enzymes in fatty acid and amino acid catabolism, form a large, pan-taxonomic protein family with at least 13 distinct subfamilies. Yet most reported ACAD members have no subfamily assigned, and little is known about the taxonomic distribution and evolution of the subfamilies. In completely sequenced genomes from approximately 210 species (eukaryotes, bacteria and archaea), we detect ACAD subfamilies by rigorous ortholog identification combining sequence similarity search with phylogeny. We then construct taxonomic subfamily-distribution profiles and build phylogenetic trees with orthologous proteins. Subfamily profiles provide unparalleled insight into the organisms' energy sources based on genome sequence alone and further predict enzyme substrate specificity, thus generating explicit working hypotheses for targeted biochemical experimentation. Eukaryotic ACAD subfamilies are traditionally considered as mitochondrial proteins, but we found evidence that in fungi one subfamily is located in peroxisomes and participates in a distinct β-oxidation pathway. Finally, we discern horizontal transfer, duplication, loss and secondary acquisition of ACAD genes during evolution of this family. Through these unorthodox expansion strategies, the ACAD family is proficient in utilizing a large range of fatty acids and amino acids—strategies that could have shaped the evolutionary history of many other ancient protein families.

INTRODUCTION

From the last two decades of intensive research especially in mammals, acyl-CoA dehydrogenases (ACADs) are now known as a large and biologically important enzyme family. Genetic defects of the corresponding genes cause severe health problems in human, including hypoglycemia, neuromuscular pathology and even death (1). While ACAD proteins occur in all three domains of life, animals possess the largest number of distinct subfamilies. In human, for example, 11 different ACAD enzymes have been recognized (2–12). These proteins, which in eukaryotes are localized in mitochondria, share up to ~50% amino acid identity among each other (Table 1) and catalyze similar biochemical reactions: the oxidation of diverse acyl-CoA compounds, produced during the degradation of fat and protein, to enoyl-CoA (Figure 1).

ACAD subfamilies are distinguished by the metabolic pathways in which they participate, and by their substrate specificity (Figure 1, Table 2). Five subfamilies participate in β-oxidation of fatty acids, with optimal activity for acyl-CoA substrates of particular chain length, short (ACAD**S**), medium (ACAD**M**), long (ACAD**L**), or very long (ACAD**V** and ACAD**V2**) (11–15). Four other subfamilies are implicated in amino acid degradation. After removal of the amino groups from isoleucine, leucine, lysine/trytophan and valine, the remaining branched acyl-CoA is dehydrogenated by short/branched chain acyl-CoA dehydrogenase (ACDSB), isovaleryl-CoA dehydrogenase (IVD), glutaryl-CoA dehydrogenase (GCDH) and isobutyryl-CoA dehydrogenase (IBD), respectively (3,5–7). The most recently identified subfamilies, ACD10 and ACD11, are of yet unknown function (8,9). Two additional subfamilies have been reported in bacteria: fadE degrades a broad range of substrates from short to long chain acyl-CoAs (16,17), while fadE12 prefers medium-chain length molecules (18). The reaction mechanism and 3D structure of ACAD enzymes have been reviewed by others (19,20).

In eukaryotes, β-oxidation involving ACAD enzymes takes place in mitochondria. Eukaryotes also possess peroxisomal β-oxidation catalyzed by acyl-CoA oxidase (ACOX) instead of ACAD proteins. The two families resemble each other in several aspects. ACOX proteins share remote yet significant sequence similarity with ACAD proteins, and also catalyze the conversion of acyl-CoA to enoyl-CoA. But unlike the ACAD family, ACOX proteins occur predominantly in eukaryotes, are

*To whom correspondence should be addressed. Tel: +1 514 343 6111 2848; Fax: +1 514 343 2210

Table 1. Pairwise sequence similarities between human ACAD subfamily members[a]

	ACD11	ACADS	ACADM	ACADL	ACADV	ACADV2	ACDSB	GCDH	IVD	IBD	fadE	fadE12
ACD10	46	30	28	25	26	26	27	25	24	24	29	25
ACD11		29	26	26	23	26	26	23	22	27	31	25
ACADS			36	31	35	36	38	30	36	35	26	27
ACADM				32	34	34	37	27	34	33	24	25
ACADL					28	30	33	27	34	32	20	27
ACADV						45	34	30	32	30	24	24
ACADV2							38	29	34	33	25	22
ACDSB								28	33	35	25	24
GCDH									28	27	34	24
IVD										32	24	23
IBD											21	22
fadE												24

[a]All subfamily members are from human, except for fadE and fadE12, which are prokaryotic subfamilies. Percentage of identical residues in aligned region by BLAST. Sequences are obtained from SwissProt. Sequence IDs are listed in Table 2.

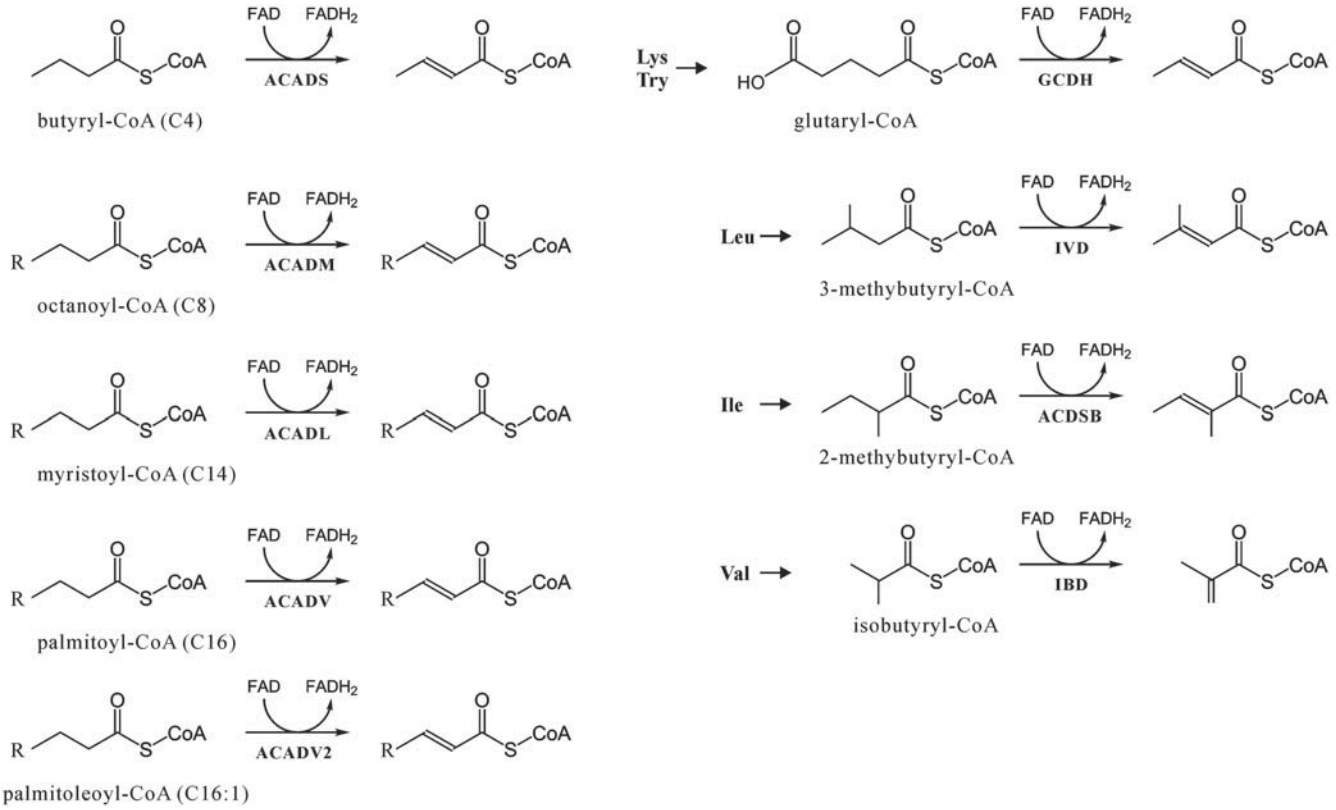


Figure 1. Optimal substrates of ACAD subfamilies. C4, etc., length of the acyl-CoA chain. C16:1, unsaturated fatty acid with one double bond. Subfamilies in the left part of the figure are involved in fatty acid degradation. Those in the right part are involved in amino acid degradation. 'R' represents straight alkyl chain.

located exclusively in peroxisomes and function by a distinct enzymatic mechanism: ACOX proteins are re-oxidized by molecular oxygen, generating H_2O_2 (20); ACAD enzymes, in contrast, having only low reactivity with molecular oxygen, are re-oxidized by electron-transferring flavoproteins, which in turn pass the electrons to the respiratory chain, generating H_2O. Insight into the origin of the ACOX family will critically depend on a better understanding of the ACAD family, which is the focus of the study reported here.

Our current knowledge about ACAD proteins is limited to a few model organisms. There has been no comprehensive survey of ACAD enzymes, except for genome-wide *in silico* screens in fungi without subfamily identification (21,22). Further, it is unclear whether the 11 subfamilies recognized in human are conserved throughout animals or even beyond. One reason for these shortcomings is that in public data repositories, sequences are generally annotated indistinctively as 'acyl-CoA dehydrogenase'. This is because in BLAST searches, remote ACAD

Table 2. Seed sequences used for BLAST searches

Protein name	Molecular function	Seed from	Sequence ID[a]	Evidence
ACADV	Oxidation of very long chain fatty acid	*Homo sapiens*	P49748	Experiment
ACADV2		*Homo sapiens*	Q9H845	Experiment
ACADL	Oxidation of long chain fatty acid	*Homo sapiens*	P28330	
		Monosiga brevicollis[a]	gi|167537125	BLAST and phylogeny
ACADM	Oxidation of medium chain fatty acid	*Homo sapiens*	P11310	
		Batrachochytrium dendrobatidis[a]	BDEG_05327	BLAST and phylogeny
		Monosiga brevicollis[a]	gi|167534479	
ACADS	Oxidation of short chain fatty acid	*Homo sapiens*	P16219	
		Monosiga brevicollis[a]	gi|167515960	BLAST and phylogeny
ACDSB	Oxidation of isoleucine	*Homo sapiens*	P45954	Experiment
		Batrachochytrium dendrobatidis[a]	BDEG_04739	BLAST and phylogeny
IVD	Oxidation of leucine	*Homo sapiens*	P26440	Experiment
		Batrachochytrium dendrobatidis[a]	BDEG_02262	BLAST and phylogeny
		Monosiga brevicollis[a]	gi|167524148	
GCDH	Oxidation of lysine and tryptophan	*Homo sapiens*	Q92947	Experiment
		Batrachochytrium dendrobatidis[a]	BDEG_05264	BLAST and phylogeny
		Monosiga brevicollis[a]	gi|167524186	
IBD	Oxidation of valine	*Homo sapiens*	Q9UKU7	Experiment
		Batrachochytrium dendrobatidis[a]	BDEG_03936	BLAST and phylogeny
		Monosiga brevicollis[a]	gi|167524677	
ACD10	Function unknown	*Homo sapiens*	Q6JQN1	cDNA
ACD11	Function unknown	*Homo sapiens*	Q709F0	cDNA
fadE	Oxidation of fatty acids of different chain length	*Escherichia coli*	Q47146	Experiment
fadE12	Oxidation of medium chain fatty acid	*Mycobacterium tuberculosis*	P71539	Experiment

[a]Seeds added in the second round of BLAST search.

homologs often match members from different subfamilies with similar scores. For example, a protein from *Janthinobacterium* (gi|152980951) shares identities of 28% with ACADS from *Mycobacterium*, 27% with ACDSB from rat and 27% with ACADV2 from human. Evidently, such a lack of distinction by similarity scores has hampered research on subfamily distribution, diversity and evolution.

As a large number of complete genome sequences from prokaryotes and eukaryotes have become available, large-scale subfamily classification and phylogenetic analysis of the ACAD family are now tractable. Our first step in this investigation was assignment of ACAD proteins to defined subfamilies. The most direct way to do so is via sequence similarity search as employed in previous protein family studies (23,24). But as illustrated above, it is difficult to distinguish members of different ACAD subfamilies by sequence similarity alone. Another widely used approach employs sequence profiles, e.g. PFAM domains (25,26) or hidden Markov models (HMMs) generated from subfamilies (27,28). But for ACAD enzymes, the number of confirmed sequences in each subfamily is not large enough to make reliable profiles. Here we identify ACAD subfamily members by rigorous ortholog detection via phylogenetic analysis, an approach successfully employed in certain genome annotation and comparison studies (29,30). Our procedure involves reiterative phylogenetic tree construction combined with a two-round BLAST search. Then, based on comprehensive subfamily assignment, we ascertain the taxonomic distribution of ACAD proteins and make inferences of their molecular function and, more generally, the energy sources of a given organism. We also attempt the inference of a global ACAD family tree, which, however, proves by far more difficult than anticipated. Still, for eukaryotic ACAD genes, we have been able to discern several recurring evolutionary patterns that we present in the last section of this article.

MATERIALS AND METHODS

Data collection

We collected the genome-deduced protein sequences of completed or coding regions-completed genome projects from 212 species, mostly taken from NCBI (http://www.ncbi.nlm.nih.gov/Genomes/), Broad Institute of MIT and Harvard (http://www.broad.mit.edu) and DOE Joint Genome Institute (http://www.jgi.doe.gov/). This dataset is composed of 91 bacteria, 29 archaea and 92 eukaryotes. To increase taxonomic coverage in phylogenetic analyses, we included proteins of 32 eukaryotes whose genome sequence is only partially completed, as well as Expressed Sequence Tag (EST) clusters from six jakobids that were generated by us in the context of the Canadian collaborative Protist EST project and retrieved from the Taxonomically Broad eukaryote EST DataBase (TBestDB) (31). Jakobids are a group of heterotrophic flagellates that are believed to diverge close to the eukaryotic origin (32–35). Since no genome sequences are available from jakobids, we included EST data of six jakobid species to obtain a more comprehensive view of the evolution of ACAD enzymes. A detailed list of species names and data sources is compiled in Table S1. Sequences of enzymatically characterized ACAD proteins were retrieved from SwissProt (Table 2) and used as seeds to search for ACAD subfamilies in collected genomes. In addition,

we included sequences of ACOX in the BLAST seed (listed in Table S2) to exclude potential mix-up of ACAD and ACOX homologs.

Subfamily assignment

Orthologs of subfamilies were identified by a two-round procedure combining BLAST search with phylogenetic inference. Each round included BLAST searches and data selection followed by phylogenetic analysis. The difference between round one and two was the set of seed sequences used for BLAST searches. In round one, the genome-deduced protein sequences from each species were compared by BLAST with known ACAD proteins (seeds listed in Table 2), at a threshold of $e = 1 \times 10^{-20}$. In total, 2258 sequences matched at least one seed under this condition. From each species, we selected up to three top matches for each ACAD subfamily and preliminarily annotated the corresponding proteins as potential homologs of the corresponding subfamily. As certain query sequences matched multiple different subfamilies, we analyzed these a second time by applying the following rule: if a given sequence matched multiple subfamilies and the *e*-values of the matches differed by more than 10-fold, then the sequence was assigned to the subfamily with the lowest *e*-value. This case applied to 1572 sequences. Otherwise, if *e*-values of the multiple matches differed less than 10-fold, all preliminary subfamily assignments were retained and the final annotation was based on the subsequent phylogenetic analysis. This category included 341 sequences. We built a maximum likelihood phylogenetic tree for each protein subfamily (see procedure below) using all sequences assigned to this subfamily. From these trees, we selected slowly evolving and unambiguous orthologs of the initial mammalian seed sequences, i.e. proteins from *Monosiga*, the closest unicellular relative of animals (36), and *Batrachochytrium*, a member of the earliest divergence in fungi (36). These, combined with the first set of seeds, formed the second set of seeds used for round two of BLAST searches (Table 2). The same screening procedure was applied as in round one, followed by construction of a phylogenetic tree for each subfamily. Inspection of the trees showed that certain species possessed multiple members of the same subfamily. These extra copies were removed from the dataset to save computational cost during subsequent analyses (especially bootstrap, see below), yielding a non-redundant data set of 861 sequences. In order to detect paralogs, phylogenetic trees were built again for each subfamily, this time with the non-redundant data set. Paralogs were removed from the subfamily until the gene trees were reconciled with the species tree (Figure 2). The sequences removed in this step (32 in total) are considered as ACAD proteins of unknown subfamily (Table S3). A special procedure was applied to ACD10 and ACD11. Numerous potential homologs of ACD10 displayed similar BLAST *e*-values to ACD11 and vice versa, and subfamily assignment as described above was possible only for a few members. The remaining proteins were grouped into a provisional subfamily termed as ACD10/11 and further analyzed by a special procedure as described in the 'Results and Discussion' section.

Phylogenetic inference

Multiple protein sequence alignments were constructed with MUSCLE (37), and alignment logos were created by WebLogo (38). For phylogenetic analysis, ambiguously aligned and highly divergent regions of the alignment were eliminated using Gblocks (39). Maximum likelihood trees were constructed using RAxML (40) with the WAG + Γ model and four discrete γ-rate categories. The statistical support of branches was evaluated by 100 bootstrap replicates.

Protein domain search

To locate functional domains in ACD10 and ACD11 homologs, we used InterProtScan (41). Protein sequences were searched against PROSITE patterns, PROSITE profiles, PRINTS, PFAM, PRODOM, SMART, TIGRFAMs, PIR SuperFamily and SUPERFAMILY.

Taxonomic distribution profiling

After the subfamily assignment of ACAD proteins, we compiled the presence/absence of subfamilies in the genome-derived proteomes of the species included in this study. This information was mapped on NCBI's taxon tree (http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy).

Targeting peptide prediction

We predicted subcellular location of subfamilies based on recognition of targeting peptide from four predictors: TargetP (42), Predotar (43), Protein Prowler (44) and MitoProt (45). Annotation-based predictors, such as PA-SUB (46), were excluded from the analysis to preclude 'prejudicial' association, because ACAD enzymes are traditionally annotated as mitochondrial proteins in public databases. All results were obtained from online servers, except for MitoProt, which was installed and run locally. For most proteins, the predictors gave contradictory results. Therefore, we integrated these predictions via YimLOC, a tool employing machine learning (MTP-DT predictor) (47) that is significantly more accurate than any of the individual predictors. To detect the peroxisomal targeting signal, we used the web service PTS1 predictor (http://mendel.imp.ac.at/mendeljsp/sat/pts1/PTS1predictor.jsp) (48).

RESULTS AND DISCUSSION

Subfamily assignment of previously unclassified ACAD proteins

To identify ACAD subfamilies in the genome-derived proteomes of 250 species (species names are listed in Table S1 and S3), we initially searched for homologs of well-characterized subfamily members by BLAST. This approach failed to distinguish subfamilies in many instances, especially when query and target sequences were from taxonomically distant species. As illustrated by the example of *Janthinobacterium* (see 'Introduction' section), remote homologs often match several subfamilies with similar scores, and the top hit may not correspond to

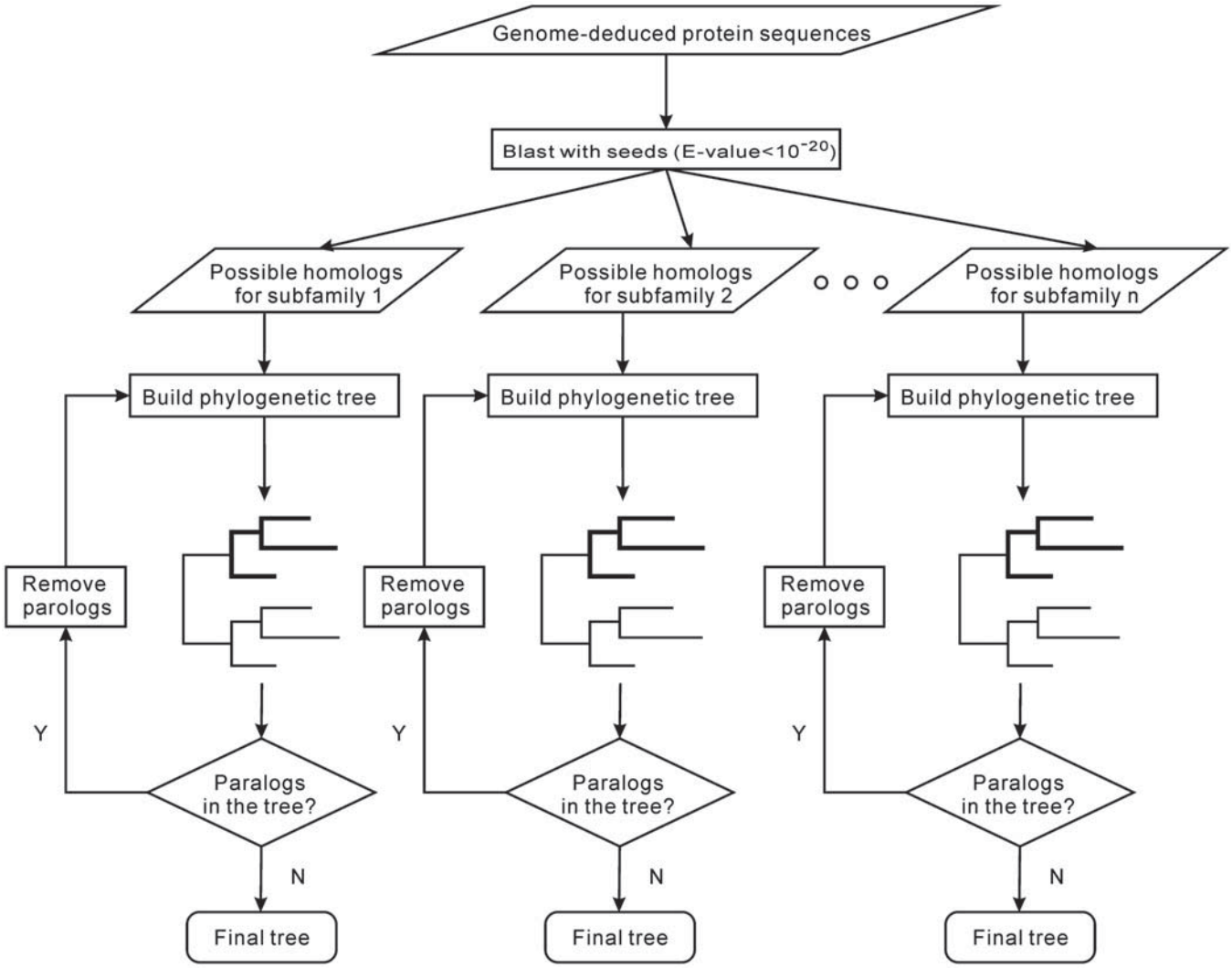


Figure 2. Flow chart of the procedure for assigning ACAD subfamilies. Blast searches combined with reconciliation of gene and species trees were used to identify orthologs (see 'Materials and Methods' section). Y, yes; N, no.

the protein's true affiliation. Therefore, we developed a procedure that combines BLAST searches with phylogenetic analysis, as described in the 'Materials and Methods' section. By this procedure, a total of 702 sequences from 177 species were unambiguously assigned to one of the ACAD subfamilies (Table S1), with the exception of ACD10 and ACD11. As all non-animal proteins that match ACD10 also match ACD11 with similar score and vice versa, these proteins were classified provisionally as ACD10/11 and further analyzed with a distinct approach (see subsequently). Only 32 (4%) proteins, mostly from prokaryotes and protists, remained unassigned (Table S3). These proteins do not form a new subfamily, because sequence similarity was observed only between proteins from the same genus and not across larger phylogenetic distances. Notably, 60 out of the 250 investigated species, predominantly bacteria, appear to completely lack ACAD genes (Table S3). The subfamily distribution across taxa is analyzed in more detail further below.

Distinction of ACD10 and ACD11

The two ACAD subfamilies of unknown function, ACD10 and ACD11, have only recently been discovered in human and a few other mammals (8,9). Our subfamily assignment procedure clearly distinguishes ACD10 and ACD11 in animals, but fails to do so for other taxa. In the tree built with all identified ACD10 and ACD11 sequences and the provisional class ACD10/11 (127 proteins from 110 taxa in total; Figure 3C and Figure S1A), only vertebrate ACD10 and ACD11 form well supported, distinct and coherent clades. Sequences from other taxa cannot be placed with confidence.

Further distinction of ACD10 and ACD11 comes from protein domain analysis. Mammalian ACD10 and ACD11 proteins are conspicuously longer than those from other ACAD subfamilies. Domain search with InterProScan shows that the human ACD11 proteins carry in their N-terminal region an aminoglycoside phosphotransferase (APH) domain. In bacteria, this domain is

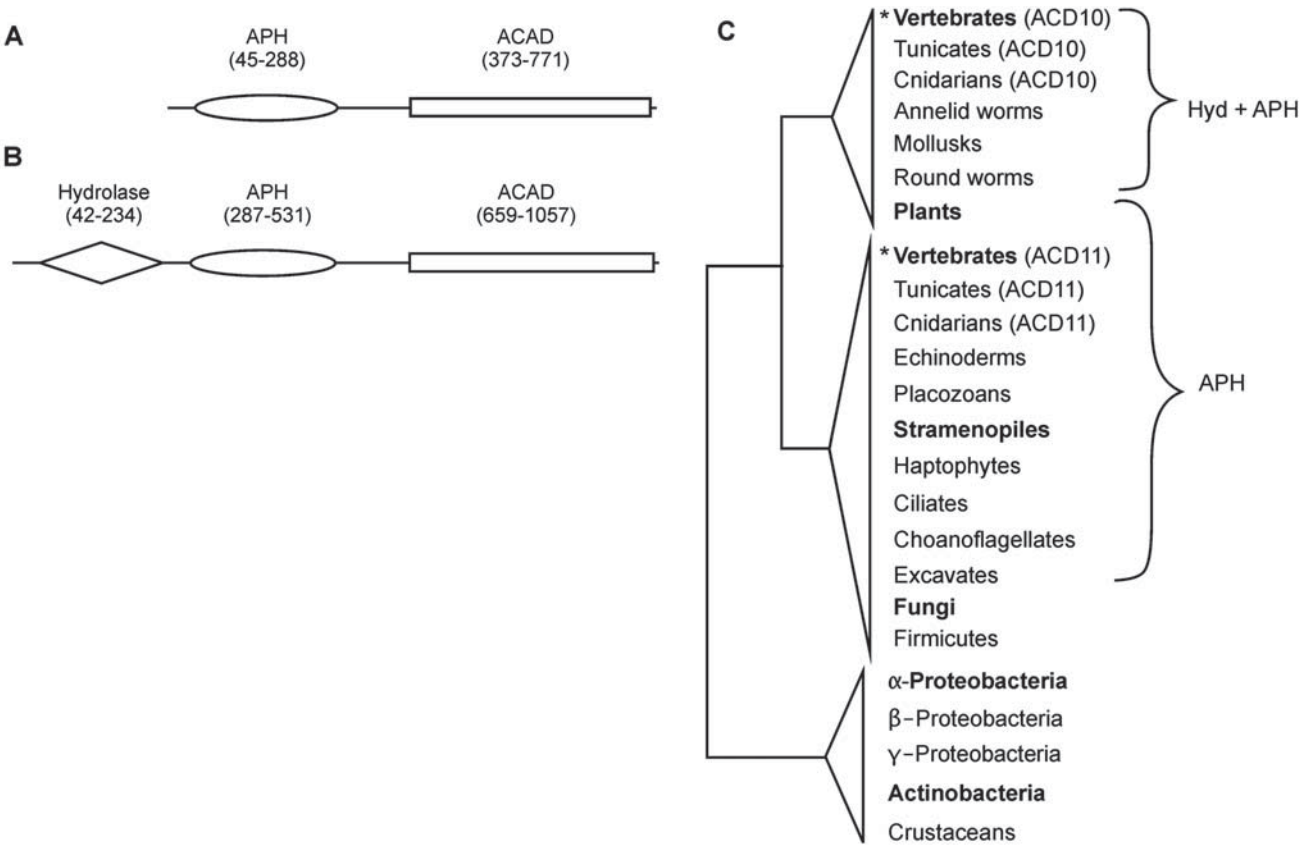


Figure 3. Distinction of ACD10 and ACD11. Protein domains of human ACD11 and ACD10 (Q709F0, **A**) and ACD10 (Q6JQN1, **B**). InterProt domain IDs are as follows: hydrolase domain, IPR005834; APH domain, IPR002575. The ACAD domain is composed of three parts: ACAD N-terminal domain, IPR013786; ACAD central domain, IPR006091; ACAD C-terminal domain, IPR013764. (**C**) Domain content of ACD10, ACD11 and provisional ACD10/11 homologs mapped onto the phylogenetic tree. Taxa representing more than three species are shown in bold. Clades with bootstrap support value >90 are labeled with asterisk. Taxa that appear twice in the tree are distinguished by the labels 'ACD10' and 'ACD11'. In animals, ACD11 includes (in addition to the common ACAD domains) an APH domain, and ACD10 possesses an APH plus a hydrolase (Hyd) domain. Exceptions are gi|115941654 of the echinoderm *Strongylocentrotus purpuratus*, which is more similar to ACD10 but lack the hydrolase domain, and jgi|Dappu1|346313 of the crustacean *Daphnia pulex*, which shares equal sequence similarity with ACD10 and ACD11 and lacks both extra domains. Homologs of other eukaryotes, which have an APH domain, but no hydrolase domain, are classified as ACD11. Sequences lacking both domains are all homologs of fungi, the green algae *Volvox carteri* and *Ostreococcus lucimarinus* and the stramenopiles *Aureococcus anophagefferens* and *Phytophthora ramorum*. Bacterial homologs also lack both domains. Those lacking both domains are classified as ACD11n, see text.

involved in antibiotic resistance (49) (Figure 3A), but its role in eukaryotes is unknown. ACD10 has in addition to APH an N-terminal hydrolase domain (Figure 3B). Both domains are absent from other ACAD families. While screening non-metazoan ACD10/11 for these domains, we did not detect a single protein including the hydrolase domain; the majority of these sequences carry APH and some lack both domains (Figure 3C). Phylogenetic trees of fungal and animal homologs place animal ACD10 and ACD11 into two monophyletic clades to the exclusion of fungal proteins, suggesting that ACD11 is an ancestral eukaryotic gene from which ACD10 has arisen in the animal lineage by gene duplication and subsequent addition of the hydrolase domain (Figure S1B). Therefore, we classify non-metazoan homologs carrying APH as ACD11 and those without either domain as ACD11n.

Are eukaryotic ACD10, ACD11 and ACD11n indeed mitochondrial proteins as traditionally assumed for the entire ACAD family? A proteomics study of rat peroxisomal proteins (50) reports the peptides whose sequence match the mouse homolog of ACD11 according to our subfamily classification (gi|28280023). In addition, an enzymatic study of peroxisomal β-oxidation in *Magnaporthe grisea* speculates that some of the fungus' ACAD proteins are imported into peroxisomes to substitute for the ACOX enzyme, whose gene is missing from the genome (51). Here we predict by *in silico* methods the subcellular localization of all ACAD subfamilies present in *Magnaporthe*. Indeed, ACD11n is the only ACAD member that has the propensity to enter peroxisomes, pinpointing ACD11n as the hypothetical protein participating in peroxisomal β-oxidation. The same situation most likely applies to other fungi that lack ACOX in their genome, specifically *Nectria haematococca*, *Hypocrea jecorina* and *Hypocrea virens* (22). This hypothesis can be readily tested experimentally, e.g. by co-localization of tagged ACAD protein with subcellular

structures. If true, the reaction mechanism of ACD11n must have undergone a fundamental adaptation to the peroxisomal environment.

Notably, peroxisomal localization is predicted for most eukaryotic members of ACD11n and the entire ACD11 subfamily (exceptions are listed in Table S4), while ACD10 displays features typical of mitochondrial proteins. The predicted subcellular localization should guide experimental approaches to elucidate these proteins' molecular function and the specific roles of the APH and hydrolase domains.

Taxonomic distribution of ACAD subfamilies

Based on the comprehensive annotation, we examined the presence/absence of ACAD subfamilies across the 250 species investigated here. The distribution profiles of subfamilies differ markedly (Figure 4). Large sets of ACAD subfamilies are typical for animals with 11 in vertebrates, as many as initially identified in mammals. In contrast, fungi possess on average only five subfamilies and these are involved in both mitochondrial β-oxidation and amino acid catabolism. A total lack of ACAD genes is observed in a few fungal lineages (*Saccharomyces*, *Encephalitozoon* and *Schizosaccharomyces*), all characterized by highly derived and reduced genomes. In Plantae, only two ACAD subfamilies, IVD and ACD11, are widely present. From the other eukaryotic lineages, there are not enough genome sequences available to infer specific profile features. Finally, Archaea have conspicuously small sets of ACAD subfamilies, and there is much variation among Bacteria.

Four subfamilies occur in all domains of life: ACADS (degrading short fatty acids), ACADM, fadE12 (both preferring medium-length fatty acids) and GCDH (involved in lysine and tryptophan catabolism). Subfamilies virtually restricted to a single domain are ACDSB, ACADV and ACADV2 in eukaryotes, and fadE in bacteria. Overall, ACAD subfamilies specialized in short-chain (straight and branched) acyl-CoAs are more broadly distributed than those preferring long-chain substrates.

We confronted the inferred ACAD subfamily profiles with experimental evidence. As mentioned earlier, animals possess 11 out of 13 ACAD subfamilies (Figure 4). Indeed, fatty acids and amino acids make up an important part of metazoan nutrition, requiring a host of specialized enzymes for degrading acyl-substrates of various length and steric structure. A comparably large repertoire of ACAD enzymes (the largest in bacteria) is present in the opportunistic human pathogen, *Pseudomonas aeruginosa*. This organism is notorious for its extraordinary metabolic versatility, capable of utilizing a wide range of organic compounds including fatty acids and amino acids as an energy source. The ACAD families in *P. aeruginosa* identified here explain the observed efficient use of fatty acids via β-oxidation (52,53). Only a single ACAD subfamily—ACADS—is found in *Clostridium botulinum*, a food-borne pathogen. Experimental studies confirm that *C. botulinum* cannot catabolize long-chain fatty acids. This explains the documented poor growth of this bacterium on ripened cheese (54). Finally, many intracellular parasites such as *Rickettsiaceae* lack all ACAD genes (Figure 4), reflecting their reliance on their host for nutrients. In sum, the above examples illustrate that the *in silico*-generated ACAD profiles are in strong agreement with, and explain well, the biochemical data. Therefore, the presence of ACAD subfamilies provides a window on an organism's biology based on genome sequence alone, when information on nutritional requirements and enzymatics are not available.

Multiple single-purpose enzymes versus single multi-purpose enzymes

Recent biochemical studies on *Aspergillus nidulans* revealed an unexpected substrate range of ACDSB, which in this organism catalyzes dehydrogenation of not only isobutyryl-CoA (derivative of isoleucine), but also 2-methyl-butyryl-CoA (derivative of valine) and short-chain acyl-CoA (55). In human, the latter two compounds are degraded by IBD and ACADS (Figure 1), two enzymes that are missing intriguingly in *Aspergillus* species and other fungi (Figure 4B).

Insight into the molecular basis of such a broad substrate range comes from directed mutagenesis experiments of the human ACDSB protein (5). This study shows that the substitutions Ser177Asn, Leu222Ile and Ala383Thr lead to significantly higher turnover rates of hexanoyl-CoA and isobutyryl-CoA (the substrates optimally degraded by ACADS and IBD, respectively). This pinpoints Ser177, Leu222 and Ala383 as substrate specificity determinants of human ACDSB (in the following, the 'specificity' residues are indicated as **NIT** and **SLA**).

To find out the substrate range of ACDSB from other organisms where experimental data are lacking, we constructed a multiple sequence alignment and also superimposed the 3D structure of ACDSB, ACADS and IBD proteins (Figure S2). These alignments show that homologs from all primates and a few other mammals are of the human type (ACDSB-**SLA**), whereas all fungal enzymes are of the *Aspergillus* type (ACDSB-**NIT**, with a single minor exception, Figure 5). From this finding, together with the subfamily distribution pattern, we predict that all fungal ACDSB unite three functions in one enzyme, i.e. the functions of human ACDSB, ACADS and IBD. (For a hypothesis on the evolution of these three subfamilies, see Figure S3.)

The above functional generalization has previously remained undetected by sequence similarity and phylogeny-based function prediction. But as exemplified here, the prediction of an enzyme's substrate range can be improved by integrating subfamily distribution profiles and function/structure data from 'model' enzymes. Such advanced function prediction provides valuable working hypotheses that can be tested by targeted biochemical experimentation.

Evolution of the ACAD family

In an attempt to unravel the origin and evolution of the ACAD protein family, we built maximum likelihood phylogenetic trees using proteins drawn from complete genome sequences and assigned to subfamilies as

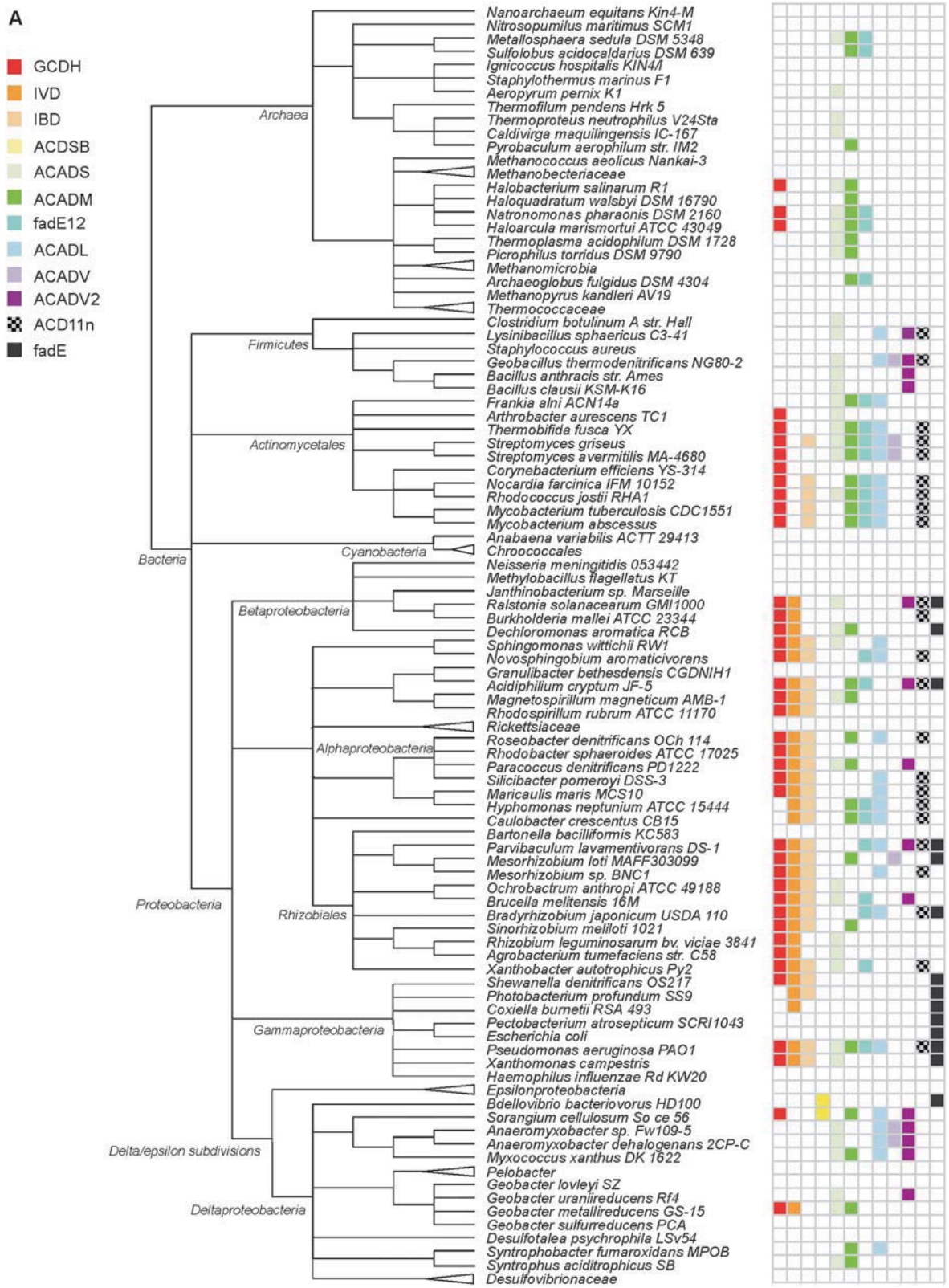


Figure 4. ACAD subfamily distribution mapped on the taxonomy hierarchy from NCBI. Only species whose genome has been completely sequenced are included in the figure. The sequence IDs are listed in Table S1. A triangle in front of a taxon name indicates that no ACAD subfamily was detected in the members of this taxon. (**A**) Subfamily distribution in prokaryotes; (**B**) subfamily distribution in eukaryotes.

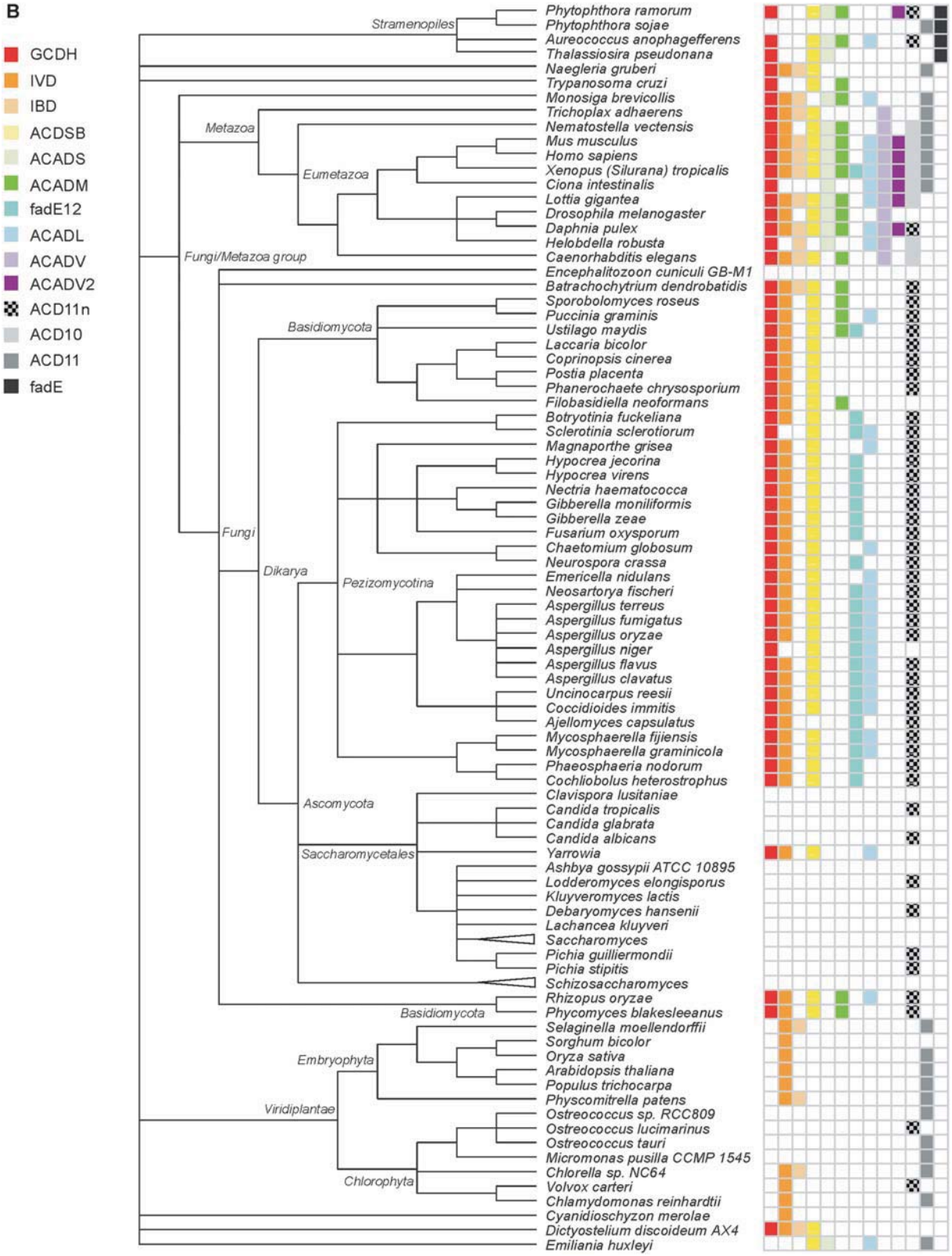


Figure 4. Continued.

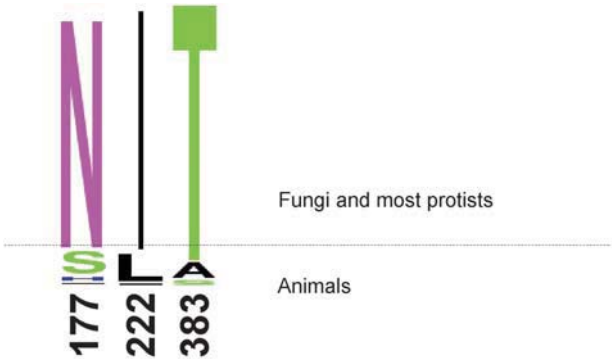


Figure 5. Alignment of residues that affect substrate specificity of human ACDSB. The multiple protein alignment was generated using eukaryotic ACDSB homologs. The numbers refer to the residue position of the mature human ACDSB (i.e. not including the mitochondrial targeting peptide). A minor deviation from the **NIT** motif is found in *Puccinia graminis*, which has a '**NIS**'.

described above (for the proteins used, see Table S1). The global trees including all ACAD subfamilies have largely unsupported topologies likely due to three factors: the extensive sequence divergence, the relatively short length of ACAD proteins (100 or less residues after Gblocks processing, see the 'Materials and Methods' section) and the immense evolutionary time spans in question. Trees of individual ACAD subfamilies suffer to a lesser degree from this problem, but only a few trees (ACDSB and ACADV2) display supported species-tree topology (Figure S1C–M). Close inspection reveals events of horizontal gene transfers, not only within bacteria but also within eukaryotes and across domains, in addition to gene losses and multiple independent gene duplications. After carefully studying each subfamily tree, we are able to discern several intriguing evolutionary patterns in eukaryotes.

Early acquisition of ACAD enzymes in eukaryotes. ACADS, GCDH, IVD and IBD occur in eukaryotes and bacteria (Figure 4). Subfamily trees unite eukaryotes mainly with α-Proteobacteria, to the exclusion of other prokaryotes. This trend is best supported by GCDH and IBD (Figure 6A). The tree topology, together with the taxonomic distribution and the predicted mitochondrial localization of these subfamilies (data not shown), suggests that eukaryotes acquired these genes from α-Proteobacteria via the endosymbiotic event leading to mitochondria. Yet, our current single-protein tree topologies have numerous branches with weak statistical support, and trees built with concatenated sequences of GCDH, IVD, IBD and ACADS do not provide more information either (Figure S1N). A rigorous test of this hypothesis would have to rule out horizontal transfer of the corresponding genes in bacteria, and improve tree robustness by substantially expanded taxon sampling.

Loss of ACAD genes in fungi and recent recruitment from α-Proteobacteria. The two ACAD subfamilies fadE12 (typically prokaryotic) and ACADM (eukaryotic) have the same substrate specificity (Figure 4) (18). Pezizomycotina (including species such as *Neurospora* and *Aspergillus*) are an intriguing exception: they lack ACADM, but possess fadE12. Phylogenetic analysis of fadE12 proteins unites fungal and α-proteobacterial sequences with high support to the exclusion of other bacteria, strongly suggesting an α-proteobacterial origin of the Pezizomycotina genes (Figure 6B and Figure S1I; see legend of Figure 6 for exceptions). Based on the phylogeny and ACAD subfamily distribution, we propose that initially all fungi possessed ACADM, but this gene was later lost in the common ancestor of Ascomycota. After the ascomycete lineages had diverged, the predecessor of Pezizomycotina acquired the functionally equivalent fadE12 via horizontal gene transfer from α-Proteobacteria. A similar history involving loss and secondary acquisition of a bacterial ACAD gene by fungi apparently applies to ACADL (Figure S1J).

Duplication of ACAD genes in mammals and recent transfer to other lineages. As mentioned earlier, the two ACAD subfamilies degrading very long chain fatty acids, ACADV and ACADV2, are predominantly present in animals, with only a few exceptions in non-mammalian eukaryotes (i.e. *Phytophthora* species) and diverse bacteria (Figure 4). The phylogenetic tree including both subfamilies separates animal ACADV and ACADV2 into two well supported, distinct and coherent clades, indicating a gene duplication event prior to the divergence of animals (Figure 6C and Figure S1O). Homologs of *Phytophthora* affiliate (with moderate support) with animal ACADV2, and the same topology, now with 99% bootstrap support, are seen in the tree based on ACADV2 only (Figure S1L). The most parsimonious explanation is that the common ancestor of these oomycetes has acquired ACADV2 from animals, and transmitted it vertically to extant *Phytophthora* species.

Conclusion

Our study of the large and biologically important ACAD protein family integrates three types of information, taxonomic distribution profiles, subfamily phylogenies as well as functional and structural data from model proteins. This allows analyses of broad scope leading to improved molecular function prediction of individual ACAD subfamilies, formulation of working hypotheses for targeted biochemical experimentation as well as to the discovery of a most 'turbulent' evolutionary history of the ACAD gene family. A study like this one relies critically on a rigorous method for identification of orthologs for each paralogous subfamily as we devised in this report.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Henner Brinkmann and Dr Nicolas Lartillot (Université de Montréal) and Yu Liu (University of

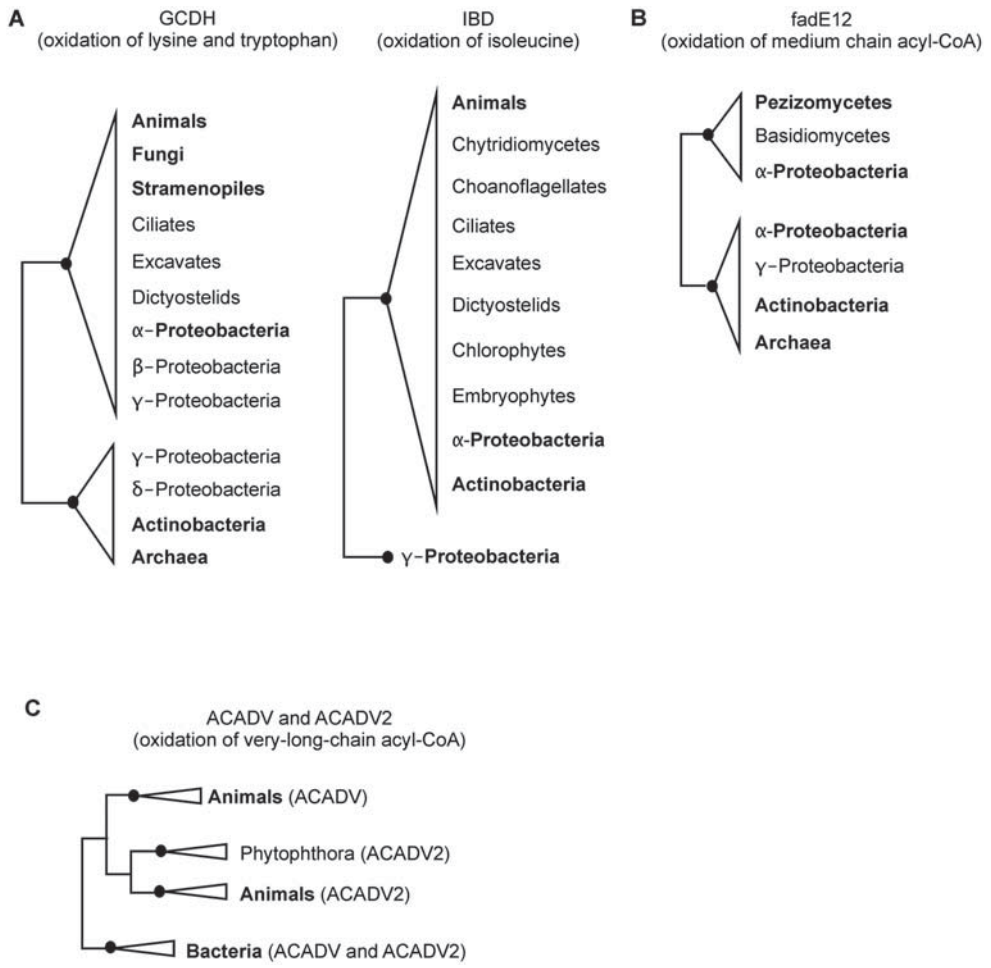


Figure 6. Schematic phylogenetic trees of ACAD subfamilies. The underlying explicit trees are provided in Figure S1. Branches with a bootstrap value ≥90 are labeled with filled dots. Taxa representing more than three species are shown in bold. (**A**) ACAD subfamilies with likely α-proteobacterial origin. In the trees of GCDH (left, Figure S1C) and IBD (right, Figure S1E), eukaryotic homologs group together with those from α-Proteobacteria. In the GCDH tree, eukaryotic and numerous α-proteobacterial taxa form a well-supported clade to the exclusion of Archaea plus Actinobacteria; both clusters include a few other Proteobacteria. The IBD tree unites eukaryotes and α-Proteobacteria to the exclusion of a few γ-Proteobacteria. (**B**) Secondary acquisition of α-proteobacterial homologs by certain fungal lineages. Some Basidiomycota and all investigated Ascomycota lack ACADM (Figure S1H). The ascomycete class Pezizomycotina possesses fadE12 (of same function as ACADM, Figure S1I) that associates strongly with α-Proteobacteria. Exceptions are *Emericella nidulans*, *M. grisea* and *Chaetomium globosum*, where fadE12 is absent. *Ustilago maydis*, the single basidiomycete possessing fadE12, likely acquired this gene from Pezizomycotina. (**C**) Gene duplication in animals and lateral transfer to other taxa. ACADV and ACADV2 are paralogs originating from a gene duplication prior to the divergence of animals (Figure S1O). The few bacterial ACADV homologs form a monophyletic clade, to the exclusion of animal proteins. ACADV2 from *Phytophthora* groups with animal homologs.

Toronto) for help in phylogenetic analyses. We also thank Emmet O'Brien (Université de Montréal) and the anonymous referees for improving the manuscript.

FUNDING

Canadian Institute for Advanced Research (CIfAR, salary support granted to G.B.). Y.-Q.S. is a Canadian Institute for Health Research (CIHR) Strategic Training Fellow in Bioinformatics, and B.F.L. holds a Canadian Research Chair. Funding for open access charge: Canadian Institute for Health Research, Institute of Genetics.

Conflict of interest statement. None declared.

REFERENCES

1. Gregersen,N., Bross,P. and Andresen,B.S. (2004) Genetic defects in fatty acid beta-oxidation and acyl-CoA dehydrogenases. Molecular pathogenesis and genotype-phenotype relationships. *Eur. J. Biochem.*, **271**, 470–482.
2. Ijlst,L. and Wanders,R.J. (1993) A simple spectrophotometric assay for long-chain acyl-CoA dehydrogenase activity measurements in human skin fibroblasts. *Ann. Clin. Biochem.*, **30(Pt 3)**, 293–297.
3. Tiffany,K.A., Roberts,D.L., Wang,M., Paschke,R., Mohsen,A.W., Vockley,J. and Kim,J.J. (1997) Structure of human isovaleryl-CoA dehydrogenase at 2.6 A resolution: structural basis for substrate specificity. *Biochemistry*, **36**, 8455–8464.
4. Udvari,S., Bross,P., Andresen,B.S., Gregersen,N. and Engel,P.C. (1999) Biochemical characterisation of mutations of human medium-chain acyl-CoA dehydrogenase. *Adv. Exp. Med. Biol.*, **466**, 387–393.

5. He,M., Burghardt,T.P. and Vockley,J. (2003) A novel approach to the characterization of substrate specificity in short/branched chain Acyl-CoA dehydrogenase. *J. Biol. Chem.*, **278**, 37974–37986.
6. Battaile,K.P., Nguyen,T.V., Vockley,J. and Kim,J.J. (2004) Structures of isobutyryl-CoA dehydrogenase and enzyme-product complex: comparison with isovaleryl- and short-chain acyl-CoA dehydrogenases. *J. Biol. Chem.*, **279**, 16526–16534.
7. Fu,Z., Wang,M., Paschke,R., Rao,K.S., Frerman,F.E. and Kim,J.J. (2004) Crystal structures of human glutaryl-CoA dehydrogenase with and without an alternate substrate: structural bases of dehydrogenation and decarboxylation reactions. *Biochemistry*, **43**, 9674–9684.
8. Ota,T., Suzuki,Y., Nishikawa,T., Otsuki,T., Sugiyama,T., Irie,R., Wakamatsu,A., Hayashi,K., Sato,H., Nagai,K. *et al.* (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet.*, **36**, 40–45.
9. Ye,X., Ji,C., Zhou,C., Zeng,L., Gu,S., Ying,K., Xie,Y. and Mao,Y. (2004) Cloning and characterization of a human cDNA ACAD10 mapped to chromosome 12q24.1. *Mol. Biol. Rep.*, **31**, 191–195.
10. Saenger,A.K., Nguyen,T.V., Vockley,J. and Stankovich,M.T. (2005) Thermodynamic regulation of human short-chain acyl-CoA dehydrogenase by substrate and product binding. *Biochemistry*, **44**, 16043–16053.
11. Merritt,J.L. 2nd, Matern,D., Vockley,J., Daniels,J., Nguyen,T.V. and Schowalter,D.B. (2006) In vitro characterization and in vivo expression of human very-long chain acyl-CoA dehydrogenase. *Mol. Genet. Metab.*, **88**, 351–358.
12. Ensenauer,R., He,M., Willard,J.M., Goetzman,E.S., Corydon,T.J., Vandahl,B.B., Mohsen,A.W., Isaya,G. and Vockley,J. (2005) Human acyl-CoA dehydrogenase-9 plays a novel role in the mitochondrial ß-oxidation of unsaturated fatty acids. *J. Biol. Chem.*, **280**, 32309–32316.
13. Ikeda,Y., Okamura-Ikeda,K. and Tanaka,K. (1985) Purification and characterization of short-chain, medium-chain, and long-chain acyl-CoA dehydrogenases from rat liver mitochondria. Isolation of the holo- and apoenzymes and conversion of the apoenzyme to the holoenzyme. *J. Biol. Chem.*, **260**, 1311–1325.
14. Izai,K., Uchida,Y., Orii,T., Yamamoto,S. and Hashimoto,T. (1992) Novel fatty acid beta-oxidation enzymes in rat liver mitochondria. I. Purification and properties of very-long-chain acyl-coenzyme A dehydrogenase. *J. Biol. Chem.*, **267**, 1027–1033.
15. Roe,C.R. and Ding,J. (2001) *The Metabolic and Molecular Bases of Inherited Disease*. McGraw-Hill Book Co., New York.
16. Pauli,G. and Overath,P. (1972) ato Operon: a highly inducible system for acetoacetate and butyrate degradation in *Escherichia coli*. *Eur. J. Biochem.*, **29**, 553–562.
17. Iram,S.H. and Cronan,J.E. (2006) The beta-oxidation systems of *Escherichia coli* and *Salmonella enterica* are not functionally equivalent. *J. Bacteriol.*, **188**, 599–608.
18. Mahadevan,U. and Padmanaban,G. (1998) Cloning and expression of an acyl-CoA dehydrogenase from *Mycobacterium tuberculosis*. *Biochem. Biophys. Res. Commun.*, **244**, 893–897.
19. Ghisla,S. and Thorpe,C. (2004) Acyl-CoA dehydrogenases. A mechanistic overview. *Eur. J. Biochem.*, **271**, 494–508.
20. Kim,J.J. and Miura,R. (2004) Acyl-CoA dehydrogenases and acyl-CoA oxidases. Structural basis for mechanistic similarities and differences. *Eur. J. Biochem.*, **271**, 483–493.
21. Cornell,M.J., Alam,I., Soanes,D.M., Wong,H.M., Hedeler,C., Paton,N.W., Rattray,M., Hubbard,S.J., Talbot,N.J. and Oliver,S.G. (2007) Comparative genome analysis across a kingdom of eukaryotic organisms: specialization and diversification in the fungi. *Genome Res.*, **17**, 1809–1822.
22. Shen,Y.Q. and Burger,G. (2009) Plasticity of a key metabolic pathway in fungi. *Funct. Integr. Genomics*, **9**, 145–151.
23. Kondrashov,F.A., Koonin,E.V., Morgunov,I.G., Finogenova,T.V. and Kondrashova,M.N. (2006) Evolution of glyoxylate cycle enzymes in Metazoa: evidence of multiple horizontal transfer events and pseudogene formation. *Biol. Direct*, **1**, 31.
24. Pereto,J., Lopez-Garcia,P. and Moreira,D. (2005) Phylogenetic analysis of eukaryotic thiolases suggests multiple proteobacterial origins. *J. Mol. Evol.*, **61**, 65–74.
25. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
26. Brown,D.P., Krishnamurthy,N. and Sjolander,K. (2007) Automated protein subfamily identification and classification. *PLoS Comput. Biol.*, **3**, e160.
27. Hannenhalli,S.S. and Russell,R.B. (2000) Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.*, **303**, 61–76.
28. Weston,J., Leslie,C., Ie,E., Zhou,D., Elisseeff,A. and Noble,W.S. (2005) Semi-supervised protein classification using cluster kernels. *Bioinformatics*, **21**, 3241–3247.
29. Hubbard,T.J., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
30. Remm,M., Storm,C.E. and Sonnhammer,E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.
31. O'Brien,E.A., Koski,L.B., Zhang,Y., Yang,L., Wang,E., Gray,M.W., Burger,G. and Lang,B.F. (2007) TBestDB: a taxonomically broad database of expressed sequence tags (ESTs). *Nucleic Acids Res.*, **35**, D445–D451.
32. Gray,M.W., Lang,B.F., Cedergren,R., Golding,G.B., Lemieux,C., Sankoff,D., Turmel,M., Brossard,N., Delage,E., Littlejohn,T.G. *et al.* (1998) Genome structure and gene content in protist mitochondrial DNAs. *Nucleic Acids Res.*, **26**, 865–878.
33. Gray,M.W., Burger,G. and Lang,B.F. (1999) Mitochondrial evolution. *Science*, **283**, 1476–1481.
34. Lang,B.F., Burger,G., O'Kelly,C.J., Cedergren,R., Golding,G.B., Lemieux,C., Sankoff,D., Turmel,M. and Gray,M.W. (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature*, **387**, 493–497.
35. Palmer,J.D. (1997) Genome evolution. The mitochondrion that time forgot. *Nature*, **387**, 454–455.
36. Keeling,P.J., Burger,G., Durnford,D.G., Lang,B.F., Lee,R.W., Pearlman,R.E., Roger,A.J. and Gray,M.W. (2005) The tree of eukaryotes. *Trends Ecol. Evol.*, **20**, 670–676.
37. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
38. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
39. Castresana,J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540–552.
40. Stamatakis,A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.
41. Mulder,N. and Apweiler,R. (2007) InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol. Biol.*, **396**, 59–70.
42. Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
43. Small,I., Peeters,N., Legeai,F. and Lurin,C. (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, **4**, 1581–1590.
44. Boden,M. and Hawkins,J. (2005) Prediction of subcellular localization using sequence-biased recurrent networks. *Bioinformatics*, **21**, 2279–2286.
45. Claros,M.G. and Vincens,P. (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.*, **241**, 779–786.
46. Lu,Z., Szafron,D., Greiner,R., Lu,P., Wishart,D.S., Poulin,B., Anvik,J., Macdonell,C. and Eisner,R. (2004) Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics*, **20**, 547–556.
47. Shen,Y.Q. and Burger,G. (2007) 'Unite and conquer': enhanced prediction of protein subcellular localization by integrating multiple specialized tools. *BMC Bioinformatics*, **8**, 420.
48. Neuberger,G., Maurer-Stroh,S., Eisenhaber,B., Hartig,A. and Eisenhaber,F. (2003) Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. *J. Mol. Biol.*, **328**, 581–592.

49. Nurizzo,D., Shewry,S.C., Perlin,M.H., Brown,S.A., Dholakia,J.N., Fuchs,R.L., Deva,T., Baker,E.N. and Smith,C.A. (2003) The crystal structure of aminoglycoside-3′-phosphotransferase-IIa, an enzyme responsible for antibiotic resistance. *J. Mol. Biol.*, **327**, 491–506.
50. Kikuchi,M., Hatano,N., Yokota,S., Shimozawa,N., Imanaka,T. and Taniguchi,H. (2004) Proteomic analysis of rat liver peroxisome: presence of peroxisome-specific isozyme of Lon protease. *J. Biol. Chem.*, **279**, 421–428.
51. Wang,Z.Y., Soanes,D.M., Kershaw,M.J. and Talbot,N.J. (2007) Functional analysis of lipid metabolism in Magnaporthe grisea reveals a requirement for peroxisomal fatty acid beta-oxidation during appressorium-mediated plant infection. *Mol. Plant Microbe Interact.*, **20**, 475–491.
52. Kang,Y., Nguyen,D.T., Son,M.S. and Hoang,T.T. (2008) The *Pseudomonas aeruginosa* PsrA responds to long-chain fatty acid signals to regulate the fadBA5 beta-oxidation operon. *Microbiology*, **154**, 1584–1598.
53. Son,M.S., Matthews,W.J. Jr, Kang,Y., Nguyen,D.T. and Hoang,T.T. (2007) In vivo evidence of *Pseudomonas aeruginosa* nutrient acquisition and pathogenesis in the lungs of cystic fibrosis patients. *Infect. Immun.*, **75**, 5313–5324.
54. Grecz,N., Wagenaar,R.O. and Dack,G.M. (1959) Relation of fatty acids to the inhibition of *Clostridium botulinum* in aged surface ripened cheese. *Appl. Microbiol.*, **7**, 228–234.
55. Maggio-Hall,L.A., Lyne,P., Wolff,J.A. and Keller,N.P. (2008) A single acyl-CoA dehydrogenase is required for catabolism of isoleucine, valine and short-chain fatty acids in *Aspergillus nidulans*. *Fungal Genet. Biol.*, **45**, 180–189.

Supplementary tables

(Supplementary Table 1 and Table 3 mainly compile sequence IDs, and each table is more than two pages long. Therefore instead of the tables, their links are provided.)

TableS1 List of proteins assigned to ACAD subfamilies in this study

http://nar.oxfordjournals.org/content/vol0/issue2009/images/data/gkp566/DC1/nar-01024-h-2009-File009.xls

Table S2 ACOX sequences used as BLAST seeds

ACOX subfamilies	SwissPort ID	Species name
ACOX1	O65202	*Arabidopsis thaliana*
ACOX1	Q3SZP5	*Bos taurus*
ACOX1	Q9Z1N0	*Cavia porcellus*
ACOX1	Q15067	*Homo sapiens*
ACOX1	Q9R0H0	*Mus musculus*
ACOX1	Q8HYL8	*Phascolarctos cinereus*
ACOX1	Q5RC19	*Pongo abelii*
ACOX1	P07872	*Rattus norvegicus*
ACOX1	O74934	*Yarrowia lipolytica*
ACOX2	O65201	*Arabidopsis thaliana*
ACOX2	Q00468	*Candida maltosa*
ACOX2	P11356	*Candida tropicalis*
ACOX2	O64894	*Cucurbita maxima*
ACOX2	Q99424	*Homo sapiens*
ACOX2	Q9QXD1	*Mus musculus*
ACOX2	O02767	*Oryctolagus cuniculus*
ACOX2	P97562	*Rattus norvegicus*
ACOX2	O74935	*Yarrowia lipolytica*
ACOX3	Q9LLH9	*Arabidopsis thaliana*
ACOX3	O15254	*Homo sapiens*
ACOX3	Q9EPL9	*Mus musculus*
ACOX3	Q5RAU0	*Pongo abelii*
ACOX3	Q63448	*Rattus norvegicus*
ACOX3	O74936	*Yarrowia lipolytica*
ACOX4	Q96329	*Arabidopsis thaliana*
ACOX4	P05335	*Candida maltosa*
ACOX4	P06598	*Candida tropicalis*
ACOX5	P08790	*Candida tropicalis*
ACOXL	Q9NUZ1	*Homo sapiens*
ACOXL	Q9DBS4	*Mus musculus*
ACOX	Q756A9	*Ashbya gossypii*
ACOX	P34355	*Caenorhabditis elegans*
ACOX	Q6FY63	*Candida glabrata*
ACOX	Q6BRD5	*Debaryomyces hansenii*
ACOX	Q6CKK7	*Kluyveromyces lactis*
ACOX	Q9Y7B1	*Pichia pastoris*
ACOX	P13711	*Saccharomyces cerevisiae*

Table S3 species without ACAD genes, and BLAST results for ACAD members of unknown subfamily

http://nar.oxfordjournals.org/content/vol0/issue2009/images/data/gkp566/DC1/nar-01024-h-2009-File011.xls

Table S4. ACD10, ACD11, and ACD11n homologs that have different targeting siganls than the majority of their group

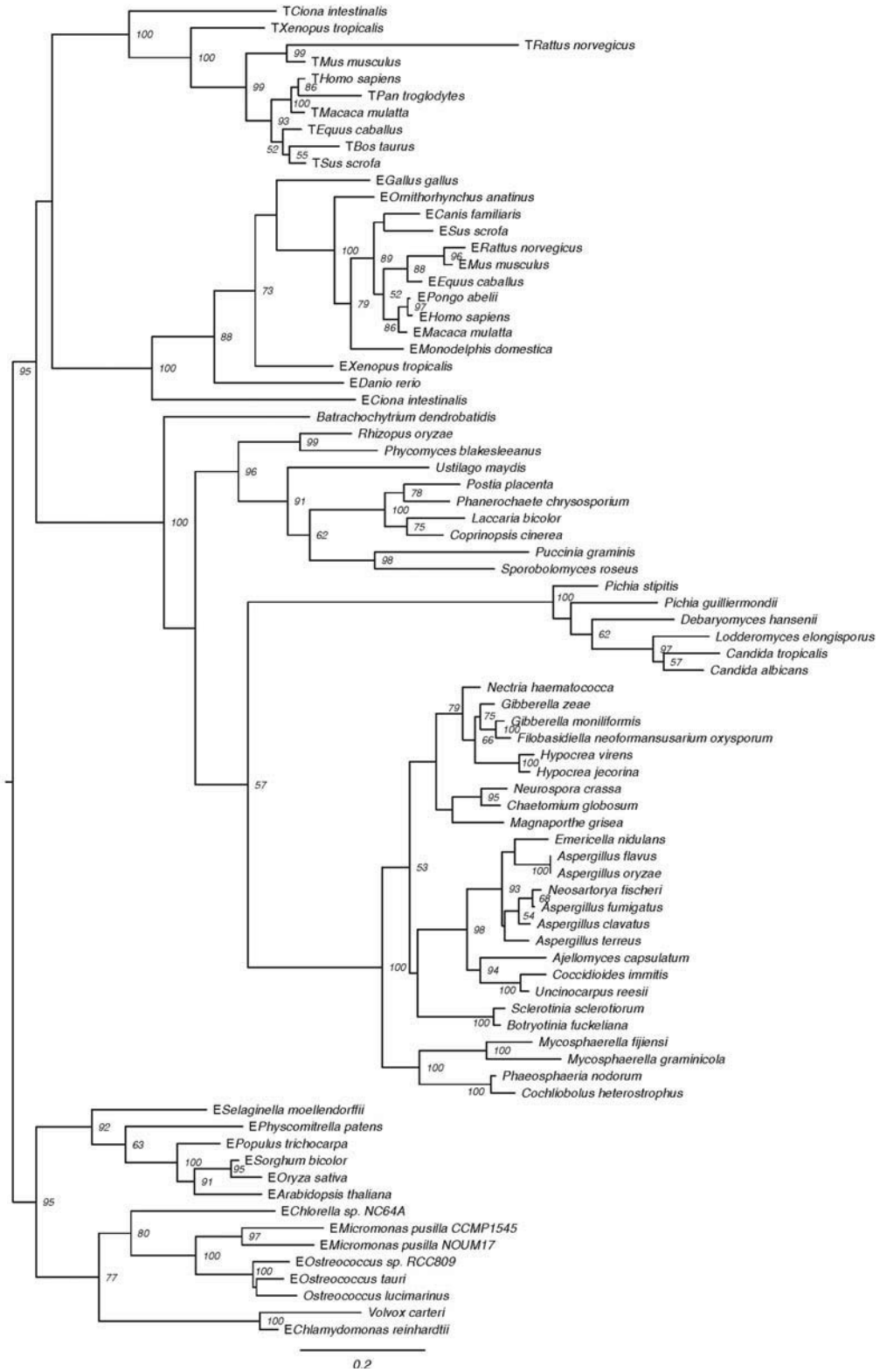
Species name	Subfamily	MTP	PTS
Ciona intestinalis	ACD10	n	n
Daphnia pulex	ACD10	n	n
Nematostella vectensis	ACD10	n	n
Rattus norvegicus	ACD10	n	n
Strongylocentrotus purpuratus	ACD10	n	n
Equus caballus	ACD11	n	n
Strongylocentrotus purpuratus	ACD11	n	n
Micromonas pusilla CCMP 1545	ACD11	y	n
Ostreococcus sp. RCC809	ACD11	y	y
Chlamydomonas reinhardtii	ACD11	n	n
Selaginella moellendorffii	ACD11	n	n
Ajellomyces capsulatus	ACD11n	n	n
Aspergillus flavus	ACD11n	n	n
Aspergillus oryzae	ACD11n	n	n
Batrachochytrium dendrobatidis	ACD11n	n	n
Candida tropicalis	ACD11n	n	n
Chaetomium globosum	ACD11n	n	n
Coccidioides immitis	ACD11n	n	n
Coprinopsis cinerea	ACD11n	n	n
Hypocrea virens	ACD11n	y	y
Laccaria bicolor	ACD11n	n	n
Pichia guilliermondii	ACD11n	n	n
Pichia stipitis	ACD11n	n	n
Postia placenta	ACD11n	n	n
Puccinia graminis	ACD11n	n	n
Uncinocarpus reesii	ACD11n	n	n
Ustilago maydis	ACD11n	n	n
Tetrahymena thermophila	ACD11	n	n
Aureococcus anophagefferens	ACD11n	y	n
Emiliania huxleyi	ACD11	y	n

MTP: mitochondrial targeting peptide; PST: peroxisomal targeting peptide

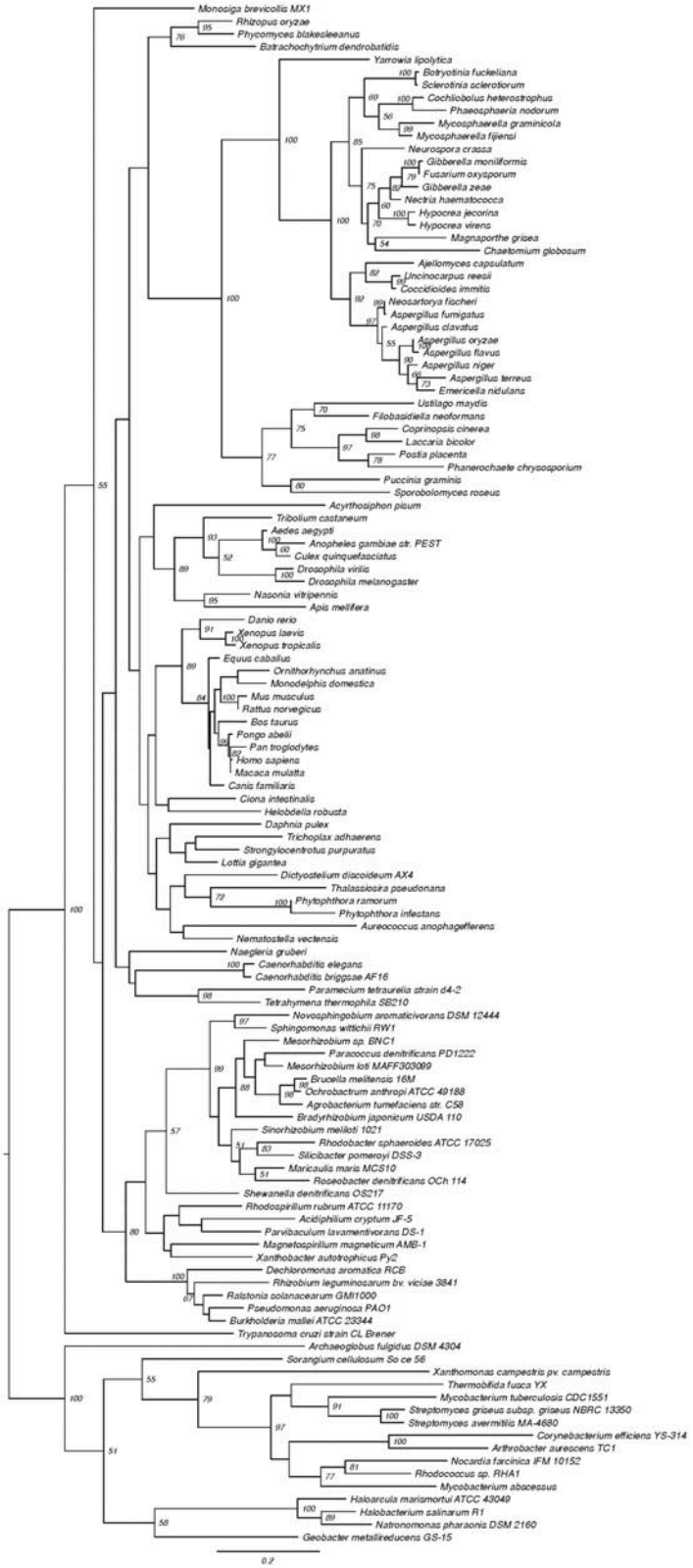
Supplementary figures

A ACD10, ACD11, and ACD11n

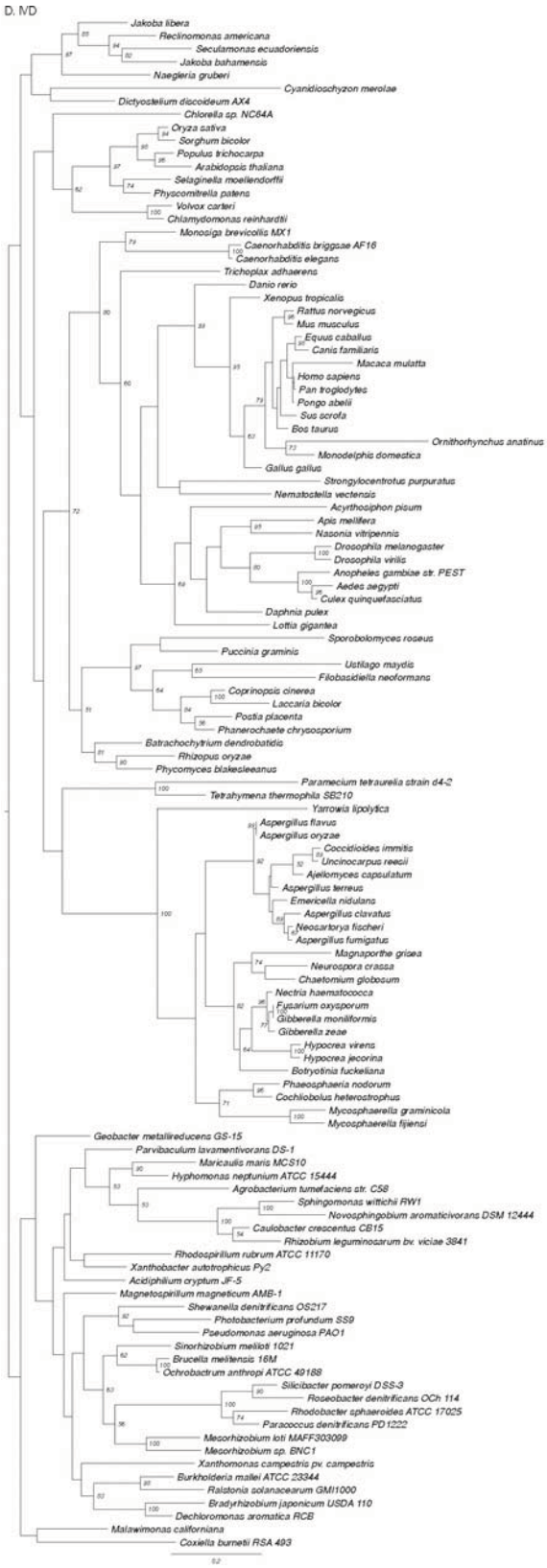
B ACD10, ACD11, and ACD11n in animals, plants, and fungi

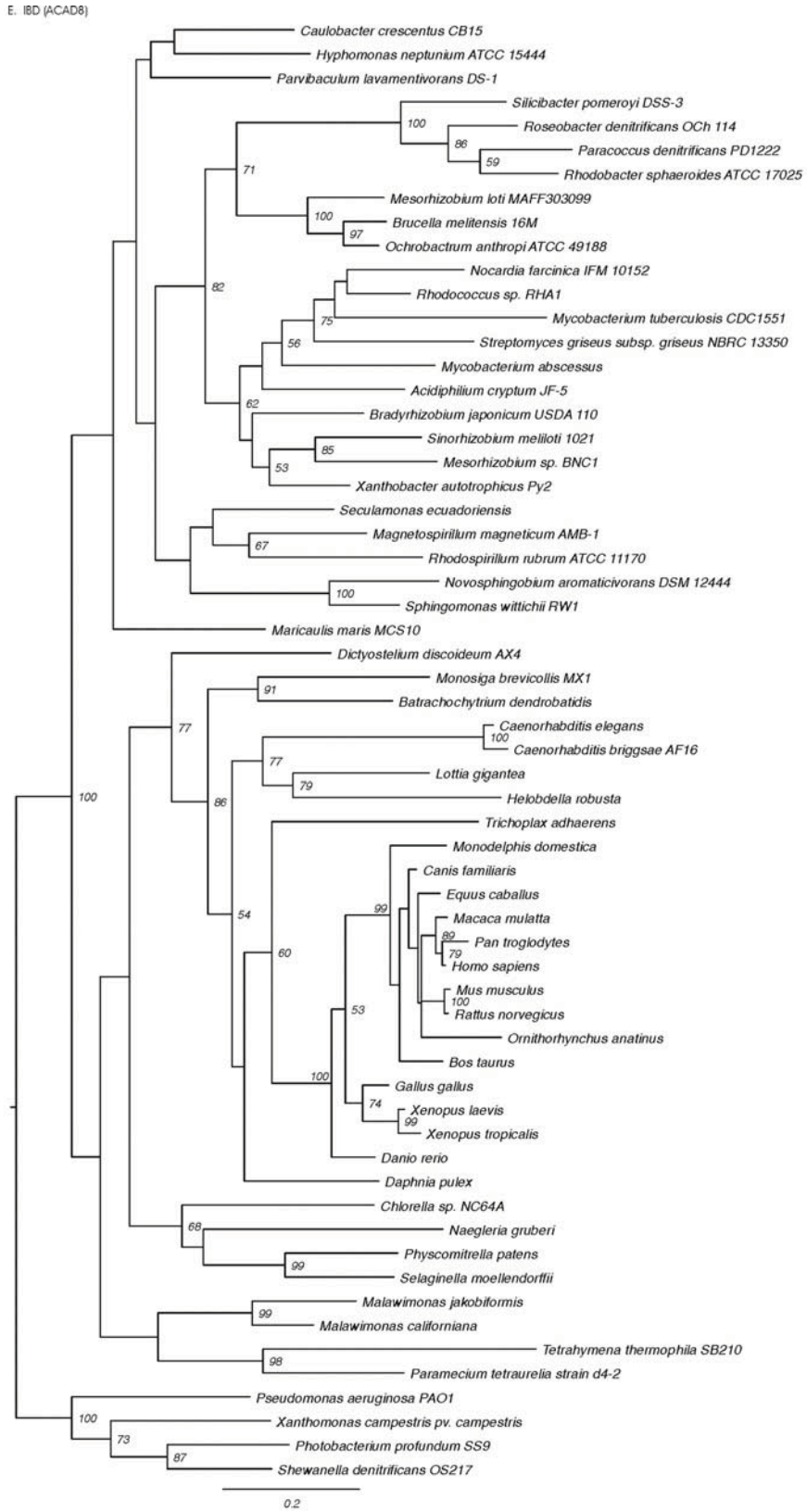


C. GCDH

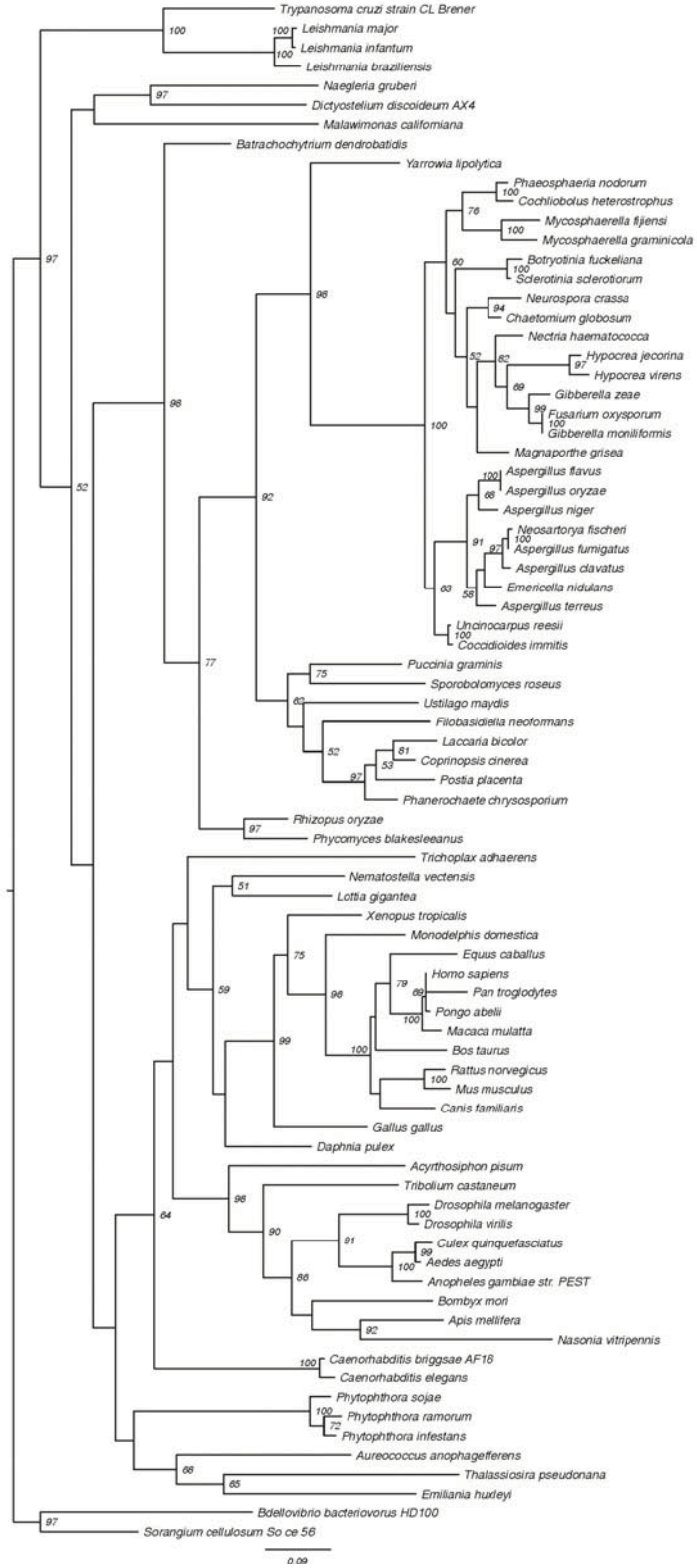

Monosiga brevicollis MX1
Rhizopus oryzae
Phycomyces blakesleeanus
Batrachochytrium dendrobatidis
Yarrowia lipolytica
Botryotinia fuckeliana
Sclerotinia sclerotiorum
Cochliobolus heterostrophus
Phaeosphaeria nodorum
Mycosphaerella graminicola
Mycosphaerella fijiensi
Neurospora crassa
Gibberella moniliformis
Fusarium oxysporum
Gibberella zeae
Nectria haematococca
Hypocrea jecorina
Hypocrea virens
Magnaporthe grisea
Chaetomium globosum
Ajellomyces capsulatum
Uncinocarpus reesii
Coccidioides immitis
Neosartorya fischeri
Aspergillus fumigatus
Aspergillus clavatus
Aspergillus oryzae
Aspergillus flavus
Aspergillus niger
Aspergillus terreus
Emericella nidulans
Ustilago maydis
Filobasidiella neoformans
Coprinopsis cinerea
Laccaria bicolor
Postia placenta
Phanerochaete chrysosporium
Puccinia graminis
Sporobolomyces roseus
Acyrthosiphon pisum
Tribolium castaneum
Aedes aegypti
Anopheles gambiae str. PEST
Culex quinquefasciatus
Drosophila virilis
Drosophila melanogaster
Nasonia vitripennis
Apis mellifera
Danio rerio
Xenopus laevis
Xenopus tropicalis
Equus caballus
Ornithorhynchus anatinus
Monodelphis domestica
Mus musculus
Rattus norvegicus
Bos taurus
Pongo abelii
Pan troglodytes
Homo sapiens
Macaca mulatta
Canis familiaris
Ciona intestinalis
Helobdella robusta
Daphnia pulex
Trichoplax adhaerens
Strongylocentrotus purpuratus
Lottia gigantea
Dictyostelium discoideum AX4
Thalassiosira pseudonana
Phytophthora ramorum
Phytophthora infestans
Aureococcus anophagefferens
Nematostella vectensis
Naegleria gruberi
Caenorhabditis elegans
Caenorhabditis briggsae AF16
Paramecium tetraurelia strain d4-2
Tetrahymena thermophila SB210
Novosphingobium aromaticivorans DSM 12444
Sphingomonas wittichii RW1
Mesorhizobium sp. BNC1
Paracoccus denitrificans PD1222
Mesorhizobium loti MAFF303099
Brucella melitensis 16M
Ochrobactrum anthropi ATCC 49188
Agrobacterium tumefaciens str. C58
Bradyrhizobium japonicum USDA 110
Sinorhizobium meliloti 1021
Rhodobacter sphaeroides ATCC 17025
Silicibacter pomeroyi DSS-3
Maricaulis maris MCS10
Roseobacter denitrificans OCh 114
Shewanella denitrificans OS217
Rhodospirillum rubrum ATCC 11170
Acidiphilium cryptum JF-5
Parvibaculum lavamentivorans DS-1
Magnetospirillum magneticum AMB-1
Xanthobacter autotrophicus Py2
Dechloromonas aromatica RCB
Rhizobium leguminosarum bv. viciae 3841
Ralstonia solanacearum GMI1000
Pseudomonas aeruginosa PAO1
Burkholderia mallei ATCC 23344
Trypanosoma cruzi strain CL Brener
Archaeoglobus fulgidus DSM 4304
Sorangium cellulosum So ce 56
Xanthomonas campestris pv. campestris
Thermobifida fusca YX
Mycobacterium tuberculosis CDC1551
Streptomyces griseus subsp. griseus NBRC 13350
Streptomyces avermitilis MA-4680
Corynebacterium efficiens YS-314
Arthrobacter aurescens TC1
Nocardia farcinica IFM 10152
Rhodococcus sp. RHA1
Mycobacterium abscessus
Haloarcula marismortui ATCC 43049
Halobacterium salinarum R1
Natronomonas pharaonis DSM 2160
Geobacter metallireducens GS-15
0.2

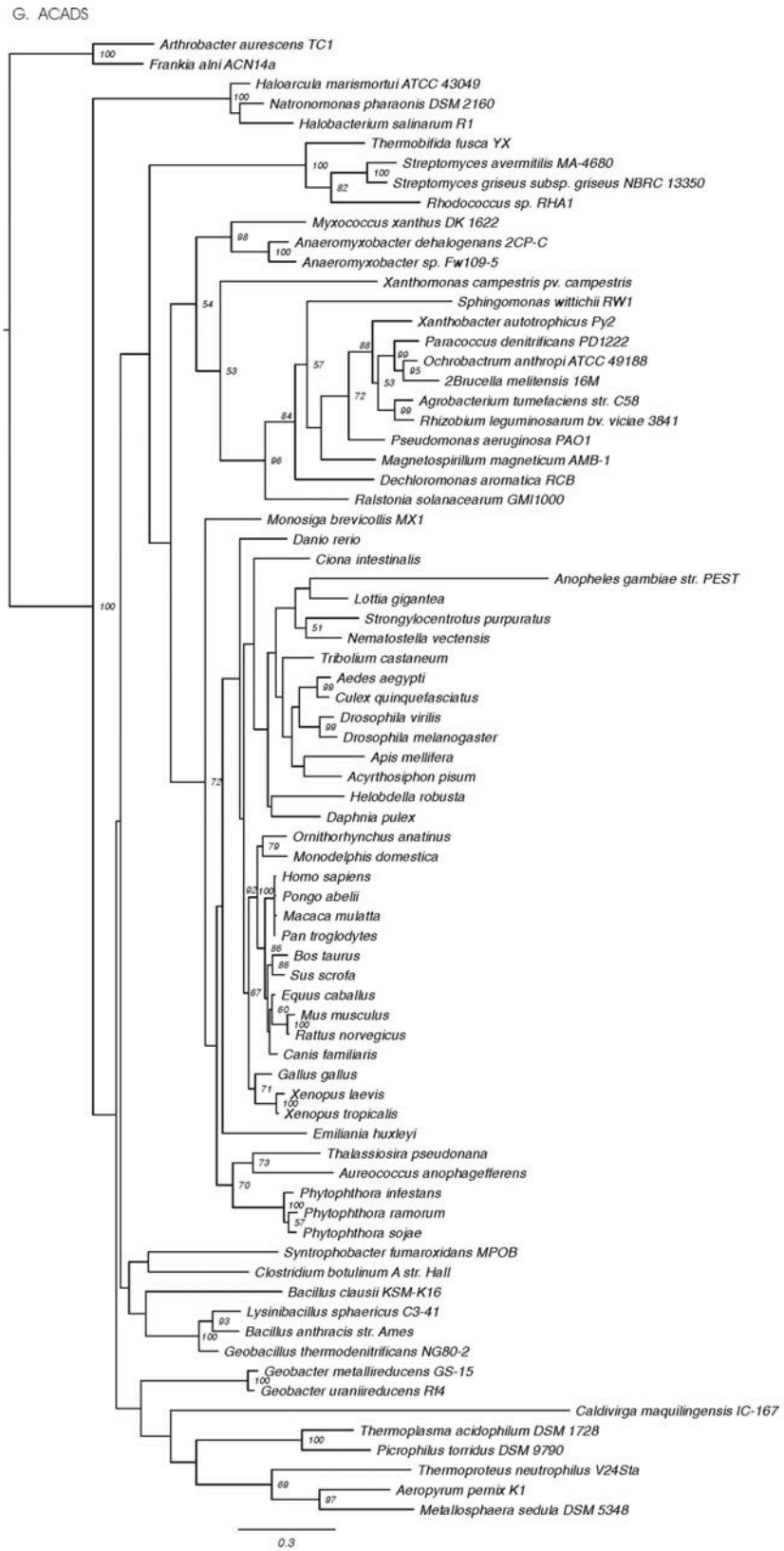
D. IVD

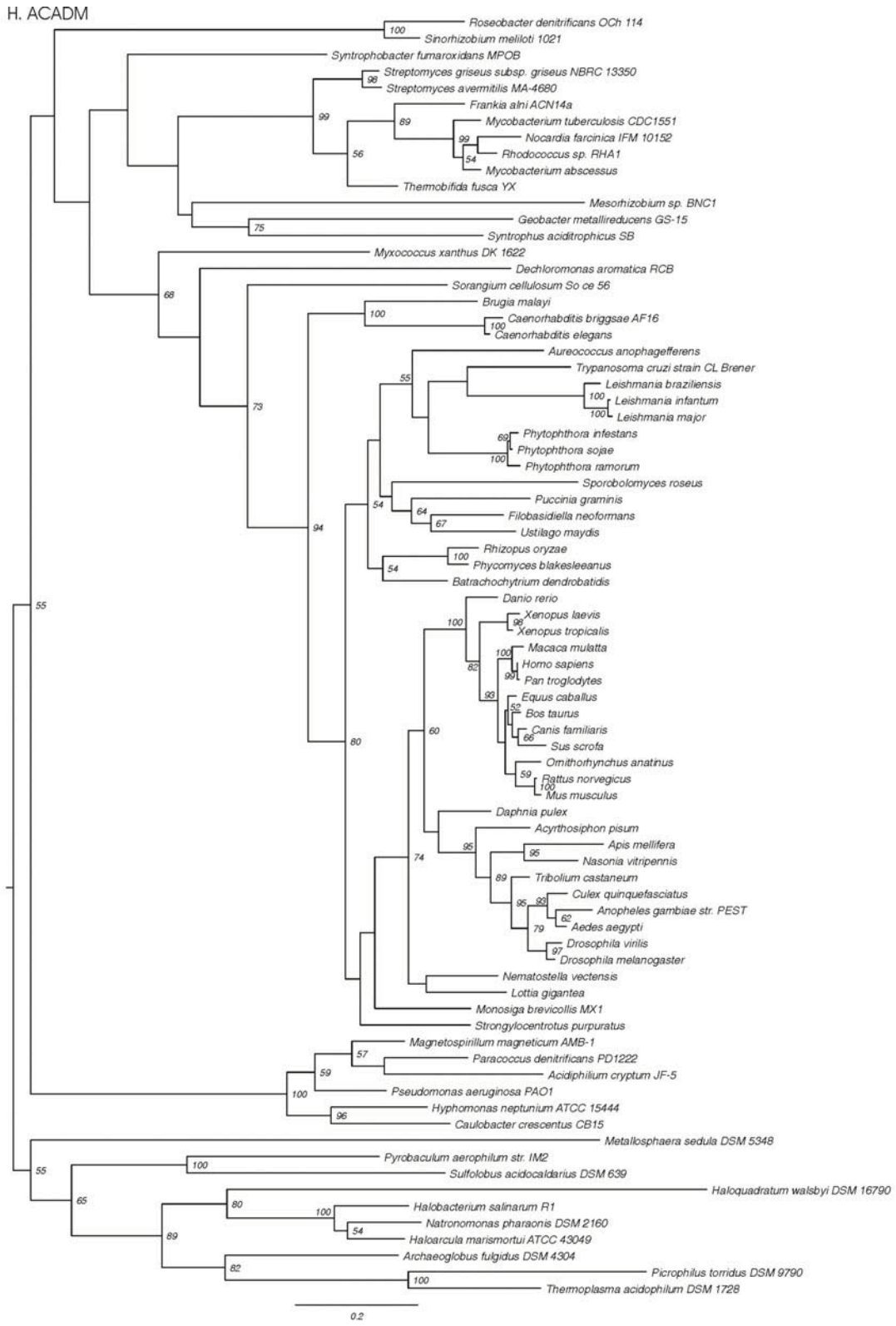

Jakoba libera
Reclinomonas americana
Seculamonas ecuadoriensis
Jakoba bahamensis
Naegleria gruberi
Cyanidioschyzon merolae
Dictyostelium discoideum AX4
Chlorella sp. NC64A
Oryza sativa
Sorghum bicolor
Populus trichocarpa
Arabidopsis thaliana
Selaginella moellendorffii
Physcomitrella patens
Volvox carteri
Chlamydomonas reinhardtii
Monosiga brevicollis MX1
Caenorhabditis briggsae AF16
Caenorhabditis elegans
Trichoplax adhaerens
Danio rerio
Xenopus tropicalis
Rattus norvegicus
Mus musculus
Equus caballus
Canis familiaris
Macaca mulatta
Homo sapiens
Pan troglodytes
Pongo abelii
Sus scrofa
Bos taurus
Ornithorhynchus anatinus
Monodelphis domestica
Gallus gallus
Strongylocentrotus purpuratus
Nematostella vectensis
Acyrthosiphon pisum
Apis mellifera
Nasonia vitripennis
Drosophila melanogaster
Drosophila virilis
Anopheles gambiae str. PEST
Aedes aegypti
Culex quinquefasciatus
Daphnia pulex
Lottia gigantea
Sporobolomyces roseus
Puccinia graminis
Ustilago maydis
Filobasidiella neoformans
Coprinopsis cinerea
Laccaria bicolor
Postia placenta
Phanerochaete chrysosporium
Batrachochytrium dendrobatidis
Rhizopus oryzae
Phycomyces blakesleeanus
Paramecium tetraurelia strain d4-2
Tetrahymena thermophila SB210
Yarrowia lipolytica
Aspergillus flavus
Aspergillus oryzae
Coccidioides immitis
Uncinocarpus reedii
Ajellomyces capsulatum
Aspergillus terreus
Emericella nidulans
Aspergillus clavatus
Neosartorya fischeri
Aspergillus fumigatus
Magnaporthe grisea
Neurospora crassa
Chaetomium globosum
Nectria haematococca
Fusarium oxysporum
Gibberella moniliformis
Gibberella zeae
Hypocrea virens
Hypocrea jecorina
Botryotinia fuckeliana
Phaeosphaeria nodorum
Cochliobolus heterostrophus
Mycosphaerella graminicola
Mycosphaerella fijiensi
Geobacter metallireducens GS-15
Parvibaculum lavamentivorans DS-1
Maricaulis maris MCS10
Hyphomonas neptunium ATCC 15444
Agrobacterium tumefaciens str. C58
Sphingomonas wittichii RW1
Novosphingobium aromaticivorans DSM 12444
Caulobacter crescentus CB15
Rhizobium leguminosarum bv. viciae 3841
Rhodospirillum rubrum ATCC 11170
Xanthobacter autotrophicus Py2
Acidiphilium cryptum JF-5
Magnetospirillum magneticum AMB-1
Shewanella denitrificans OS217
Photobacterium profundum SS9
Pseudomonas aeruginosa PAO1
Sinorhizobium meliloti 1021
Brucella melitensis 16M
Ochrobactrum anthropi ATCC 49188
Silicibacter pomeroyi DSS-3
Roseobacter denitrificans OCh 114
Rhodobacter sphaeroides ATCC 17025
Paracoccus denitrificans PD1222
Mesorhizobium loti MAFF303099
Mesorhizobium sp. BNC1
Xanthomonas campestris pv. campestris
Burkholderia mallei ATCC 23344
Ralstonia solanacearum GMI1000
Bradyrhizobium japonicum USDA 110
Dechloromonas aromatica RCB
Malawimonas californiana
Coxiella burnetii RSA 493
0.2


E. IBD (ACAD8)
Caulobacter crescentus CB15
Hyphomonas neptunium ATCC 15444
Parvibaculum lavamentivorans DS-1
Silicibacter pomeroyi DSS-3
Roseobacter denitrificans OCh 114
Paracoccus denitrificans PD1222
Rhodobacter sphaeroides ATCC 17025
Mesorhizobium loti MAFF303099
Brucella melitensis 16M
Ochrobactrum anthropi ATCC 49188
Nocardia farcinica IFM 10152
Rhodococcus sp. RHA1
Mycobacterium tuberculosis CDC1551
Streptomyces griseus subsp. griseus NBRC 13350
Mycobacterium abscessus
Acidiphilium cryptum JF-5
Bradyrhizobium japonicum USDA 110
Sinorhizobium meliloti 1021
Mesorhizobium sp. BNC1
Xanthobacter autotrophicus Py2
Seculamonas ecuadoriensis
Magnetospirillum magneticum AMB-1
Rhodospirillum rubrum ATCC 11170
Novosphingobium aromaticivorans DSM 12444
Sphingomonas wittichii RW1
Maricaulis maris MCS10
Dictyostelium discoideum AX4
Monosiga brevicollis MX1
Batrachochytrium dendrobatidis
Caenorhabditis elegans
Caenorhabditis briggsae AF16
Lottia gigantea
Helobdella robusta
Trichoplax adhaerens
Monodelphis domestica
Canis familiaris
Equus caballus
Macaca mulatta
Pan troglodytes
Homo sapiens
Mus musculus
Rattus norvegicus
Ornithorhynchus anatinus
Bos taurus
Gallus gallus
Xenopus laevis
Xenopus tropicalis
Danio rerio
Daphnia pulex
Chlorella sp. NC64A
Naegleria gruberi
Physcomitrella patens
Selaginella moellendorffii
Malawimonas jakobiformis
Malawimonas californiana
Tetrahymena thermophila SB210
Paramecium tetraurelia strain d4-2
Pseudomonas aeruginosa PAO1
Xanthomonas campestris pv. campestris
Photobacterium profundum SS9
Shewanella denitrificans OS217
0.2

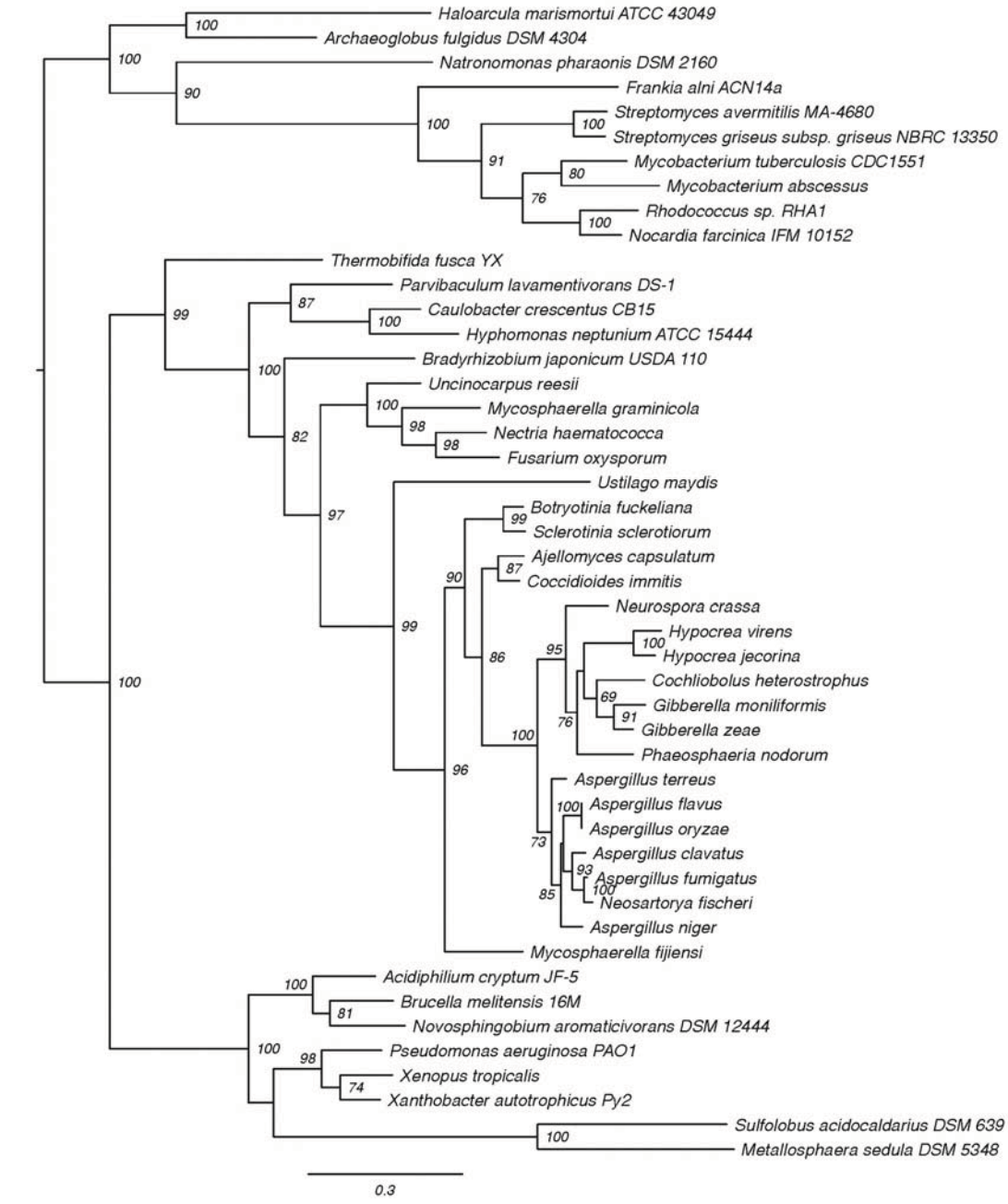
F. ACDSB

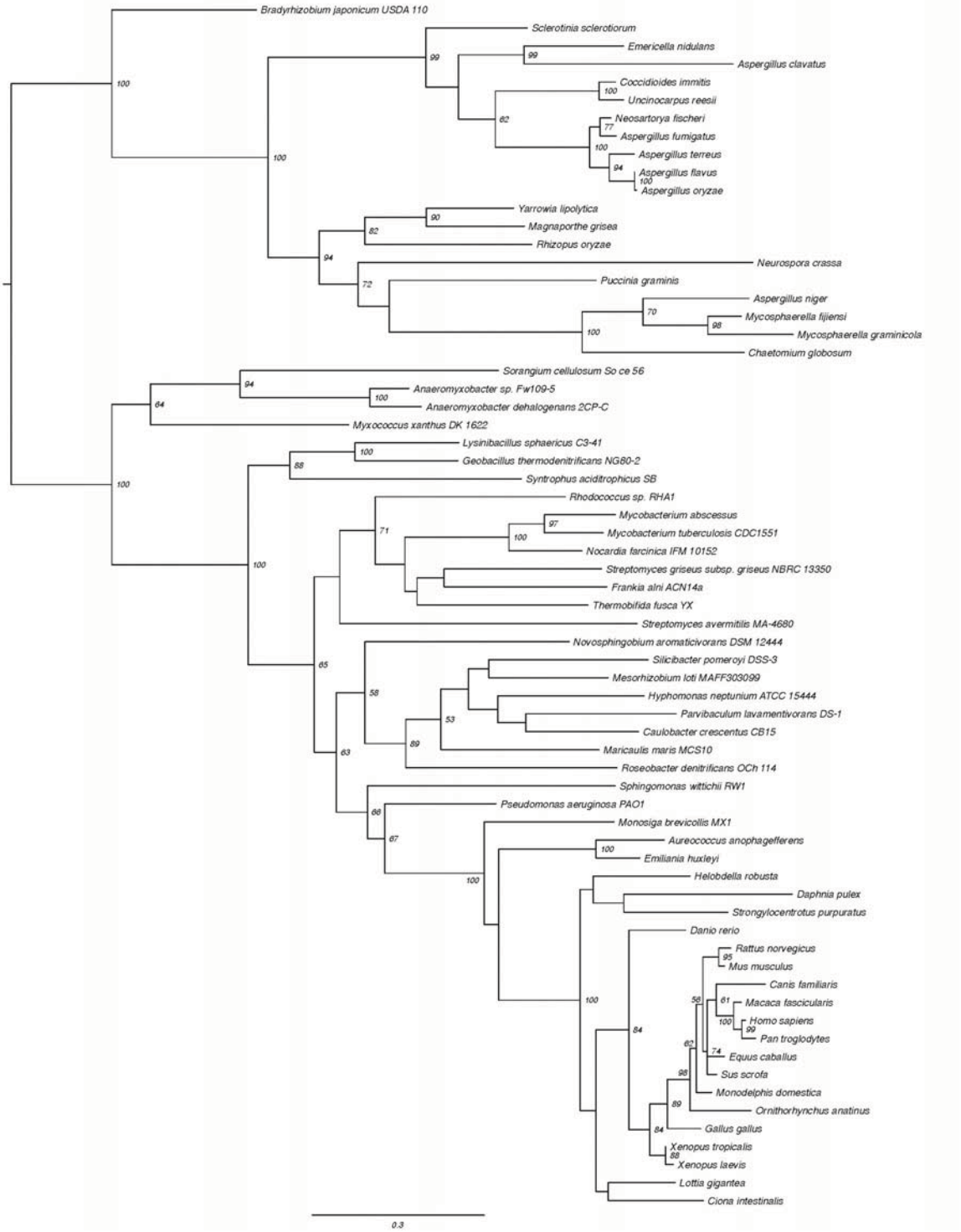

Trypanosoma cruzi strain CL Brener
Leishmania major
Leishmania infantum
Leishmania braziliensis
Naegleria gruberi
Dictyostelium discoideum AX4
Malawimonas californiana
Batrachochytrium dendrobatidis
Yarrowia lipolytica
Phaeosphaeria nodorum
Cochliobolus heterostrophus
Mycosphaerella fijiensi
Mycosphaerella graminicola
Botryotinia fuckeliana
Sclerotinia sclerotiorum
Neurospora crassa
Chaetomium globosum
Nectria haematococca
Hypocrea jecorina
Hypocrea virens
Gibberella zeae
Fusarium oxysporum
Gibberella moniliformis
Magnaporthe grisea
Aspergillus flavus
Aspergillus oryzae
Aspergillus niger
Neosartorya fischeri
Aspergillus fumigatus
Aspergillus clavatus
Emericella nidulans
Aspergillus terreus
Uncinocarpus reesii
Coccidioides immitis
Puccinia graminis
Sporobolomyces roseus
Ustilago maydis
Filobasidiella neoformans
Laccaria bicolor
Coprinopsis cinerea
Postia placenta
Phanerochaete chrysosporium
Rhizopus oryzae
Phycomyces blakesleeanus
Trichoplax adhaerens
Nematostella vectensis
Lottia gigantea
Xenopus tropicalis
Monodelphis domestica
Equus caballus
Homo sapiens
Pan troglodytes
Pongo abelii
Macaca mulatta
Bos taurus
Rattus norvegicus
Mus musculus
Canis familiaris
Gallus gallus
Daphnia pulex
Acyrthosiphon pisum
Tribolium castaneum
Drosophila melanogaster
Drosophila virilis
Culex quinquefasciatus
Aedes aegypti
Anopheles gambiae str. PEST
Bombyx mori
Apis mellifera
Nasonia vitripennis
Caenorhabditis briggsae AF16
Caenorhabditis elegans
Phytophthora sojae
Phytophthora ramorum
Phytophthora infestans
Aureococcus anophagefferens
Thalassiosira pseudonana
Emiliania huxleyi
Bdellovibrio bacteriovorus HD100
Sorangium cellulosum So ce 56

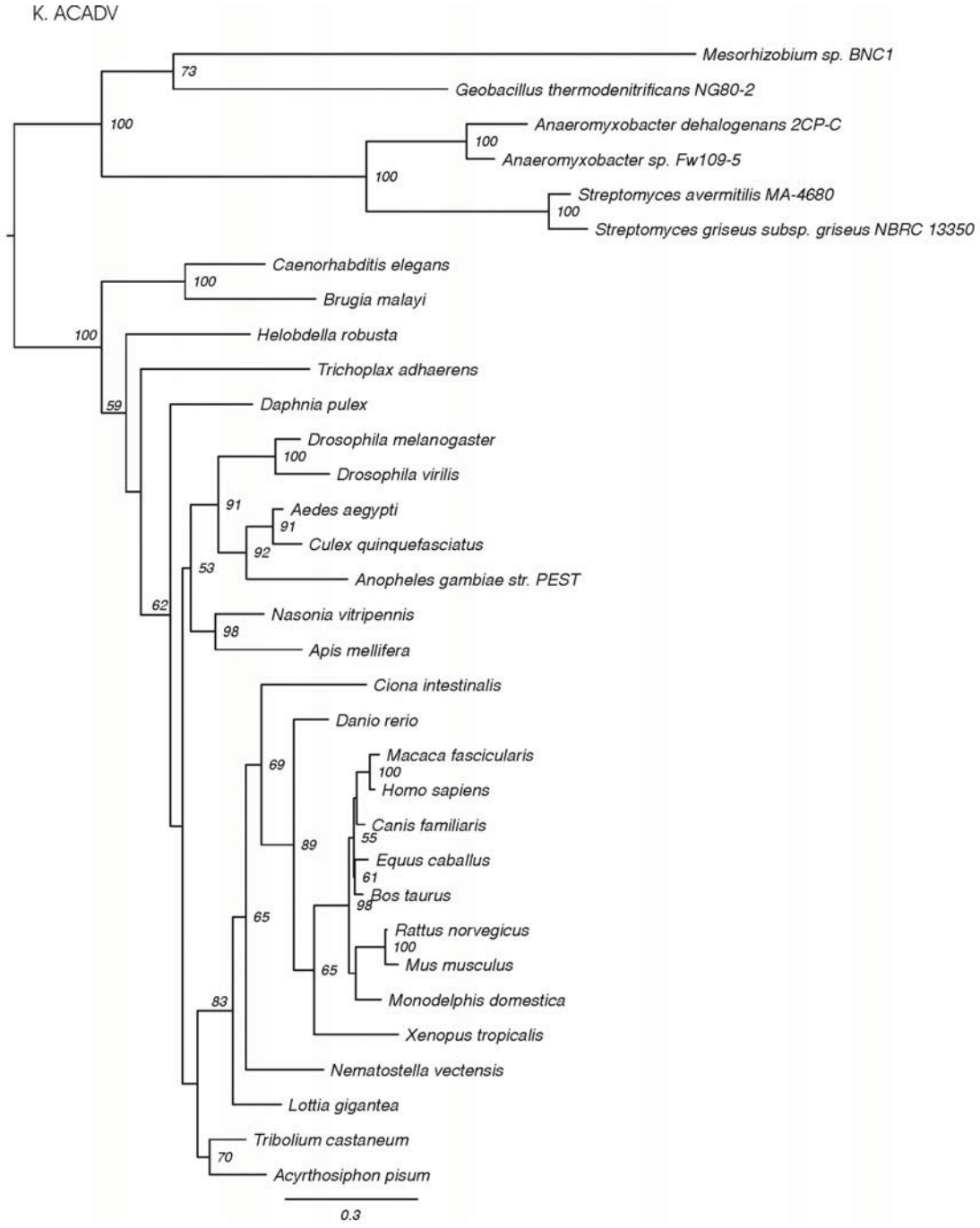

G. ACADS
Arthrobacter aurescens TC1
Frankia alni ACN14a
Haloarcula marismortui ATCC 43049
Natronomonas pharaonis DSM 2160
Halobacterium salinarum R1
Thermobifida fusca YX
Streptomyces avermitilis MA-4680
Streptomyces griseus subsp. griseus NBRC 13350
Rhodococcus sp. RHA1
Myxococcus xanthus DK 1622
Anaeromyxobacter dehalogenans 2CP-C
Anaeromyxobacter sp. Fw109-5
Xanthomonas campestris pv. campestris
Sphingomonas wittichii RW1
Xanthobacter autotrophicus Py2
Paracoccus denitrificans PD1222
Ochrobactrum anthropi ATCC 49188
2Brucella melitensis 16M
Agrobacterium tumefaciens str. C58
Rhizobium leguminosarum bv. viciae 3841
Pseudomonas aeruginosa PAO1
Magnetospirillum magneticum AMB-1
Dechloromonas aromatica RCB
Ralstonia solanacearum GMI1000
Monosiga brevicollis MX1
Danio rerio
Ciona intestinalis
Anopheles gambiae str. PEST
Lottia gigantea
Strongylocentrotus purpuratus
Nematostella vectensis
Tribolium castaneum
Aedes aegypti
Culex quinquefasciatus
Drosophila virilis
Drosophila melanogaster
Apis mellifera
Acyrthosiphon pisum
Helobdella robusta
Daphnia pulex
Ornithorhynchus anatinus
Monodelphis domestica
Homo sapiens
Pongo abelii
Macaca mulatta
Pan troglodytes
Bos taurus
Sus scrofa
Equus caballus
Mus musculus
Rattus norvegicus
Canis familiaris
Gallus gallus
Xenopus laevis
Xenopus tropicalis
Emiliania huxleyi
Thalassiosira pseudonana
Aureococcus anophagefferens
Phytophthora infestans
Phytophthora ramorum
Phytophthora sojae
Syntrophobacter fumaroxidans MPOB
Clostridium botulinum A str. Hall
Bacillus clausii KSM-K16
Lysinibacillus sphaericus C3-41
Bacillus anthracis str. Ames
Geobacillus thermodenitrificans NG80-2
Geobacter metallireducens GS-15
Geobacter uraniireducens Rf4
Caldivirga maquilingensis IC-167
Thermoplasma acidophilum DSM 1728
Picrophilus torridus DSM 9790
Thermoproteus neutrophilus V24Sta
Aeropyrum pernix K1
Metallosphaera sedula DSM 5348
0.3

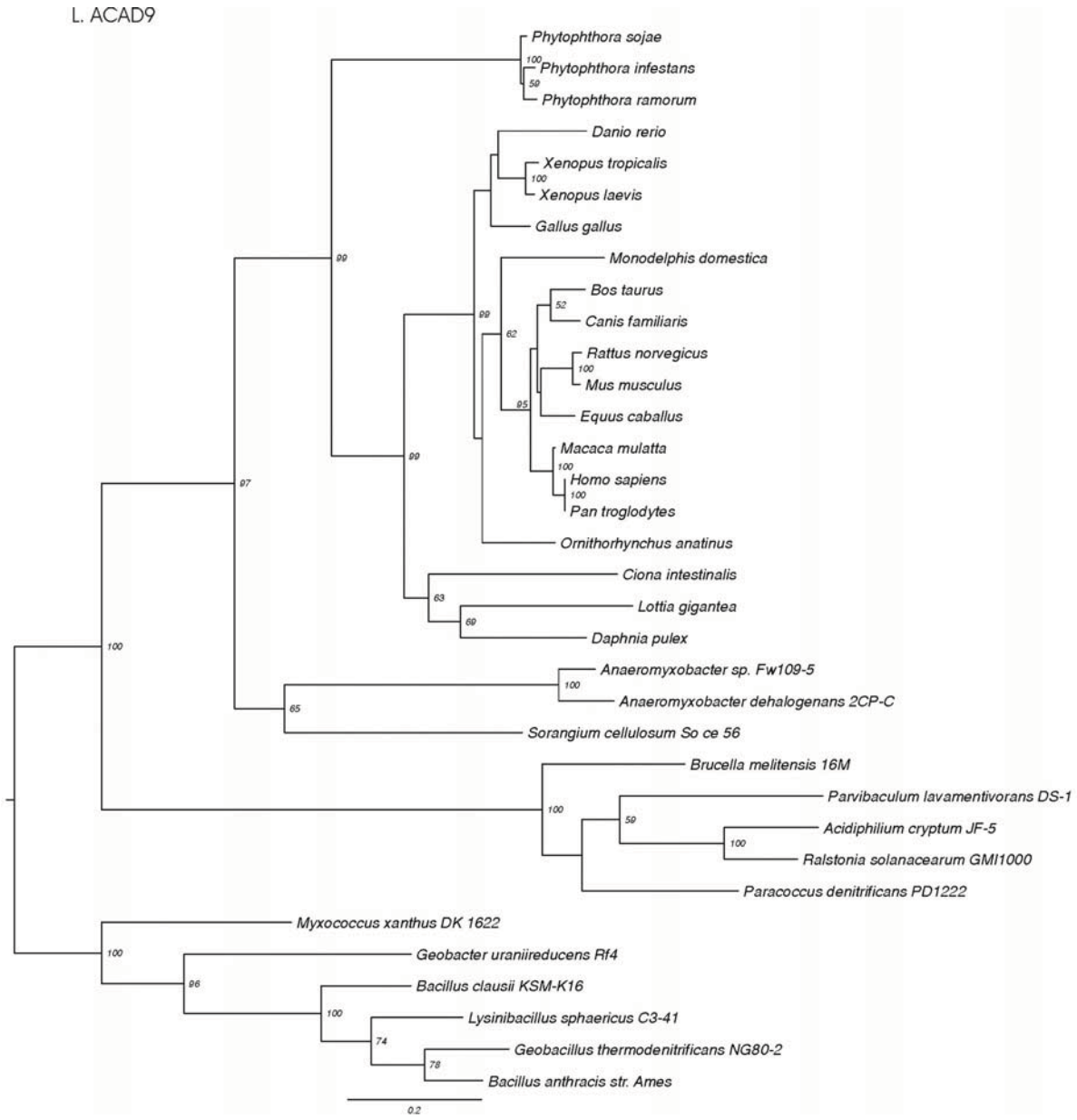

H. ACADM
Roseobacter denitrificans OCh 114
Sinorhizobium meliloti 1021
Syntrophobacter fumaroxidans MPOB
Streptomyces griseus subsp. griseus NBRC 13350
Streptomyces avermitilis MA-4680
Frankia alni ACN14a
Mycobacterium tuberculosis CDC1551
Nocardia farcinica IFM 10152
Rhodococcus sp. RHA1
Mycobacterium abscessus
Thermobifida fusca YX
Mesorhizobium sp. BNC1
Geobacter metallireducens GS-15
Syntrophus aciditrophicus SB
Myxococcus xanthus DK 1622
Dechloromonas aromatica RCB
Sorangium cellulosum So ce 56
Brugia malayi
Caenorhabditis briggsae AF16
Caenorhabditis elegans
Aureococcus anophagefferens
Trypanosoma cruzi strain CL Brener
Leishmania braziliensis
Leishmania infantum
Leishmania major
Phytophthora infestans
Phytophthora sojae
Phytophthora ramorum
Sporobolomyces roseus
Puccinia graminis
Filobasidiella neoformans
Ustilago maydis
Rhizopus oryzae
Phycomyces blakesleeanus
Batrachochytrium dendrobatidis
Danio rerio
Xenopus laevis
Xenopus tropicalis
Macaca mulatta
Homo sapiens
Pan troglodytes
Equus caballus
Bos taurus
Canis familiaris
Sus scrofa
Ornithorhynchus anatinus
Rattus norvegicus
Mus musculus
Daphnia pulex
Acyrthosiphon pisum
Apis mellifera
Nasonia vitripennis
Tribolium castaneum
Culex quinquefasciatus
Anopheles gambiae str. PEST
Aedes aegypti
Drosophila virilis
Drosophila melanogaster
Nematostella vectensis
Lottia gigantea
Monosiga brevicollis MX1
Strongylocentrotus purpuratus
Magnetospirillum magneticum AMB-1
Paracoccus denitrificans PD1222
Acidiphilium cryptum JF-5
Pseudomonas aeruginosa PAO1
Hyphomonas neptunium ATCC 15444
Caulobacter crescentus CB15
Metallosphaera sedula DSM 5348
Pyrobaculum aerophilum str. IM2
Sulfolobus acidocaldarius DSM 639
Haloquadratum walsbyi DSM 16790
Halobacterium salinarum R1
Natronomonas pharaonis DSM 2160
Haloarcula marismortui ATCC 43049
Archaeoglobus fulgidus DSM 4304
Picrophilus torridus DSM 9790
Thermoplasma acidophilum DSM 1728
0.2

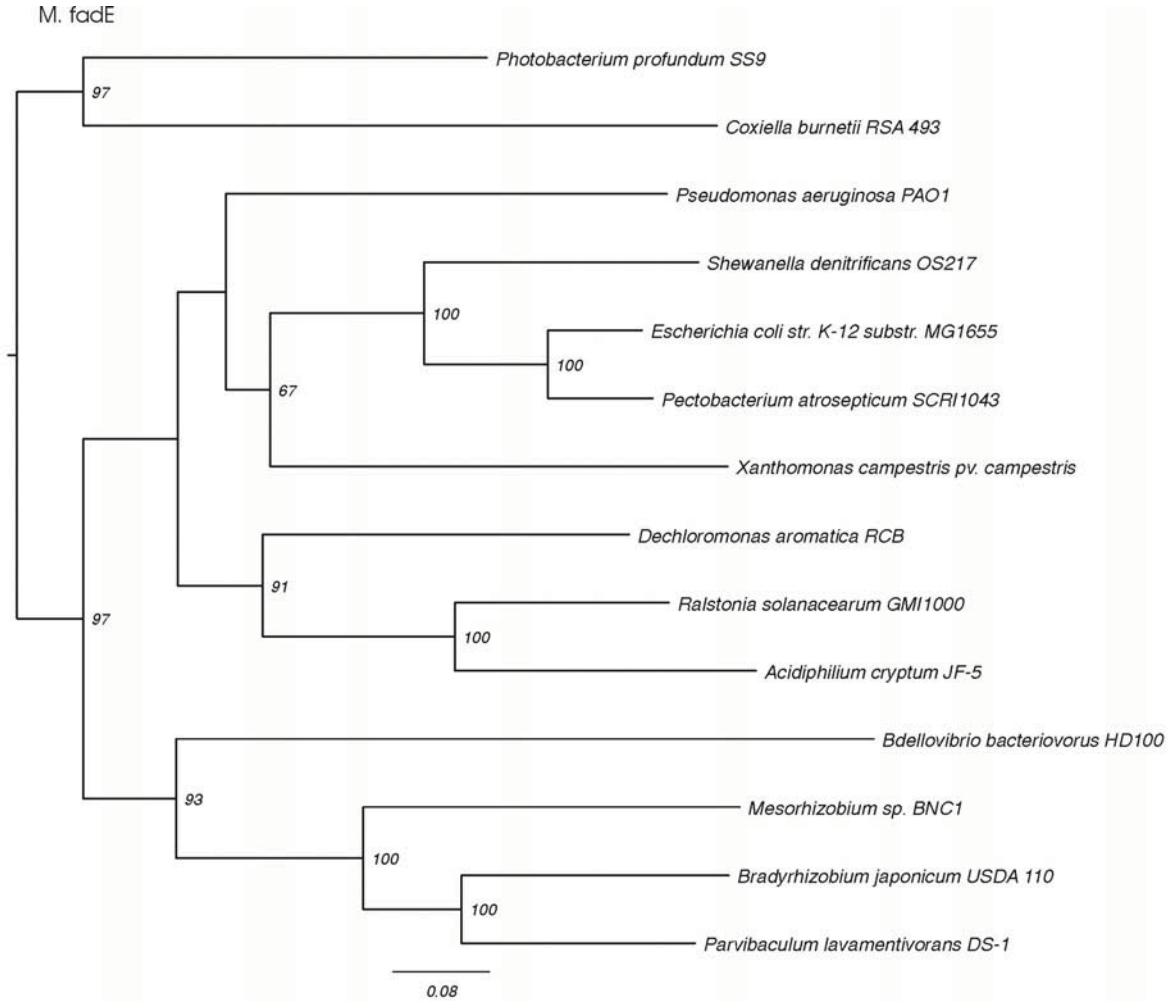
I. fadE12


Haloarcula marismortui ATCC 43049
Archaeoglobus fulgidus DSM 4304
Natronomonas pharaonis DSM 2160
Frankia alni ACN14a
Streptomyces avermitilis MA-4680
Streptomyces griseus subsp. griseus NBRC 13350
Mycobacterium tuberculosis CDC1551
Mycobacterium abscessus
Rhodococcus sp. RHA1
Nocardia farcinica IFM 10152
Thermobifida fusca YX
Parvibaculum lavamentivorans DS-1
Caulobacter crescentus CB15
Hyphomonas neptunium ATCC 15444
Bradyrhizobium japonicum USDA 110
Uncinocarpus reesii
Mycosphaerella graminicola
Nectria haematococca
Fusarium oxysporum
Ustilago maydis
Botryotinia fuckeliana
Sclerotinia sclerotiorum
Ajellomyces capsulatum
Coccidioides immitis
Neurospora crassa
Hypocrea virens
Hypocrea jecorina
Cochliobolus heterostrophus
Gibberella moniliformis
Gibberella zeae
Phaeosphaeria nodorum
Aspergillus terreus
Aspergillus flavus
Aspergillus oryzae
Aspergillus clavatus
Aspergillus fumigatus
Neosartorya fischeri
Aspergillus niger
Mycosphaerella fijiensi
Acidiphilium cryptum JF-5
Brucella melitensis 16M
Novosphingobium aromaticivorans DSM 12444
Pseudomonas aeruginosa PAO1
Xenopus tropicalis
Xanthobacter autotrophicus Py2
Sulfolobus acidocaldarius DSM 639
Metallosphaera sedula DSM 5348
0.3

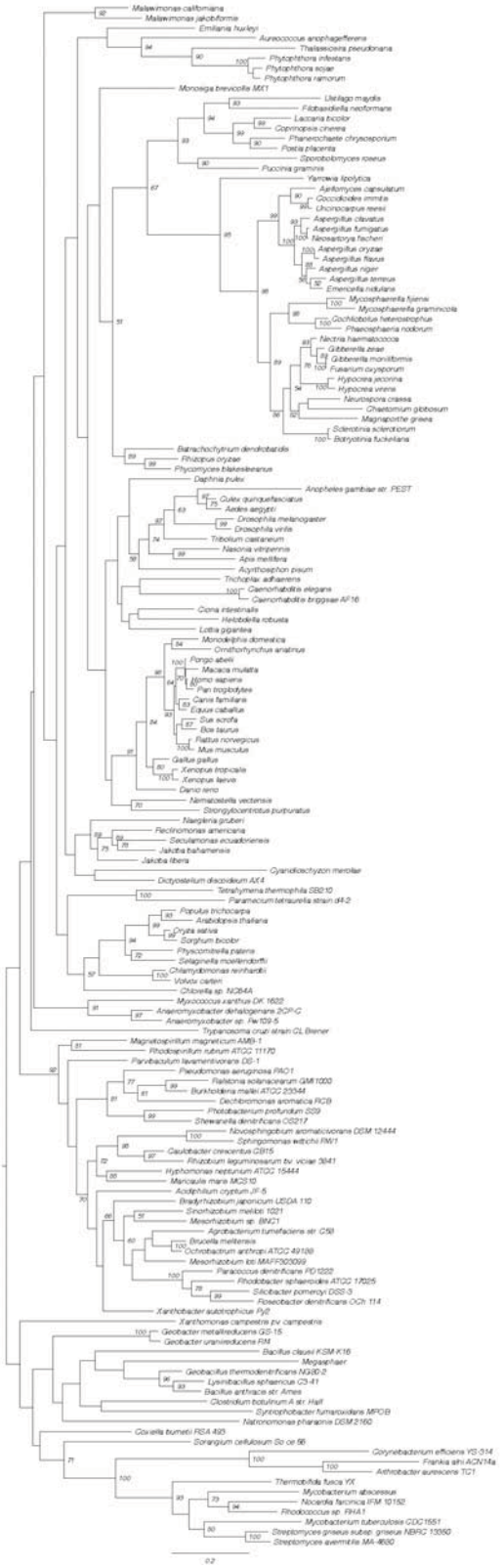

Bradyrhizobium japonicum USDA 110
Sclerotinia sclerotiorum
Emericella nidulans
Aspergillus clavatus
Coccidioides immitis
Uncinocarpus reesii
Neosartorya fischeri
Aspergillus fumigatus
Aspergillus terreus
Aspergillus flavus
Aspergillus oryzae
Yarrowia lipolytica
Magnaporthe grisea
Rhizopus oryzae
Neurospora crassa
Puccinia graminis
Aspergillus niger
Mycosphaerella fijiensi
Mycosphaerella graminicola
Chaetomium globosum
Sorangium cellulosum So ce 56
Anaeromyxobacter sp. Fw109-5
Anaeromyxobacter dehalogenans 2CP-C
Myxococcus xanthus DK 1622
Lysinibacillus sphaericus C3-41
Geobacillus thermodenitrificans NG80-2
Syntrophus aciditrophicus SB
Rhodococcus sp. RHA1
Mycobacterium abscessus
Mycobacterium tuberculosis CDC1551
Nocardia farcinica IFM 10152
Streptomyces griseus subsp. griseus NBRC 13350
Frankia alni ACN14a
Thermobifida fusca YX
Streptomyces avermitilis MA-4680
Novosphingobium aromaticivorans DSM 12444
Silicibacter pomeroyi DSS-3
Mesorhizobium loti MAFF303099
Hyphomonas neptunium ATCC 15444
Parvibaculum lavamentivorans DS-1
Caulobacter crescentus CB15
Maricaulis maris MCS10
Roseobacter denitrificans OCh 114
Sphingomonas wittichii RW1
Pseudomonas aeruginosa PAO1
Monosiga brevicollis MX1
Aureococcus anophagefferens
Emiliania huxleyi
Helobdella robusta
Daphnia pulex
Strongylocentrotus purpuratus
Danio rerio
Rattus norvegicus
Mus musculus
Canis familiaris
Macaca fascicularis
Homo sapiens
Pan troglodytes
Equus caballus
Sus scrofa
Monodelphis domestica
Ornithorhynchus anatinus
Gallus gallus
Xenopus tropicalis
Xenopus laevis
Lottia gigantea
Ciona intestinalis
0.3


K. ACADV
Mesorhizobium sp. BNC1
73
Geobacillus thermodenitrificans NG80-2
100
Anaeromyxobacter dehalogenans 2CP-C
100
Anaeromyxobacter sp. Fw109-5
100
Streptomyces avermitilis MA-4680
100
Streptomyces griseus subsp. griseus NBRC 13350
Caenorhabditis elegans
100
Brugia malayi
100
Helobdella robusta
Trichoplax adhaerens
59
Daphnia pulex
Drosophila melanogaster
100
Drosophila virilis
91
Aedes aegypti
91
Culex quinquefasciatus
92
53
Anopheles gambiae str. PEST
62
Nasonia vitripennis
98
Apis mellifera
Ciona intestinalis
Danio rerio
69
Macaca fascicularis
100
Homo sapiens
Canis familiaris
55
89
Equus caballus
61
Bos taurus
98
65
Rattus norvegicus
100
Mus musculus
65
Monodelphis domestica
83
Xenopus tropicalis
Nematostella vectensis
Lottia gigantea
Tribolium castaneum
70
Acyrthosiphon pisum
0.3


L. ACAD9
Phytophthora sojae
Phytophthora infestans
Phytophthora ramorum
Danio rerio
Xenopus tropicalis
Xenopus laevis
Gallus gallus
Monodelphis domestica
Bos taurus
Canis familiaris
Rattus norvegicus
Mus musculus
Equus caballus
Macaca mulatta
Homo sapiens
Pan troglodytes
Ornithorhynchus anatinus
Ciona intestinalis
Lottia gigantea
Daphnia pulex
Anaeromyxobacter sp. Fw109-5
Anaeromyxobacter dehalogenans 2CP-C
Sorangium cellulosum So ce 56
Brucella melitensis 16M
Parvibaculum lavamentivorans DS-1
Acidiphilium cryptum JF-5
Ralstonia solanacearum GMI1000
Paracoccus denitrificans PD1222
Myxococcus xanthus DK 1622
Geobacter uraniireducens Rf4
Bacillus clausii KSM-K16
Lysinibacillus sphaericus C3-41
Geobacillus thermodenitrificans NG80-2
Bacillus anthracis str. Ames
0.2


M. fadE
Photobacterium profundum SS9
97
Coxiella burnetii RSA 493
Pseudomonas aeruginosa PAO1
Shewanella denitrificans OS217
100
Escherichia coli str. K-12 substr. MG1655
100
67
Pectobacterium atrosepticum SCRI1043
Xanthomonas campestris pv. campestris
Dechloromonas aromatica RCB
91
Ralstonia solanacearum GMI1000
97
100
Acidiphilium cryptum JF-5
Bdellovibrio bacteriovorus HD100
93
Mesorhizobium sp. BNC1
100
Bradyrhizobium japonicum USDA 110
100
Parvibaculum lavamentivorans DS-1
0.08

N. ACADS+GCDH+IBD+IVD concatenated tree



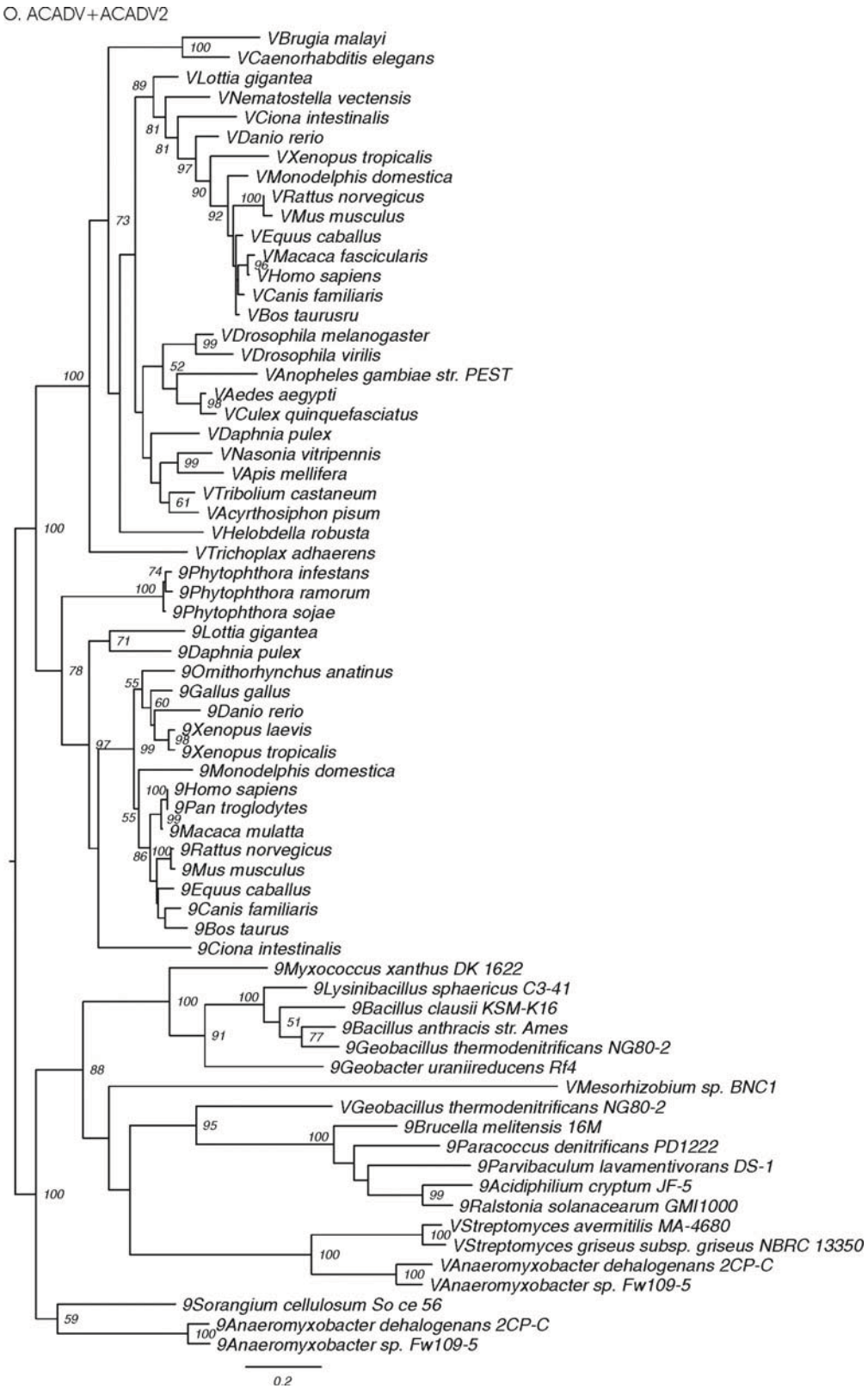

O. ACADV+ACADV2
VBrugia malayi
VCaenorhabditis elegans
VLottia gigantea
VNematostella vectensis
VCiona intestinalis
VDanio rerio
VXenopus tropicalis
VMonodelphis domestica
VRattus norvegicus
VMus musculus
VEquus caballus
VMacaca fascicularis
VHomo sapiens
VCanis familiaris
VBos taurusru
VDrosophila melanogaster
VDrosophila virilis
VAnopheles gambiae str. PEST
VAedes aegypti
VCulex quinquefasciatus
VDaphnia pulex
VNasonia vitripennis
VApis mellifera
VTribolium castaneum
VAcyrthosiphon pisum
VHelobdella robusta
VTrichoplax adhaerens
9Phytophthora infestans
9Phytophthora ramorum
9Phytophthora sojae
9Lottia gigantea
9Daphnia pulex
9Ornithorhynchus anatinus
9Gallus gallus
9Danio rerio
9Xenopus laevis
9Xenopus tropicalis
9Monodelphis domestica
9Homo sapiens
9Pan troglodytes
9Macaca mulatta
9Rattus norvegicus
9Mus musculus
9Equus caballus
9Canis familiaris
9Bos taurus
9Ciona intestinalis
9Myxococcus xanthus DK 1622
9Lysinibacillus sphaericus C3-41
9Bacillus clausii KSM-K16
9Bacillus anthracis str. Ames
9Geobacillus thermodenitrificans NG80-2
9Geobacter uraniireducens Rf4
VMesorhizobium sp. BNC1
VGeobacillus thermodenitrificans NG80-2
9Brucella melitensis 16M
9Paracoccus denitrificans PD1222
9Parvibaculum lavamentivorans DS-1
9Acidiphilium cryptum JF-5
9Ralstonia solanacearum GMI1000
VStreptomyces avermitilis MA-4680
VStreptomyces griseus subsp. griseus NBRC 13350
VAnaeromyxobacter dehalogenans 2CP-C
VAnaeromyxobacter sp. Fw109-5
9Sorangium cellulosum So ce 56
9Anaeromyxobacter dehalogenans 2CP-C
9Anaeromyxobacter sp. Fw109-5
0.2

Figure S1. Phylogenetic trees of individual and combined ACAD protein subfamilies. Numbers at branches are bootstraps values. Only values >50 are shown. **A:** tree built with ACD10, ACD11, and ACD11n sequences from animals, plant, and fungi, suggesting that ACD11 is an ancestral eukaryotic gene from which ACD10 has arisen in the animal lineage by gene duplication and subsequent addition of the hydrolase domain. Members of ACD11 are labeled with E ('Eleven'), and those of ACD10 with T ('Ten') preceding the species name, the rest are ACD11n. Genomes of a few species encode more than one ACD11n member. These homologs are distinguished by numbers preceding the corresponding species names. **B:** tree built with ACD10, ACD11, and ACD11n sequences from eukaryotes and bacteria. **C-M:** phylogenetic trees of individual subfamilies. C, GCDH; D, IVD; E, IBD; F, ACDSB; G, ACADS; H, ACADM; I, fadE12; J, ACADL; K, ACADV; L, ACADV2; M, fadE. **N:** phylogenetic tree built with concatenated sequences of GCDH, IVD, IBD, and ACADS, the four subfamilies considered more ancient. **O:** tree built with ACADV and ACADV2 sequences. Subfamily members are labeled with V (ACAD**V**) or 9 (ACADV2, synonym ACAD**9**) preceding the species name. Bacterial ACADV homologs are monophyletic with strong support, but their relationships to one another are incompatible with the species tree. This indicates a single gene transfer from animals to bacteria followed by further horizontal transfer events within bacteria.

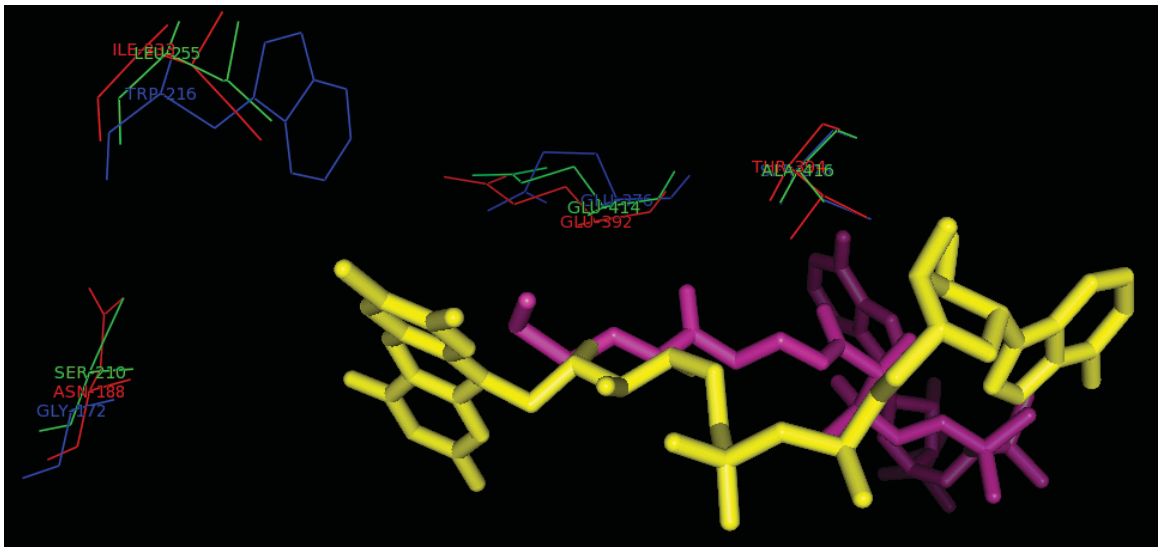


Figure S2. Three-dimensional structure alignment of the catalytic center of the human ACDSB, ACADS, and IBD proteins. The structures of ACDSB (PDB ID: 2jif), ACADS (PDB ID: 2vig) and IBD (PDB ID: 1rx0) from RCSB Protein Data Bank (http://www.rcsb.org) were superimposed by STRAP (http://www.charite.de/bioinf/strap/), and the alignment was displayed by PyMOL (http://www.pymol.org/). Key residues that determine the substrate specificity are as follows: in IBD: Gly172, Trp216, and Ser378 (blue); in ACADS: Asn188, Ile233, and Thr394 (red); and in ACDSB: Ser210, Leu255 and Ala416 (green). After removal of the transit peptide, the three latter residues correspond to Ser177, Leu222, and Ala383 in the mature protein. The catalytic reside Glu (the same in three sequences) is also shown. FAD, yellow; coenzyme A persulfide, purple. The alignment shows that the key residues of ACADS and ACDSB are more similar to one another, compared with those of IBD.

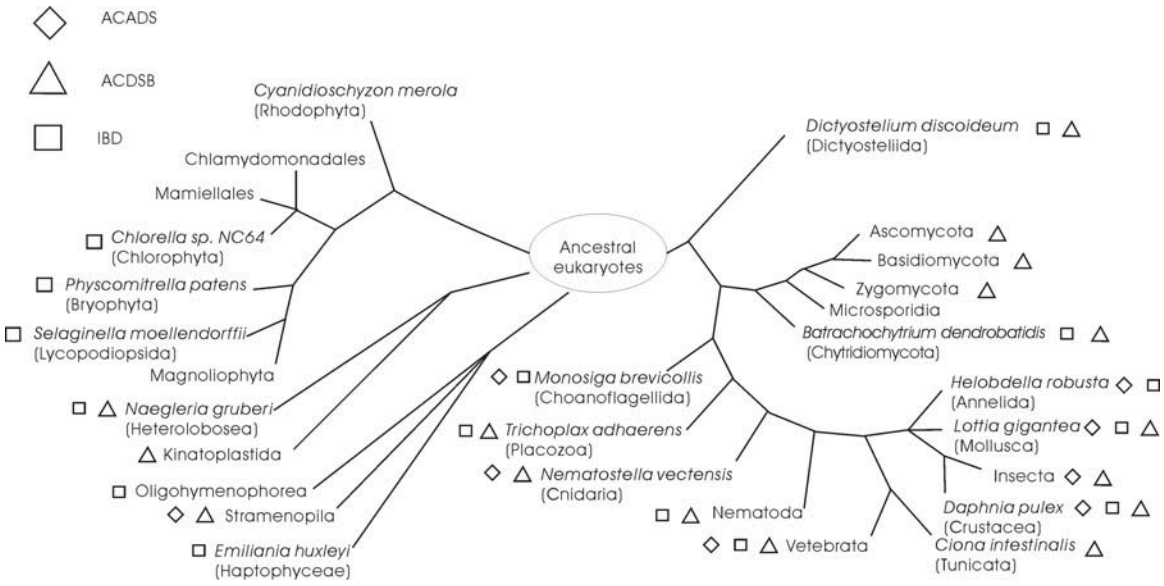


Figure S3. Distribution of ACADS, ACDSB, and IBD, mapped on a schematic eukaryotic species tree. When only one species from a given group appears in the tree, the species name is indicated with the group name in parenthesis; otherwise, the group name is indicated. Note that confidently alignable regions are too short for phylogenetic analysis of the three subfamilies, individually or combined. Our hypothesis on the evolutionary relationship of ACDSB, ACADS and IBD is based on several observations. At the level of overall sequence similarity, ACDSB, ACADS, and IBD are equally distant from one another (see Table 1 of main text), but at the level of substrate specificity-determinant residues, ACDSB and ACADS are more similar to each other than to IBD (Figure S2) and therefore may originate from a gene duplication event. ACADS has probably given rise to ACDSB, since taxonomically the former is much more broadly distributed than the latter (see Figure 4 of main text). Taken together, we speculate that at the outset, eukaryotes possessed only ACADS and IBD. Subsequent gene duplication of ACADS and paralog

divergence gave rise to the ACDSB subfamily in an ancestral eukaryote. Certain animal lineages such as crustaceans, mollusks, and vertebrates have held on to all three subfamilies for efficient degradation of a large range of fatty acids. In fungi and probably kinetoplastids as well, ACDSB acquired the catalytic activities of ACADS and IBD, and functional redundancy led to the loss of the two latter genes. Most variable subfamily combinations across taxa suggest multiple, independent events of functional generalization and gene loss.

Chapter 4 *In silico* identification of mitochondrion-targeted proteins using EST data

Current localization prediction methods are designed for full-length proteins, which are usually inferred from the complete gene sequence. These methods have only limited application to ESTs (sequences partially covering coding region of genes) which form a rich source to identify mitochondrial proteins in species whose genome sequence is not available. To solve this problem, we devised a tool TESTLoc, which is tailored to predict the subcellular localization from EST-derived peptides.

In silico identification of mitochondrion-targeted proteins using EST data

Yao-Qing Shen* and Gertraud Burger

Robert Cedergren Center for Bioinformatics and Genomics

Biochemistry Department, Université de Montréal

2900 Edouard-Montpetit

Montreal, QC, H3T 1J4, Canada

Corresponding author

email: yaoqing.shen@umontreal.ca

Abstract

The majority of mitochondrial proteins are encoded by the nuclear genome, and imported into mitochondria during or after translation. However, little is known about the makeup of the mitochondrial proteome across eukaryotes. In particular, knowledge about the mitochondrial proteome from primitive eukaryotes is paramount for understanding the evolutionary transition from an endosymbiont to the mitochondrial organelle. Expressed Sequence Tag (EST) data constitute the largest and taxonomically most comprehensive body of sequence information on eukaryotes. For example, the newly generated ~50,000 ESTs from jakobid flagellates, one of the most early diverging eukaryotes, provide a rich source for *in silico* inference of mitochondrial proteins (mit-proteins). Yet, bioinformatics tools do not perform well in predicting a protein's subcellular localization based on ESTs. Therefore, we developed a new predictor, TESTLoc, specifically for this purpose. TESTLoc predicts subcellular localization using partial protein sequences (EST-peptides), conceptually translated from ESTs. We encoded the ESTs peptides with different features such as amino acid composition, and used Support Vector Machine (SVM) as computational method. The correct reading frame is predicted using an existing algorithm, Prot4EST, which requires genome data for training. TESTLoc trained with data from plants (TESTLoc-plants) identifies mit-proteins from *Arabidopsis* (deduced from ESTs) at high sensitivity (93.5%) and positive predictive value (68%). We applied TESTLoc-plants to jakobids, which predicted known mit-proteins with high sensitivity (93%), but low positive predictive value (25%), likely due to the large phylogenetic distance between plants and

jakobids. In conclusion, our approach is well suited to predict the mitochondrial protein based on EST data, but predictor performance depends critically on the availability of training data from closely related taxa.

Key words: subcellular localization prediction, mitochondria, expressed sequence tags, support vector machine

Introduction

Mitochondria play an important role in the eukaryotic cell. They are involved in key processes such as energy production, metabolism, and regulation of cell death (reviewed in (Lang, Gray et al. 1999; Burger, Gray et al. 2003)). In addition, mitochondria bear clues about eukaryotic evolution. It is now widely accepted that mitochondria originated from an endosymbiotic α-proteobacterium that gradually transformed into an organelle as we know it today. During this evolutionary transition, some α-proteobacterial genes migrated to the host genome, while others were lost for good. Those genes that took up residence in the nucleus are expressed in the cytosol, and their gene products are imported back into mitochondria. Indeed, phylogenetic studies have identified a number of nucleus-encoded genes of clear α-bacterial origin, which are most likely contributed by the mitochondrial ancestor (Brown 2003; Doolittle, Boucher et al. 2003).

The exact makeup of the mitochondrial proteome is still unclear. Depending on the species, between three to ~70 proteins are contributed by the mitochondrial genome (Lang, Gray et al. 1999), while the majority of mitochondrial proteins, probably >1,000 (Meisinger, Sickmann et al. 2008), are encoded by the nucleus and imported into the organelle. Current experimental studies on the mitochondrial proteome focus on a few species, such as human, *Arabidopsis*, rice, yeast and *Tetrahymena thermophila* (Heazlewood, Howell et al. 2003; Sickmann, Reinders et al. 2003; Taylor, Fahy et al. 2003; Heazlewood, Tonti-Filippini et al. 2004; Forner, Foster et al. 2006; Johnson, Harris et al. 2007; Reinders and Sickmann 2007; Smith, Gawryluk et al. 2007; Li, Cai et al. 2009),

which are all phylogenetically derived. As a consequence, knowledge learnt from these species will not likely give insight into the origin of the mitochondrial proteome and how it changed in evolutionary time.

Jakobids are a group of unicellular heterotrophic flagellates. Both their mitochondrial gene complement and protein-based phylogenetic studies suggest that mitochondria of jakobids are the most primitive ones known, i.e., most closely related to α-proteobacteria (Lang, Gray et al. 1999; Rodriguez-Ezpeleta, Brinkmann et al. 2007). Therefore the jakobid mitochondrial proteome may reveal an intermediate stage of evolution, providing insight about gene losses and migrations, as well as protein recruitments to mitochondria, events that took place in the course of evolutionary time.

For a large number of species, including jakobids, experimental identification of the mitochondrial proteome still remains too expensive or unfeasible. This has set the stage for bioinformatics approaches to predict mit-proteins *in silico*. Currently around 47,000 jakobid Expressed Sequence Tags (ESTs) have been generated by the Protist EST Project (PEP), a trans-Canadian collaboration (http://megasun.bch.umontreal.ca/pepdb/pep.html). This large body of data provides a unique window into the jakobid mitochondrial proteome.

Today, over 20 computational tools are available to predict the subcellular localization of proteins (Supplementary Table 1). They predict the localization of a given sequence based on either annotation or the sequence itself. Annotation information from query sequences or their homologs, including textual description in SWISSPROT database, Gene Ontology database, or pubmed literatures, has been exploited by several predictors

(Nair and Rost 2002; Lu, Szafron et al. 2004; Shatkay, Hoglund et al. 2007). Co-occurrence of functional motifs or structural domains in proteins has also been used for localization predictions (Scott, Thomas et al. 2004; Guda and Subramaniam 2005). Sequence-based tools recognize specific targeting signals that guide proteins to different cellular compartments (Claros and Vincens 1996; Emanuelsson, Nielsen et al. 2000; Bannai, Tamada et al. 2002; Small, Peeters et al. 2004; Boden and Hawkins 2005), or classify proteins according to the frequency of each amino acid (Reinhardt and Hubbard 1998; Hua and Sun 2001), dipeptide, and gapped amino acid pair composition (Chou 2001; Chou and Cai 2002; Park and Kanehisa 2003; Huang and Li 2004), or physicochemical properties of amino acids (Sarda, Chua et al. 2005). More recently, a number of predictors combine different protein features (Bhasin and Raghava 2004; Guda, Fahy et al. 2004; Nair and Rost 2005; Xie, Li et al. 2005), or incorporate annotation with sequences-based prediction (Blum, Briesemeister et al. 2009). Meta predictors that integrate the prediction from heterogeneous tools have also been developed (Liu, Kang et al. 2007; Shen and Burger 2007; Assfalg, Gong et al. 2009).

A recent review evaluated the performance of available localization predictors using a dataset containing only sequences not included in, or similar to, those in the training set of a particular evaluated predictor (Casadio, Martelli et al. 2008). The study suggested five best performing tools: BaCelLo (Pierleoni, Martelli et al. 2006), LOCtree (Nair and Rost 2005), Protein Prowler (Boden and Hawkins 2005), TargetP (Emanuelsson, Nielsen et al. 2000), and Wolf PSORT (Horton, Park et al. 2007) (for sequence features and

computational methods they use, see Supplementary Table 1). These tools recognize mitochondrial proteins with sensitivities ranging from 47%-90% for animal, 35%-80% for fungi, and 29%-71% for plant sequences.

Available localization prediction programs are all built from full-length proteins. When these tools are tested on protein sequences conceptually translated from ESTs, the accuracy of recognizing mitochondrial proteins is very low (<30%, Figure 1). ESTs often represent only partial proteins (referred to as EST-peptides from here on) and typically lack the N-terminal region. The poor performance of traditional tools is likely due to their dependence on protein features (such as targeting signal and sequence motifs) which EST-peptides lack. In addition, being partial sequences, their amino acid composition may differ from that of the full-length proteins. Therefore, we developed a method that is optimized for predicting mitochondrial proteins using EST data. Our approach is readily applicable to build predictors for all subcellular compartments.

Materials and Methods

1. Data sets

1.1 *Arabidopsis* ESTs data set

We collected from SWISSPROT full-length *Arabidopsis* proteins localized in eight known subcellular compartments: cytosol (cyt), endoplasmic reticulum (end), extracellular space (ext), mitochondria (mit), nucleus (nuc), peroxisomes (per), plasma membrane (pla), and vacuole (vac). These sequences were selected by the following criteria: 1) they are encoded

by the nucleus; 2) their subcellular localization is experimentally verified; and 3) the localization annotation is not ambiguous (i.e. terms like “probable” or “possible” are absent from their annotation of subcellular localization). Since we are more interested in mitochondrial proteins, we enriched the mitochondrial data by adding sequences from *Arabidopsis* Mitochondrial Protein DataBase (AMPDB) (Heazlewood and Millar 2005). Furthermore, the mitochondrial data were divided into two classes: mitochondrial globular proteins, and mitochondrial membrane proteins. The ESTs corresponding to these proteins were found via a similarity search by BLASTX in dbEST of GenBank, as illustrated in Figure 2. The selected ESTs were clustered by Phrap (default parameters, http://www.phrap.org), and translated into proteins in the frame indicated by BLASTX alignment. Sequence redundancy was reduced so that no pair of sequences shares more than 60% similarity. We obtained a dataset of 289 EST-peptides. Table 1 compiles all datasets generated in the context of this work.

1.2 Plant ESTs data set

Using the same procedure for *Arabidopsis* EST selection and translation, EST-peptides corresponding to all plant proteins from SWISSPROT with known localization were combined to create a dataset of 1108 sequences in total (Table 1).

1.3 Expanded ESTs data

The collected plant full-length protein sequences were processed according to the following rules, as illustrated in Figure 3:

If the sequence is shorter than 200 aa, it remains unchanged.

If the length of the sequence is between 200 to 400 aa, fragments at length ranging from 140 to 260 aa will be taken from both the N-terminus and C-terminus. The range was based on a survey of the length distribution of ESTs, with means being ~600 nt and standard deviation being ~180 nt. The N-terminal fragment will start within 80 aa from the starting Met, while the C-terminal fragment will contain the last amino acid. This is to simulate the nature of ESTs which usually contains the intact C-terminus, but misses the N-terminus.

If the sequence is longer than 400 aa, an additional middle fragment will be taken, which starts after the first 80 aa, but before the middle position of the original sequence.

The fragmented sequences were combined with plant EST-derived peptides to form the expanded EST data. The data was clustered using a threshold of 60%, after clustering, 80% of the remaining sequences share a sequence similarity lower than 50% (Figure 4).

1.4 Jakobid ESTs with "known" subcellular location

The 20,683 EST clusters from six early diverging jakobid protists from four different genera, *Jakoba bahamensis*, *J. libera*, *Malawimonas californiana*, *M. jakobiformis*, *Seculamonas ecuadoriensis*, and *Reclinomonas americana*, have already been functionally annotated (O'Brien, Koski et al. 2007). The annotation including GO terms of these ESTs

from TBestDB was extracted. From the GO annotations, we collected the jakobid sequences from nine subcellular compartments, as listed in Table 1. All annotations are inferred from sequence similarity with experimentally validated proteins, but none of the jakobids proteins has been validated experimentally.

2. Implementation of SVM

2.1 Attributes used to represent the sequence as input for SVM

Physicochemical properties. Physicochemical properties of the amino acids can be represented by amino acid indices (AAindex), developed by the Amino Acid Index Database (http://www.genome.jp/dbget-bin/show_man?aaindex). The database currently contains 494 features for each amino acid (alpha-helix, turn and beta-sheet propensity, hydrophobicity, bulkiness, etc.). For each feature, its value was calculated for the whole sequence, and normalized by the sequence length. All 494 features were calculated, and each EST-peptide was converted into a 494-dimension vector.

Amino acid composition. Six different types of amino acid compositions were calculated. They are the frequency of: individual amino acid (1^{st}-order), di-peptide (2^{nd}-order), tri-peptide (3^{rd}-order), tetra-peptide (4^{th}-order), penta-peptide (5^{th}-order), and hexa-peptide (6^{th}-order) in the input sequence. For the amino acid composition of T^{th}-order, the input sequence is represented by a vector of size 20^T.

Grouped amino acid composition. Amino acids are grouped according to their properties (Table 2). The alphabet of size 20 for amino acids are replaced by an alphabet of size eight (group C) and size ten (group D). Group C classifies amino acids according to their chemical properties, which has shown good performance for localization prediction of full-length proteins in CELLO (Yu, Chen et al. 2006). Group D classifies amino acids according to their structure. EST sequences were converted to the new alphabets, and the composition of the amino acid groups was used to encode the sequences. Compositions from 1^{st}-order to 8^{th}-order were calculated.

Gapped amino acid composition. This feature represents the frequency of two amino acids (or amino acid groups) separated by x amino acids (or grouped amino acids) between them, x being the gap length. We experimented with the gap length from 1 to 6. Depending on the alphabet used, each sequence was represented by a vector of size 400, 64, or 100.

2.2 Parameter selection and evaluation of Support Vector Machine (SVM) predictions

The SVM package LIBSVM was employed for this study (Fan, Chen et al. 2005), with the radial basis function (RBF) adopted as kernel function: $K(x_i, x_j) = \exp(-\gamma||x_i,-x_j||^2)$, $\gamma > 0$, which requires the selection of kernel parameter γ, and penalty parameter C. To select the optimal parameters for SVM training and to evaluate the predictions, we performed a 10-fold cross validation (10-fold CV), followed by a 10-fold independent evaluation (Figure 5). We first randomly divided the whole data set into ten groups of equal size. For each

iteration of the ten folds, nine groups were combined to build SVM models, and the remaining group was used for evaluation. The combined nine groups were again divided into ten parts, nine parts were combined and used to train SVM with given C and γ, while the remaining part was used to find the optimal combination of the two parameters. Then the SVM with selected C and γ was assessed with the evaluation data.

2.4 Performance evaluation

The overall accuracy for all classes, as well the sensitivity (SN), specificity (SP), positive predictive value (PPV), and Matthews correlation coefficient (MCC) of each class are calculated as follows:

$$\text{Overall Accuracy (acc)} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FN_i)} * 100$$ i: the i-th class; n: total number of classes

For each class i:

$$\text{Sensitivity } (SN_i) = \frac{TP_i}{TP_i + FN_i} * 100$$

$$\text{Specificity } (SP_i) = \frac{TN_i}{TN_i + FP_i} * 100$$

$$\text{Positive predictive value } (PPV_i) = \frac{TP_i}{TP_i + FP_i} * 100$$

Matthews correlation coefficient (MCCi) =

$$\frac{TP_i * TN_i - FP_i * FN_i}{\sqrt{(TP_i + FP_i) * (FP_i + FN_i) * (TN_i + FP_i) * (TN_i + FN_i)}}$$

3 Open reading frame (ORF) prediction for ESTs

Prot4EST (Barbe, Lundberg et al. 2008) was used for the prediction of open reading frame (ORF) in ESTs. The non-redundant protein sequence database from NCBI was used for BLASTX search, and ORFs were first inferred from the alignment. For ESTs without significant matches from BLAST, the tool ESTScan incorporated with Prot4EST was used to predict the ORF. ESTScan was trained with annotated *Arabidopsis* genomic and mRNA data collected from the European Molecular Biology Laboratory (EMBL) database, to generate the score matrix that represents the species-dependent hexanucleotide biases between coding and non-coding regions.

Results and Discussion

Selection of training set

To develop a prediction method for ESTs from early diverging eukaryotes such as jakobids, one of the challenges lies in the limited training and test data. We performed several pilot studies to form the best plant training set. Predictors perform well when trained with experimentally verified data and tested with data from the same species. But in jakobids, few mitochondrial proteins have been experimentally verified. From a few model species in which well characterized localization data are available, we need to choose the one that is

phylogenetically closest to jakobids. Both phylogenetic studies, and the mitochondrial genome suggest that plant mitochondria more closely resemble the alpha-proteobateria ancestor than animals and fungi, and should therefore serve as better training data than the latter two.

On the other hand, plant cells contain chloroplast proteins, while jakobids lack this organelle. Since both mitochondria and chloroplasts possess proteins included in replication, transcription, translation and energy production that are derived from bacterial ancestors, they are likely to share common sequence features. This likely leads to misclassification when the model is trained with sequences from an organism bearing chloroplasts and tested with those from an organism lacking this organelle. Therefore we removed the chloroplast as a class in the prediction.

A number of proteins with known subcellular location are absent from the collected plant EST data set, because they have no corresponding EST sequences in public databases (for example, 289 *Arabidopsis* EST-peptide selected from 1035 full-length proteins). To construct a training set with optimal coverage, the missing EST-peptides were substituted by artificial ones, generated by breaking up the corresponding full-length proteins into overlapping pieces of ~200 residues.

Using this expanded dataset, we have experimented with different sequence features, including amino acid composition, grouped amino acid composition reflecting the physicochemical properties, gapped amino acid composition capturing the spatial context,

the physicochemical properties of amino acids, as well as their combination. The prediction scheme with best performance was tested with jakobid data.

Performance of predictors based on individual features

Out of the 41 sequence features we investigated, the best prediction was obtained by the SVM based on the 4^{th}-order of amino acid composition. For mitochondrial proteins, it yielded a sensitivity of 91%, PPV of 68%, and MCC of 0.8. SVMs based on the 6^{th}-order of group C amino acid composition, and 7^{th}-order of group D amino acid composition had similar performance (Figure 6A).

The same trend was observed for all three kinds of composition: the higher the order of composition, the higher the sensitivity for larger classes, i.e., mitochondria (Figure 6B). But the trends of PPV (Figure 6C) and MCC are different. At first their values increase with the order, reach a peak, then drop when the order continues to increase. This suggests that higher order composition makes the scheme remember the instances in the training procedure, causing bias towards larger classes to achieve the highest overall accuracy. This phenomenon is reflected by increased sensitivity but decreased PPV of the largest class, a sign of overfitting.

Other sequence features such as the gapped amino acid composition and physicochemical properties represented by AAindex, did not give satisfying results: all of their MCCs are below 0.4 (Supplementary Table 2).

Integration of different features

Previous studies show that integration of multiple sequence features improves the performance of localization prediction (Shen and Burger 2007; Blum, Briesemeister et al. 2009). We integrated all the 41 sequence features described in the Methods section in two ways: integration of prediction results, and integration of attributes. The integration of prediction results was achieved by a two-layer SVM (Figure 7). The first layer consists of SVMs based on each sequence feature, which yielded as output the probability of the query sequence belonging to each class. The outputs of all first-layer SVM were combined, and served as input of the second-layer SVM. Therefore each sequence was converted to a vector of size 369 (nine predictions for each of the 41 features). Instead, the integration of attributes combined the vectors of each sequence feature, and the combined feature was used as input for SVM. Both integrations yielded much lower performance than the best individual predictor (Table 3). It seems that instead of combining the strength of individual features, their integration merely averaged their performance. The small improvement of integrated prediction suggests that the individual features tested in this study have no addictive or complementary effects.

When only the top three features (4^{th}-order amino acid composition, 6^{th}-order group-C composition and 7^{th}-order group-D composition) were combined, the prediction accuracy was slightly improved compared to the prediction based on the best individual feature. The integration of predictions made from the top three features shows similar performance as

integration of the three features themselves as attributes (Table 3), but is computationally much faster than the latter approach and is therefore more practical in application.

Implementation of prediction methods and validation with AT data

Another challenge in the localization prediction based on EST-derived peptides is the correct translation. Unlike genomic data, ESTs often lack start codon and 5'-UTR which otherwise help detecting the correct reading frame. In addition, ESTs are products of single-pass reads containing low quality regions with sequencing errors that increase the difficulty to find correct open reading frame (ORF). Several tools have been developed for ORF identification in ESTs (Iseli, Jongeneel et al. 1999; Hatzigeorgiou, Fiziev et al. 2001; Barbe, Lundberg et al. 2008). Among them, we chose the Prot4EST (Barbe, Lundberg et al. 2008) for EST translation, which combines similarity-based and machine-learning-based prediction of ORF. Prot4EST first searches ESTs against protein databases by BLAST. The protein-EST alignment indicates the correct translation frame. The translation is extended beyond the aligned region till the start or stop codon. For ESTs of which no similar proteins are found, ESTScan integrated within Prot4EST is used to predict the ORF based on Hidden Markov Model, which recognizes the species-specific bias in hexanucleotide composition associated with coding and non-coding regions (Iseli, Jongeneel et al. 1999), and generates a score matrix to represent such bias. To get the score matrix specific to

Arabidopsis, we trained ESTScan with *Arabidopsis* data, as described in the Methods section.

We built a tool named TESTLoc, which takes the EST translation from Prot4EST, and predicts the subcellular localization of these translated peptides. The sequence feature to use is an option that can be chosen by users. The performance of TESTLoc was evaluated with *Arabidopsis* ESTs, which correspond to *Arabidopsis* proteins of known localization. We used the predictor built with integration of top three attributes. The result showed satisfying prediction for most classes, except for the small class of mitochondrial membrane peptides, which were predicted as globular mitochondrial proteins (Table 3). All the four proteins are ATP synthase subunits (two subunit β, subunit δ, and the 24-kD subunit). Although the ATP synthase is annotated as inner membrane proteins in SWISSPROT, the F1 part of ATP synthase is not embedded within the membrane. So the apparent misclassification was an artifact caused by imprecise annotation, which when rectified, increased prediction accuracy of this class to 100%.

Application to jakobids ESTs with known location

We applied TESTLoc on jakobids ESTs of known (inferred) localization (see Methods). The results show a two times higher sensitivity in recognizing mitochondrial proteins compared to available tools or by homology to *Arabidopsis* mitochondrial proteins (Figure 8). But the false positive rate is high (75%). If TESTLoc is applied to infer jakobid mitochondrial proteome, most mitochondrial proteins would be identified, but the

prediction would also include proteins from other subcellular compartments. Obviously, the models trained with plant data perform less well with jakobids than with plant sequences. This is not surprising. Already the known mitochondrial targeting signals are poorly conserved across larger phylogenetic distances, and the same will likely apply to the hidden signals detected by SVM. Therefore, the availability of training data from taxonomically close (or moderately distant) relatives is crucial for building a predictor with good performance.

Conclusion

The results described above show that the SVM machine learning method, together with sequence features representing various amino acid compositions, predicts the sub-location of plant EST-derived proteins two times more accurately compared to existing tools. We implemented TESTLoc as a pipeline combining EST ORF prediction and localization prediction. This tool performs well for *Arabidopsis* data, and is able to identify mitochondrial proteins from jakobids with high sensitivity, but with low positive predictive value. The prediction of the jakobid proteome will improve with two types of new data: (i) experimentally validated training data from species closer related to jakobids than plants; (ii) non-coding genomic sequences to build species-specific model for ORF prediction by Prot4EST, which will soon become available as the Reclinomonas genome project is already well underway.

Acknowledgements

This work was supported by the Canadian Institute for Health Research (CIHR) Strategic Training Grant in Bioinformatics. We would like to thank Geneviève Galarneau for the work on TESTLoc, and Jean-François Théroux for implementation of gapped amino acid composition in the context of an undergraduate research internship in bioinformatics at the Université de Montréal.

References:

Assfalg, J., J. Gong, et al. (2009). "Supervised ensembles of prediction methods for subcellular localization." J Bioinform Comput Biol **7**(2): 269-85.

Bannai, H., Y. Tamada, et al. (2002). "Extensive feature detection of N-terminal protein sorting signals." Bioinformatics **18**(2): 298-305.

Barbe, L., E. Lundberg, et al. (2008). "Toward a confocal subcellular atlas of the human proteome." Mol Cell Proteomics **7**(3): 499-508.

Bhasin, M. and G. P. Raghava (2004). "ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST." Nucleic Acids Res **32**(Web Server issue): W414-9.

Blum, T., S. Briesemeister, et al. (2009). "MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction." BMC Bioinformatics **10**: 274.

Blum, T., S. Briesemeister, et al. (2009). "MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction." BMC Bioinformatics **10**(1): 274.

Boden, M. and J. Hawkins (2005). "Prediction of subcellular localization using sequence-biased recurrent networks." Bioinformatics **21**(10): 2279-86.

Brown, J. R. (2003). "Ancient horizontal gene transfer." Nat Rev Genet **4**(2): 121-32.

Burger, G., M. W. Gray, et al. (2003). "Mitochondrial genomes: anything goes." Trends Genet **19**(12): 709-16.

Casadio, R., P. L. Martelli, et al. (2008). "The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation." Brief Funct Genomic Proteomic **7**(1): 63-73.

Chou, K. C. (2001). "Prediction of protein cellular attributes using pseudo-amino acid composition." Proteins **43**(3): 246-55.

Chou, K. C. and Y. D. Cai (2002). "Using functional domain composition and support vector machines for prediction of protein subcellular location." J Biol Chem **277**(48): 45765-9.

Claros, M. G. and P. Vincens (1996). "Computational method to predict mitochondrially imported proteins and their targeting sequences." Eur J Biochem **241**(3): 779-86.

Devlin, T. M. (1992). The Textbook of Biochemistry New York, Wiley-Liss Inc.

Doolittle, W. F., Y. Boucher, et al. (2003). "How big is the iceberg of which organellar genes in nuclear genomes are but the tip?" Philos Trans R Soc Lond B Biol Sci **358**(1429): 39-57; discussion 57-8.

Emanuelsson, O., H. Nielsen, et al. (2000). "Predicting subcellular localization of proteins based on their N-terminal amino acid sequence." J Mol Biol **300**(4): 1005-16.

Fan, R. E., P. H. Chen, et al. (2005). "Working set selection using the second order information for training SVM." Journal of Machine Learning Research **6**: 1889-1918.

Forner, F., L. J. Foster, et al. (2006). "Quantitative proteomic comparison of rat mitochondria from muscle, heart, and liver." Mol Cell Proteomics **5**(4): 608-19.

Guda, C., E. Fahy, et al. (2004). "MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins." Bioinformatics **20**(11): 1785-94.

Guda, C. and S. Subramaniam (2005). "pTARGET [corrected] a new method for predicting protein subcellular localization in eukaryotes." Bioinformatics **21**(21): 3963-9.

Hatzigeorgiou, A. G., P. Fiziev, et al. (2001). "DIANA-EST: a statistical analysis." Bioinformatics **17**(10): 913-9.

Heazlewood, J. L., K. A. Howell, et al. (2003). "Towards an analysis of the rice mitochondrial proteome." Plant Physiol **132**(1): 230-42.

Heazlewood, J. L. and A. H. Millar (2005). "AMPDB: the Arabidopsis Mitochondrial Protein Database." Nucleic Acids Res **33**(Database issue): D605-10.

Heazlewood, J. L., J. S. Tonti-Filippini, et al. (2004). "Experimental analysis of the *Arabidopsis* mitochondrial proteome highlights signaling and regulatory components, provides assessment of targeting prediction programs, and indicates plant-specific mitochondrial proteins." Plant Cell **16**(1): 241-56.

Horton, P., K. J. Park, et al. (2007). "WoLF PSORT: protein localization predictor." Nucleic Acids Res **35**(Web Server issue): W585-7.

Hua, S. and Z. Sun (2001). "Support vector machine approach for protein subcellular localization prediction." Bioinformatics **17**(8): 721-8.

Huang, Y. and Y. Li (2004). "Prediction of protein subcellular locations using fuzzy k-NN method." Bioinformatics **20**(1): 21-8.

Iseli, C., C. V. Jongeneel, et al. (1999). "ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences." Proc Int Conf Intell Syst Mol Biol: 138-48.

Johnson, D. T., R. A. Harris, et al. (2007). "Tissue heterogeneity of the mammalian mitochondrial proteome." Am J Physiol Cell Physiol **292**(2): C689-97.

Kumar, A., S. Agarwal, et al. (2002). "Subcellular localization of the yeast proteome." Genes Dev **16**(6): 707-19.

Lang, B. F., M. W. Gray, et al. (1999). "Mitochondrial genome evolution and the origin of eukaryotes." Annu Rev Genet **33**: 351-97.

Li, J., T. Cai, et al. (2009). "Proteomic analysis of mitochondria from *Caenorhabditis elegans*." Proteomics.

Liu, J., S. Kang, et al. (2007). "Meta-prediction of protein subcellular localization with reduced voting." Nucleic Acids Res **35**(15): e96.

Lu, Z., D. Szafron, et al. (2004). "Predicting subcellular localization of proteins using machine-learned classifiers." Bioinformatics **20**(4): 547-56.

Maximo, V., J. Lima, et al. (2009). "Mitochondria and cancer." Virchows Arch **454**(5): 481-95.

Meisinger, C., A. Sickmann, et al. (2008). "The mitochondrial proteome: from inventory to function." Cell **134**(1): 22-4.

Mootha, V. K., J. Bunkenborg, et al. (2003). "Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria." Cell **115**(5): 629-40.

Nair, R. and B. Rost (2002). "Inferring sub-cellular localization through automated lexical analysis." Bioinformatics **18 Suppl 1**: S78-86.

Nair, R. and B. Rost (2005). "Mimicking cellular sorting improves prediction of subcellular localization." J Mol Biol **348**(1): 85-100.

O'Brien, E. A., L. B. Koski, et al. (2007). "TBestDB: a taxonomically broad database of expressed sequence tags (ESTs)." Nucleic Acids Res **35**(Database issue): D445-51.

Park, K. J. and M. Kanehisa (2003). "Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs." Bioinformatics **19**(13): 1656-63.

Pierleoni, A., P. L. Martelli, et al. (2006). "BaCelLo: a balanced subcellular localization predictor." Bioinformatics **22**(14): e408-16.

Reinders, J. and A. Sickmann (2007). "Proteomics of yeast mitochondria." Methods Mol Biol **372**: 543-57.

Reinhardt, A. and T. Hubbard (1998). "Using neural networks for prediction of the subcellular location of proteins." Nucleic Acids Res **26**(9): 2230-6.

Rodriguez-Ezpeleta, N., H. Brinkmann, et al. (2007). "Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans." Curr Biol **17**(16): 1420-5.

Sarda, D., G. H. Chua, et al. (2005). "pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties." BMC Bioinformatics **6**: 152.

Scott, M. S., D. Y. Thomas, et al. (2004). "Predicting subcellular localization via protein motif co-occurrence." Genome Res **14**(10A): 1957-66.

Shatkay, H., A. Hoglund, et al. (2007). "SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data." Bioinformatics.

Shen, Y. Q. and G. Burger (2007). "'Unite and conquer': enhanced prediction of protein subcellular localization by integrating multiple specialized tools." BMC Bioinformatics **8**: 420.

Shen, Y. Q. and G. Burger (2009). "Plasticity of a key metabolic pathway in fungi." Funct Integr Genomics **9**(2): 145-51.

Sickmann, A., J. Reinders, et al. (2003). "The proteome of *Saccharomyces cerevisiae* mitochondria." Proc Natl Acad Sci U S A **100**(23): 13207-12.

Small, I., N. Peeters, et al. (2004). "Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences." Proteomics **4**(6): 1581-90.

Smith, D. G., R. M. Gawryluk, et al. (2007). "Exploring the mitochondrial proteome of the ciliate protozoon *Tetrahymena thermophila*: direct analysis by tandem mass spectrometry." J Mol Biol **374**(3): 837-63.

Taylor, S. W., E. Fahy, et al. (2003). "Characterization of the human heart mitochondrial proteome." Nat Biotechnol **21**(3): 281-6.

Wang, Z. Y., D. M. Soanes, et al. (2007). "Functional analysis of lipid metabolism in *Magnaporthe grisea* reveals a requirement for peroxisomal fatty acid beta-oxidation during appressorium-mediated plant infection." Mol Plant Microbe Interact **20**(5): 475-91.

Xie, D., A. Li, et al. (2005). "LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST." Nucleic Acids Res **33**(Web Server issue): W105-10.

Yu, C. S., Y. C. Chen, et al. (2006). "Prediction of protein subcellular localization." Proteins.

Tables

Table 1. Number of sequences (from *Arabidopsis* and all plants tested) used in this study

Classes	cyt	end	ext	mit	mitm	nuc	per	pla	vac	total
Arab ESTs (after clustering)	53	4	9	167	4	40	4	3	5	289
Plant ESTs	96	6	23	182	32	137	8	6	16	506
Expanded ESTs	343	26	94	448	70	369	24	8	41	1423
Expanded ESTs (after clustering)	122	11	48	309	36	260	12	7	29	834
Jakobid ESTs (before clustering)	65	38	3	114	45	216	13	16	3	513
Jakobid ESTs (after clustering)	44	38	3	103	40	204	13	16	3	464

Abbreviations: cyt, cytosol; end, endoplasmatic reticulum; ext, extracellular; mit, mitochondrion; mitm, mitochondrial membrane; nuc, nucleus; per, peroxisome; pla, plasma membrane; vac, vacuole; No.Seq: number of sequences; Arab, *Arabidopsis*

Table 2. Amino acids grouped according to their chemical properties or structures

Group C, showed good performance in CELLO		Group D: Devlin (Devlin 1992) classifies amino acids along structural lines		
Property	Amino acid	Superstructure	Structure	Amino Acid
Acidic	D, E	Monoamino, moncarboxylic		G,A
Basic	H,K,R		Unsubstituted	V,L,I
Aromatic	F,W, Y		Heterocyclic	P,F
Small hydroxyl	S,T		Aromatic	W,Y
Sulphur-containing	C,M		Thioether	M
Aliphatic1	A,G,P		Hydroxy	S, T
Aliphatic2	I, L,V		Mercapto	C
Amide	N,Q		Carboxamide	N,Q
		Monamino, dicarboxylic		D,E
		Diamino, monocarboxylic		H, K, R

Table 3 The independent evaluation results of different prediction schemes

Prediction schemes		cyt	end	ext	mit	mitm	nuc	per	pla	vac
Best result from individual features[1]	SN	53.94	20	81	90.64	72.5	83.1	30	50	80.01
	PP	87.7	20	96	68.34	93.5	89.03	30	50	97.5
	MCC	0.64	0.2	0.87	0.63	0.80	0.79	0.3	0.5	0.87
Integration of predictions	SN	19.32	20	46	82.17	45	77.31	10	0	19.99
	PP	46.83	20	71.17	58.01	91.67	72.06	10	0	50
	MCC	0.21	0.2	0.54	0.42	0.62	0.57	0.09	0	0.30
Integration of attributes	SN	9.22	40	58.5	80.87	42.5	77.31	0	0	16.66
	PP	27.83	40	81.88	55.31	88.33	73.09	0	0	40
	MCC	0.06	0.4	0.65	0.38	0.59	0.58	0	0	0.25
Integration of predictions from top three features	SN	48.03	20	77	94.18	62.5	79.99	20	50	66.67
	PP	88.79	20	98	65.82	93.5	95.16	20	33.5	100
	MCC	0.61	0.2	0.86	0.63	0.74	0.81	0.2	0.38	0.79
Integration of attributes from top three features	SN	52.2	20	83	93.54	72.5	82.7	30	50	76.68
	PP	88.62	20	96	68.06	93.5	92.56	30	50	100
	MCC	0.64	0.2	0.88	0.65	0.81	0.81	0.3	0.5	0.86
Arabidopsis validation	SN	60.4	100	100	93.4	0[2]	90.2	80	100	100
	PP	86.5	100	100	84.3	0	97.4	100	100	100
	MCC	0.67	1	1	0.72	-0.02	0.93	0.75	0.89	1

1 The result was obtained from SVM trained with 4^{th}-order amino acid composition

2 the misclassification of *Arabidopsis* mitochondrial membrane proteins was an artifact caused by imprecise annotation

Figure legends

Figure 1. Prediction performance of top-ranked available tools and TESTLoc for mitochondrial proteins from plant EST-derived peptides. Desired results should be located in the top left region of the plot area, indicating high true positive rate and low false positive rate. True positive rate=sensitivity, false positive rate=1- positive predictive value

Figure 2. Selection of *Arabidopsis* ESTs.

Figure 3. Fragmentation of plant ESTs to expand the EST-peptide data. Bars: full-length proteins; lines: fragments of the proteins. Proteins shorter than 200 aa remain unchanged. Proteins ranging from 200 to 400 aa were fragmented into two parts, each part ranging from 140 to 260 aa: the N-terminus part which starts within 80 aa of the starting methionine, and C-terminal part including the C-terminus amino acid. Proteins longer than 400 aa were fragmented into three parts. In addition to the two terminal parts, a middle-region piece was created, whose first amino acid was in the N-terminal half part of the protein, but after the 80^{th} amino acid.

Figure 4. Sequence identities within expanded data set, calculated from BLASTP alignment.

Figure 5. Training and evaluation of SVM. The procedure in each dash box was repeated ten times. The whole expanded plant EST-peptides were randomly divided into ten parts, with nine parts combined and used to construct the SVM model, and the remaining one to evaluate the model. The construction-evaluation procedure was repeated ten times. In each round, the combined data for model construction was further divided randomly into ten data sets, in which nine sets combined as training data, the rest one as test data. Selection of

SVM parameters was done by ten-fold cross validation, and SVM built with optimal parameters was assessed by the evaluation data.

Figure 6. The independent evaluation for SVMs based on different orders of amino acid composition. A. sensitivity of each class. B. positive predictive value (PPV) of each class. C. matthews correlation coefficient (MCC) of each class. Sensitivity for mitochondrial proteins increased with order of composition, but dropped for other classes after the 4^{th}-order. As to PPV, different classes show different trends. The PPV for mitochondrial proteins reaches the peak at the 4^{th}-order. The peak of MCC for most classes was obtained at the 4^{th}-order. Similar trends were observed for group C and group D composition (see Supplementary Table 2).

Figure 7. Integration of predictions from SVM models based on individual features. From each of the 41 SVM models built with different sequence features, the probabilities of the query sequence predicted as each class were used as input for the 2^{nd}-layer SVM.

Figure 8. Comparison of available tools and TESTLoc for the prediction power on recognizing mitochondrial proteins from jakobid ESTs. TESTLoc shows much higher sensitivity, but at the cost of low specificity.

Figures

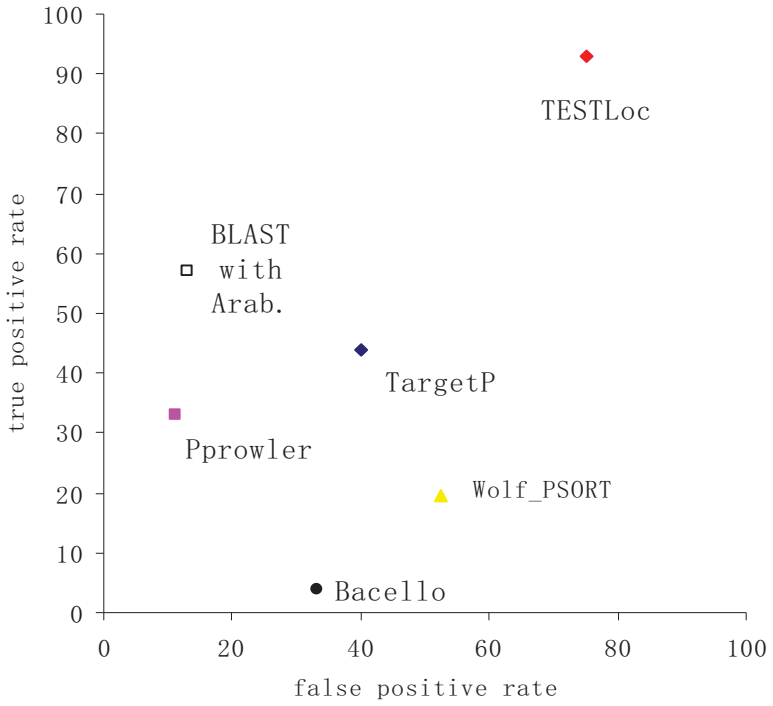


Figure 1. Prediction performance of top-ranked available tools and TESTLoc for mitochondrial proteins from plant EST-derived peptides. Desired results should be located in the top left region of the plot area, indicating high true positive rate and low false positive rate. True positive rate=sensitivity, false positive rate=1- positive predictive value

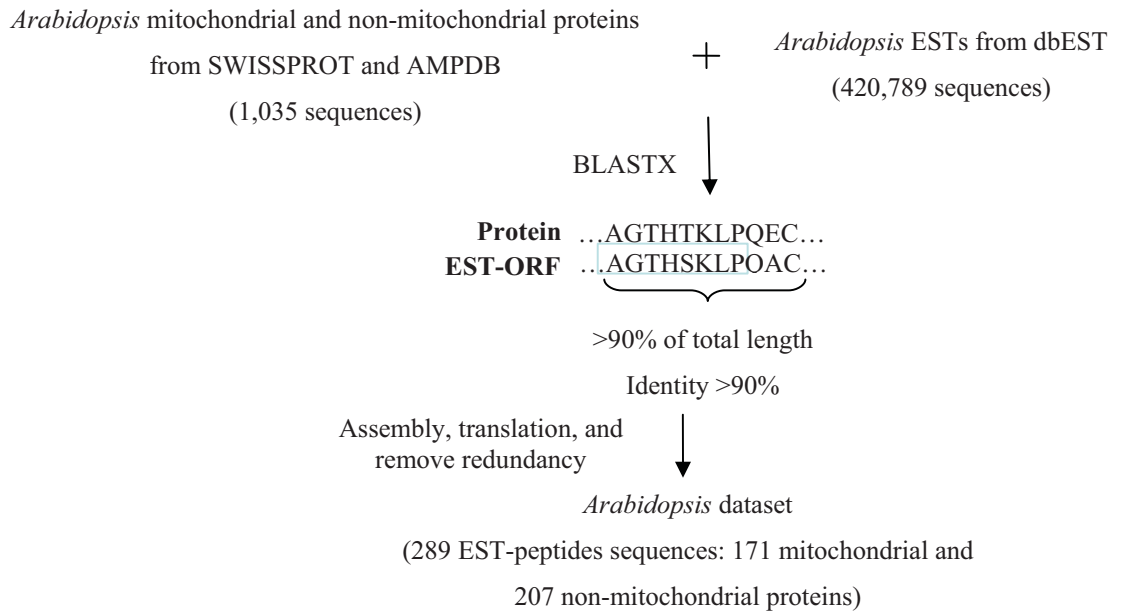


Figure 2. Selection of *Arabidopsis* ESTs.

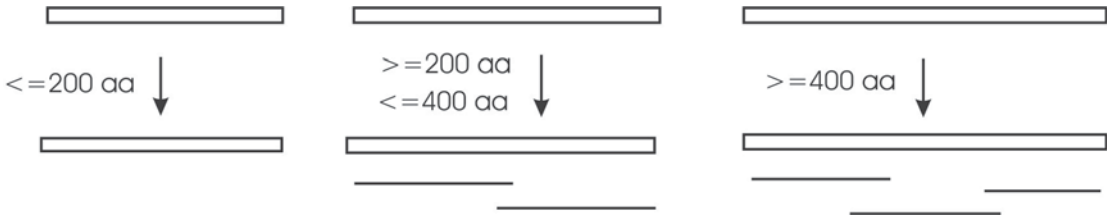


Figure 3. Fragmentation of plant ESTs to expand the EST-peptide data. Bars: full-length proteins; lines: fragments of the proteins. Proteins shorter than 200 aa remain unchanged. Proteins ranging from 200 to 400 aa were fragmented into two parts, each part ranging from 140 to 260 aa: the N-terminus part which starts within 80 aa of the starting methionine, and C-terminal part including the C-terminus amino acid. Proteins longer than 400 aa were fragmented into three parts. In addition to the two terminal parts, a middle-region piece was created, whose first amino acid was in the N-terminal half part of the protein, but after the 80th amino acid.

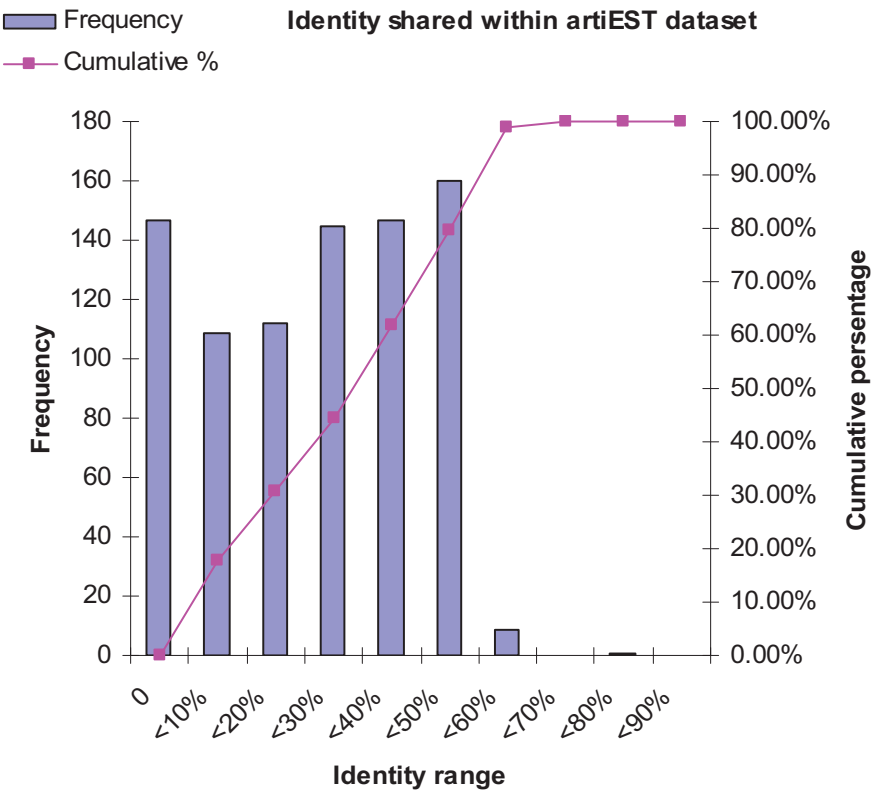


Figure 4. Sequence identities within expanded data set, calculated from BLASTP alignment.

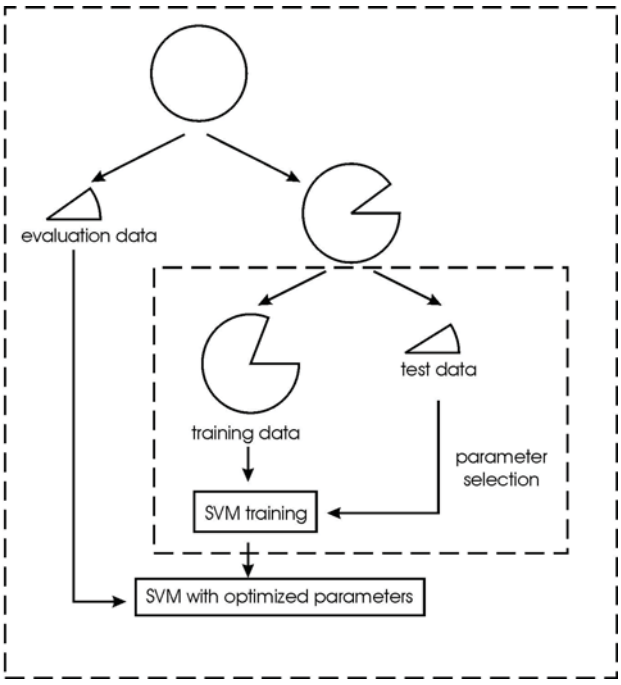
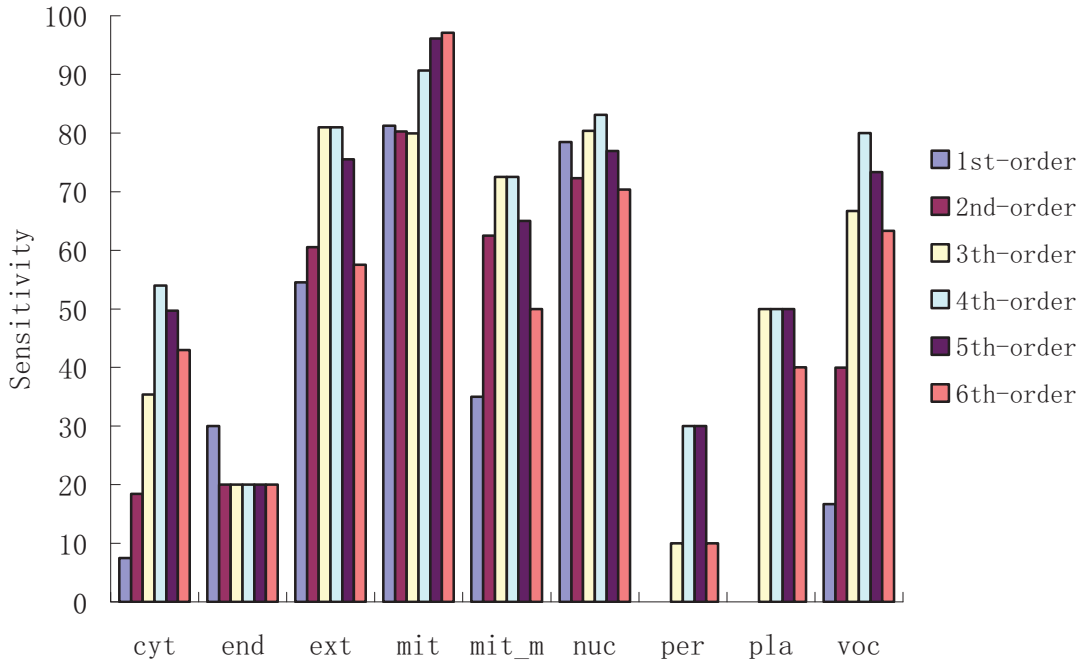
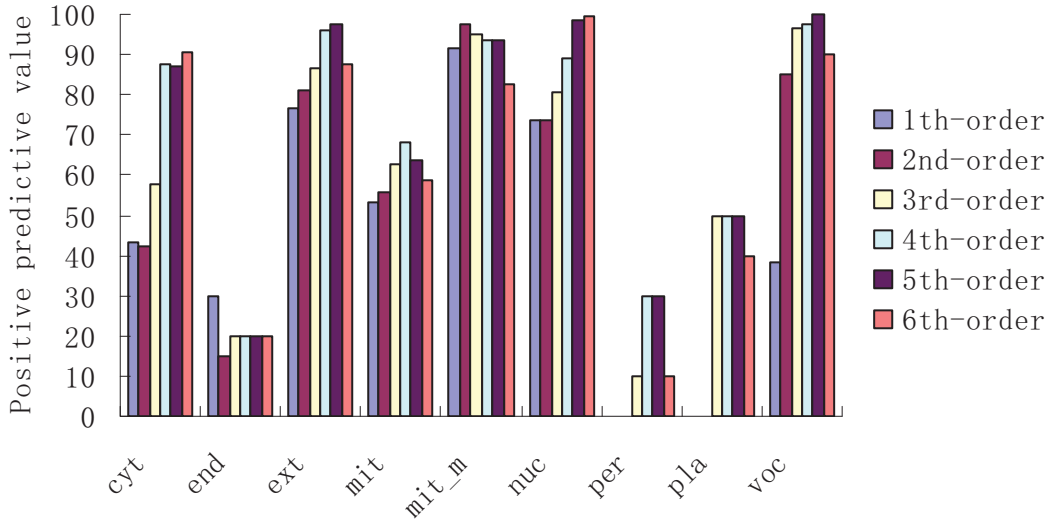


Figure 5. Training and evaluation of SVM. The procedure in each dash box was repeated ten times. The whole expanded plant EST-peptides were randomly divided into ten parts, with nine parts combined and used to construct the SVM model, and the remaining one to evaluate the model. The construction-evaluation procedure was repeated ten times. In each round, the combined data for model construction was further divided randomly into ten data sets, in which nine sets combined as training data, the rest one as test data. Selection of SVM parameters was done by ten-fold cross validation, and SVM built with optimal parameters was assessed by the evaluation data.


Sensitivity
100
90
80
70
60
50
40
30
20
10
0
cyt
end
ext
mit
mit_m
nuc
per
pla
voc
1st-order
2nd-order
3th-order
4th-order
5th-order
6th-order


Positive predictive value
100
90
80
70
60
50
40
30
20
10
0
cyt
end
ext
mit
mit_m
nuc
per
pla
voc
1th-order
2nd-order
3rd-order
4th-order
5th-order
6th-order

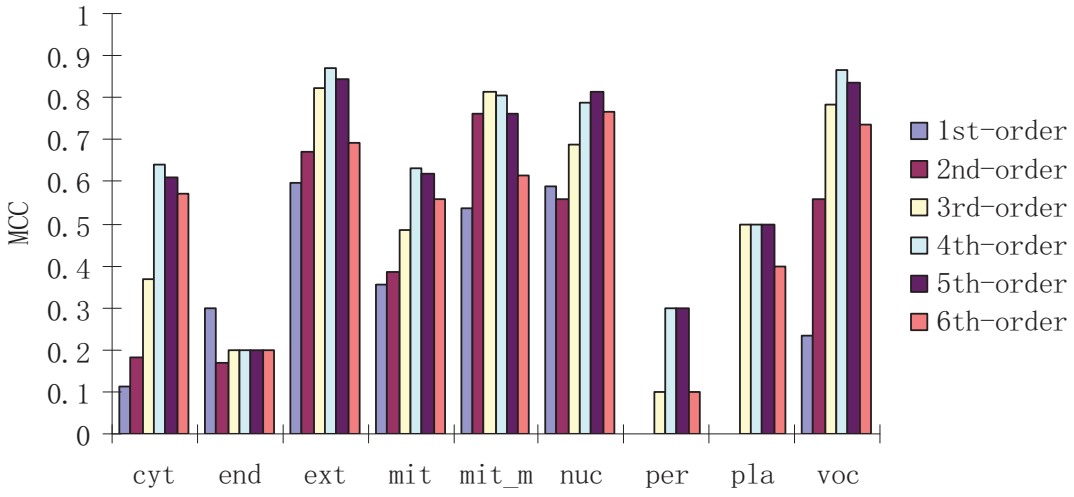


Figure 6. The independent evaluation for SVMs based on different orders of amino acid composition. A. sensitivity of each class. B. positive predictive value (PPV) of each class. C. matthews correlation coefficient (MCC) of each class. Sensitivity for mitochondrial proteins increased with order of composition, but dropped for other classes after the 4th-order. As to PPV, different classes show different trends. The PPV for mitochondrial proteins reaches the peak at the 4th-order. The peak of MCC for most classes was obtained at the 4th-order. Similar trends were observed for group C and group D composition (see Supplementary Table 2).

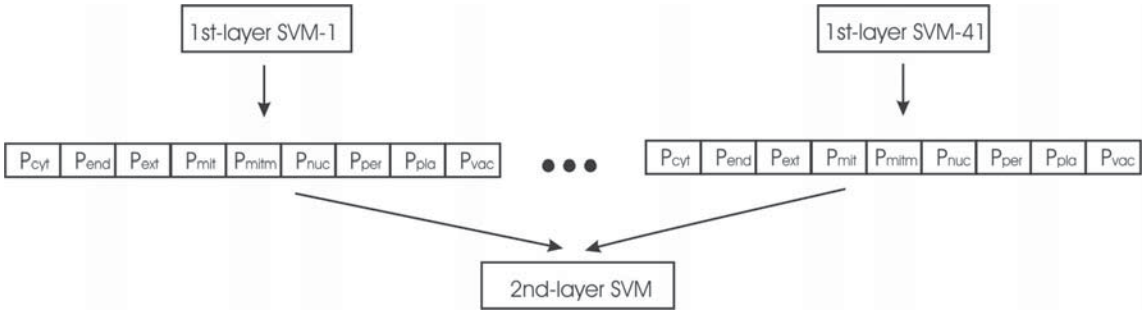


Figure 7. Integration of predictions from SVM models based on individual features. From each of the 41 SVM models built with different sequence features, the probabilities of the query sequence predicted as each class were used as input for the 2nd-layer SVM.

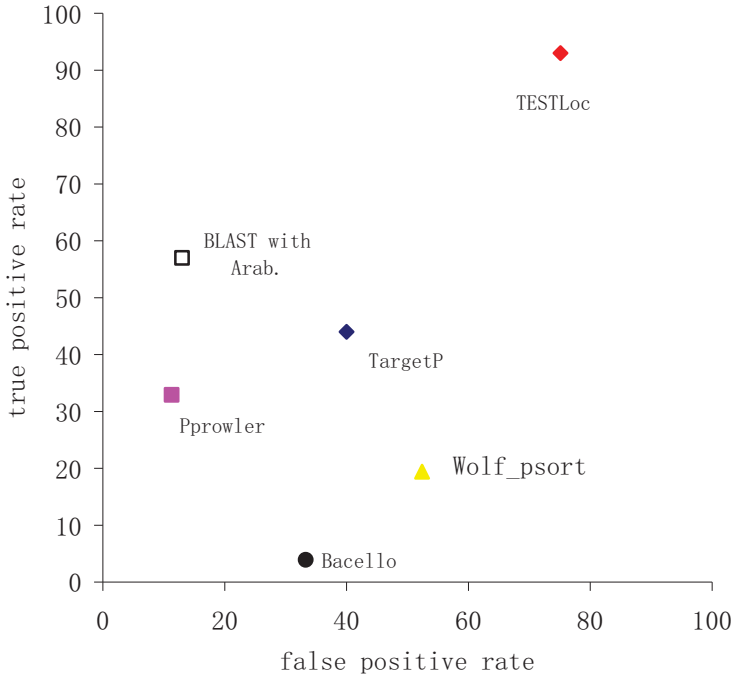


Figure 8. Comparison of available tools and TESTLoc for the prediction power on recognizing mitochondrial proteins from jakobid ESTs. The open square shows the selection of jakobid mitochondrial proteins by blasting with Arabidopsis mitochondrial proteins. TESTLoc shows much higher sensitivity, but at the cost of low specificity.

Supplementary Table 1 List of available subcellular localization prediction methods

Sequence feature		Name of the predictor or author	Computational methods
Sequence similarity and text annotation		PA-SUB	Naïve Bayes
		EpiLoc	Support Vector Machine
Gene Ontology term		ProLoc-GO	Genetic algorithm based method combined with SVM
InterPro domains and specific membrane domains		PSLT	Likelihood calculated by Bayes' rule
Targeting peptide		Predotar	Neural network
		TargetP	Neural network
		iPSORT	Alphabet indexing and pattern rule
		Protein prowler	Neural network
Physicochemical properties		pSLIP	Support Vector Machine
Amino acid composition		Subloc	Support Vector Machine
		NNPSL	Neural network
Integrated	Position-specific scoring matrix + four part amino acid composition	LOCSVMPSI	Support Vector Machine
	amino acid composition 33 physicochemical properties dipeptide composition PSI-blast result Combined feature of the above	ESLpred	Support Vector Machine
	amino acid composition quasi-sequence-order (up to 13 gaps) physicochemical properties (hydrophobicity, hydrophilicity, side-chain volume)	Cai et al	Support Vector Machine
	Evolutionary profiles global amino acid composition 50N-terminal amino acid composition amino acid composition in three secondary structure states output of signalP	LOCtree	Support Vector Machine
	Amino acid composition and paired amino acid composition	Yuan	Hidden Markov Model
	A set of sequence-derived features	PSORTII	K Nearest Neighbors
	Features from iPSORT and PSORTII, together with amino acid content	WoLF PSORT	Weighted K Nearest Neighbors
	Amino acid composition +dipeptide+physicochemical	Gao, et al	K Nearest Neighbors

	properties		
	Pfam domains occurrence, amino acid composition and PI value	MITOPRED	Score of different features
	Pfam domains occurrence and amino acid composition	pTARGET	Score of different features
	targeting sequence and hydrophobicity characteristics	MitoProt	Multivariate analysis
	amino acid composition, targeting signals, motif, and text search	SherLoc	Support Vector Machine

Supplementary Table 2. Performance of predictions based on each sequence feature (Each number is averaged by 10-round evaluation)

Matthews correlation coefficient value									
natural amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
1st-order	0.11	0.3	0.59	0.35	0.54	0.584	0	0	0.23
2nd-order	0.18	0.17	0.67	0.38	0.764	0.554	0	-0.003	0.56
3rd-order	0.36	0.2	0.82	0.48	0.814	0.684	0.1	0.5	0.78
4th-order	0.64	0.2	0.86	0.63	0.804	0.784	0.3	0.5	0.87
5th-order	0.61	0.2	0.84	0.62	0.764	0.814	0.3	0.5	0.83
6th-order	0.57	0.2	0.69	0.56	0.614	0.764	0.1	0.4	0.73
group C amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
1st-order	0.05	0.1	0.3734	0.30	0.49	0.49	0	0	-0.01
2nd-order	0.008	0	0.46	0.30	0.7	0.51	0	0	0.36
3rd-order	0.05	0.17	0.66	0.37	0.73	0.54	0	0	0.56
4th-order	0.25	0.2	0.65	0.38	0.76	0.52	0	0	0.73
5th-order	0.47	0.2	0.76	0.49	0.77	0.63	0.2	0.5	0.77
6th-order	0.60	0.2	0.82	0.63	0.80	0.74	0.3	0.5	0.83
7th-order	0.64	0.2	0.86	0.59	0.74	0.73	0.3	0.5	0.83
8th-order	0.63	0.2	0.86	0.6	0.70	0.76	0.3	0.5	0.83
group D amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
1st-order	0.09	0.09	0.46	0.31	0.49	0.46	0	0	0.34
2nd-order	0.06	0.1	0.59	0.24	0.61	0.43	0	0	0.60
3rd-order	0.08	0.19	0.64	0.3	0.64	0.45	0	0	0.61
4th-order	0.27	0.17	0.82	0.41	0.78	0.56	0	0.2	0.77
5th-order	0.52	0.4	0.84	0.52	0.77	0.66	0.3	0.5	0.83
6th-order	0.59	0.2	0.89	0.61	0.80	0.72	0.3	0.5	0.85
7th-order	0.62	0.2	0.85	0.62	0.76	0.77	0.3	0.5	0.83
8th-order	0.6	0.2	0.76	0.58	0.74	0.75	0.1	0.5	0.83
natural amino acid									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
one_gap	0.13	0.39	0.58	0.34	0.67	0.54	0	0	0.45
two_gap	0.12	0.37	0.76	0.33	0.65	0.51	0	0	0.41
three_gap	0.08	0.37	0.70	0.34	0.74	0.52	0	0	0.52
four_gap	0.07	0.2	0.51	0.3	0.64	0.516	0	0	0.44
five_gap	0.10	0.2	0.53	0.28	0.68	0.51	0	0	0.47
six_gap	0.07	0.3661	0.65	0.33	0.69	0.50	0	-0.002	0.56
group C amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
one_gap	0.05	0	0.43	0.27	0.50	0.45	0	0	0.05

two_gap	0.09	0.1	0.47	0.24	0.58	0.44	0	0	-0.01
three_gap	0.16	0.2	0.42	0.25	0.54	0.45	0	0	-0.01
four_gap	0.14	0.1	0.37	0.24	0.60	0.42	0	0	0.05
five_gap	-0.01	0	0.45	0.26	0.61	0.43	0	0	0.11
six_gap	-0.007	0	0.49	0.19	0.6	0.36	0	0	0
group D amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
one_gap	0.13	0.1	0.49	0.29	0.70	0.46	0	0	0.28
two_gap	0.12	0	0.62	0.28	0.50	0.45	0	0	0.15
three_gap	0.16	0.2	0.51	0.26	0.54	0.43	0	0	0.37
four_gap	0.11	0.1	0.47	0.26	0.42	0.44	0	-0.002	0.05
five_gap	0.04	0	0.49	0.28	0.624	0.42	0	-0.004	0.42
six_gap	0.18	0	0.51	0.22	0.54	0.32	0	0	0.40
aaindex	0.10	0.3	0.60	0.39	0.60	0.58	0	0	0.31
sensitivity (%)									
natural amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
1st-order	7.5	30	54.5	81.22	35	78.47	0	0	16.65
2nd-order	18.41	20	60.5	80.24	62.5	72.3	0	0	39.98
3rd-order	35.4	20	81	79.95	72.5	80.39	10	50	66.68
4th-order	53.94	20	81	90.64	72.5	83.1	30	50	80.01
5th-order	49.69	20	75.5	96.12	65	76.92	30	50	73.34
6th-order	42.96	20	57.5	97.1	50	70.38	10	40	63.34
group C amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
1st-order	5.83	10	32	86.74	35	63.07	0	0	0
2nd-order	3.33	0	41	81.55	55	68.08	0	0	36.65
3rd-order	3.33	20	59	84.41	57.5	71.53	0	0	39.98
4th-order	17.74	20	58.5	79.93	67.5	70.39	0	0	60
5th-order	41.28	20	73	82.53	67.5	75.76	20	50	63.34
6th-order	51.36	20	79	90.94	72.5	80.01	30	50	73.34
7th-order	53.04	20	79	91.26	62.5	76.54	30	50	73.34
8th-order	52.21	20	77.5	96.13	57.5	71.15	30	50	73.34
group D amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
1st-order	10.08	10	40	82.19	30	65.77	0	0	23.32
2nd-order	8.32	10	48.5	74.46	45	65.37	0	0	53.33
3rd-order	8.32	20	59	77.04	50	66.94	0	0	56.68
4th-order	29.47	20	81	76.06	70	73.46	0	20	63.34
5th-order	43.94	40	83	82.22	67.5	78.08	30	50	73.34
6th-order	49.77	20	83	89.64	72.5	79.23	30	50	76.67
7th-order	51.36	20	77	95.48	65	75.01	30	50	73.34

8th-order	48.79	20	63	96.78	62.5	70	10	50	73.34
natural amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
one_gap	15.16	40	56.5	76.37	50	75.77	0	0	29.98
two_gap	8.41	40	69	79.29	47.5	73.08	0	0	26.65
three_gap	15.97	40	71	71.85	60	75.37	0	0	39.99
four_gap	10.82	20	46.5	75.43	50	73.47	0	0	29.98
five_gap	10.82	20	44.5	76.34	62.5	69.23	0	0	33.3
six_gap	8.33	40	60.5	80.9	55	66.52	0	0	46.65
group C amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
one_gap	5.83	0	36	79.6	37.5	69.62	0	0	3.33
two_gap	6.74	10	40	79.93	42.5	64.6	0	0	0
three_gap	16.74	20	40.5	75.74	47.5	65.76	0	0	0
four_gap	12.5	10	34	77.32	45	64.61	0	0	3.33
five_gap	0	0	28	82.17	50	66.13	0	0	6.66
six_gap	0	0	36	79.3	45	60.37	0	0	0
group D amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
one_gap	10.98	10	44	78.97	55	67.31	0	0	16.65
two_gap	8.32	0	54.5	79.63	35	68.06	0	0	9.99
three_gap	12.57	20	48	75.72	40	67.69	0	0	23.31
four_gap	8.32	10	42.5	79.61	32.5	66.15	0	0	3.33
five_gap	1.74	0	40	80.9	45	66.15	0	0	29.99
six_gap	10.98	0	40	74.75	40	61.93	0	0	26.65
aaindex	14.24	30	60.5	78.61	45	78.47	0	0	23.32
specificity (%)									
natural amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
1st-order	97.84	100	97.74	54.2	99.62	80.72	100	100	99.24
2nd-order	94.55	99.81	98.61	58.69	99.79	83.01	100	99.63	99.43
3rd-order	94.71	100	98.91	69.88	99.65	88.27	100	100	99.84
4th-order	98.37	100	99.69	74.53	99.53	94.3	100	100	99.84
5th-order	98.5	100	99.84	67.58	99.53	99.33	100	100	100
6th-order	98.95	100	99.84	58.6	99.52	99.77	100	100	100
group C amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
1st-order	97.38	100	97.91	41.99	99.32	84.27	100	100	99.57
2nd-order	97.37	100	97.66	48.42	99.79	82.7	100	100	97.69
3rd-order	98.63	99.79	98.43	52.67	100	82.21	100	100	99.79
4th-order	96.96	100	98.2	58.19	99.19	81.54	100	100	99.82
5th-order	96.58	100	98.61	68.04	99.48	87.42	100	100	100

6th-order	97.99	100	99.02	73.8	99.52	92.63	100	100	100
7th-order	98.61	100	99.84	69.23	99.54	94.4	100	100	100
8th-order	98.63	100	99.84	65.04	99.53	98.91	100	100	100
group D amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
1st-order	95.48	99.36	97.8	47.79	100	79.6	100	100	99.79
2nd-order	94.64	100	98.83	49.41	99.5	77.61	100	100	99.09
3rd-order	95.45	99.78	98.42	52.93	99.62	78.19	100	100	98.92
4th-order	93.31	99.82	98.84	65.82	99.41	83.16	100	100	100
5th-order	97.39	100	98.99	71.1	99.5	87.79	100	100	100
6th-order	98.03	100	99.84	72.74	99.53	92.04	100	100	100
7th-order	98.46	100	99.84	67.91	99.54	97.62	100	100	100
8th-order	98.46	100	99.84	62.76	99.5	98.89	100	100	100
natural amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
one_gap	93.8	99.6	97.53	58.61	99.79	79.52	100	100	99.82
two_gap	97.28	99.81	99.19	54.7	99.78	79.09	100	100	99.6
three_gap	91.46	99.81	97.56	62.59	99.81	77.68	100	100	99.8
four_gap	94.33	100	97.64	54.31	99.38	77.97	100	100	99.52
five_gap	95.25	100	98.11	52	98.55	81.46	100	100	98.97
six_gap	96.87	99.42	98.35	52.56	99.61	82.43	100	99.81	99.17
group C amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
one_gap	97.41	100	97.72	46.95	99.78	75.92	100	100	99.57
two_gap	97.98	100	97.97	43.95	99.28	79.05	100	100	99.35
three_gap	95.02	100	97.38	48.83	98	78.87	100	100	99.17
four_gap	95.49	100	97.07	46.38	99.27	77.38	100	100	99.79
five_gap	99.6	100	99.58	42.86	98.91	76.95	100	100	100
six_gap	99.78	100	98.84	39.17	98.97	76.1	100	100	100
group D amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
one_gap	95.97	99.8	98.51	49.95	99.79	78.58	100	100	99.81
two_gap	97.69	100	98.31	48.45	99.11	76.89	100	100	99.78
three_gap	96.23	100	98.23	50.49	99.56	75.7	100	100	99.4
four_gap	97.07	100	97.86	45.97	99.13	77.62	100	99.8	100
five_gap	99.58	100	98.23	46.77	99.8	75.55	100	99.57	99.55
six_gap	97.5	100	99.11	47.11	99.34	70.52	100	100	99.76
aaindex	93.45	100	96.77	61.55	99.25	80.07	100	100	98.87
positive predictive value (%)									
natural amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
1st-order	43.34	30	76.42	53.11	91.67	73.73	0	0	38.33

2nd-order	42.45	15	81.17	55.8	97.5	73.51	0	0	85
3rd-order	57.66	20	86.5	62.45	95	80.65	10	50	96.67
4th-order	87.7	20	96	68.34	93.5	89.03	30	50	97.5
5th-order	87.14	20	97.5	63.83	93.5	98.54	30	50	100
6th-order	90.47	20	87.5	58.46	82.67	99.44	10	40	90
group C amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
1st-order	21.19	10	53.17	49.05	76.67	74.28	0	0	0
2nd-order	18.33	0	65.84	50.28	96.67	73.49	0	0	41.66
3rd-order	25	15	82.81	53.26	100	72.39	0	0	86.67
4th-order	61.32	20	76	55.43	92	70.12	0	0	95
5th-order	70.99	20	85.33	62.03	93.5	78.01	20	50	100
6th-order	84.04	20	88.5	68.35	93.5	85.87	30	50	100
7th-order	89.78	20	98	64.3	92.67	88.78	30	50	100
8th-order	89.36	20	98	62.21	92.67	97.29	30	50	100
group D amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
1st-order	31.36	10	63.34	51.25	90	68.72	0	0	56.67
2nd-order	33	10	81.67	48.51	91.67	66.75	0	0	76.67
3rd-order	38.26	20	78.34	51.95	90	66.11	0	0	71.68
4th-order	48.04	15	88	59.12	93	72.49	0	20	100
5th-order	77.79	40	88.83	63.87	93.5	78.22	30	50	100
6th-order	83.79	20	98	67.27	93.5	84.49	30	50	100
7th-order	87.46	20	98	64.38	92.67	94.62	30	50	100
8th-order	87.64	20	96.67	61	92.67	97.4	10	50	100
natural amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
one_gap	39.13	40	69.5	55.07	96.67	71.04	0	0	75
two_gap	44.5	35	91.34	52.81	96.67	69.64	0	0	70
three_gap	26.6	35	75.05	56.29	96.67	68.38	0	0	75
four_gap	30.05	20	67.5	51.91	89.17	69.29	0	0	73.33
five_gap	39.25	20	76	50.4	81.84	72.57	0	0	75
six_gap	21.25	35	78	52.26	93.34	71.83	0	0	77.5
group C amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
one_gap	25.84	0	66	49.54	75	66.62	0	0	10
two_gap	33.33	10	71.01	47.52	90	68.4	0	0	0
three_gap	42.91	20	54.17	49.01	72.5	68.75	0	0	0
four_gap	47	10	52.5	48.47	90	66.84	0	0	10
five_gap	0	0	83.34	47.87	80.84	66.6	0	0	20
six_gap	0	0	79.17	44.98	88.33	64.01	0	0	0
group D amino acid composition									
	cyt	end	ext	mit	mit_m	nuc	per	pla	voc
one_gap	45.67	10	61.67	50.69	96.67	67.79	0	0	50

two_gap	44.16	0	81.54	50.01	83.33	66.68	0	0	25
three_gap	52.5	20	63	50.03	81.67	64.89	0	0	65
four_gap	44.29	10	63	49.09	63.33	67.45	0	0	10
five_gap	20	0	71.67	50	95	64.07	0	0	63.33
six_gap	62.23	0	75.84	48.51	80	57.41	0	0	65
aaindex	31.85	30	69.63	58.14	88.33	72.47	0	0	48.33

Conclusions

With the focus on mitochondria, my thesis includes the development of new effective methods for predicting the subcellular localization of proteins, together with the subsequent large-scale *in silico* study of a key mitochondrial metabolic pathway across eukaryotes. Several aspects of this work are worthwhile to discuss in a wider context than was possible in the corresponding publications or the manuscript.

1. Localization is an important aspect of protein function

Our study shows that subcellular location prediction is an important asset in the annotation of newly discovered proteins, as it bears important clues about protein function. To reveal a protein's function, many aspects need to be addressed: the biochemical reaction it catalyzes, the molecular function it performs, bio-molecules it interacts with, the physiological conditions under which it is expressed, and the subcellular compartment where it is located. All these aspects are interrelated and interdependent, among which localization information provides valuable clues to infer the other aspects of function. A good example is the analysis of the ACD11 protein, as we showed in our study. Although ACD11 has been identified as a new ACAD family member several years ago, its molecular function has remained unknown. We found that this enzyme is present in almost all eukaryotic and many bacterial groups, suggesting that it carries out universal and fundamental functions for cellular life. Our localization prediction indicates that unlike other proteins in the eukaryotic ACAD family, which are always found in mitochondria, ACD11 is imported into peroxisomes in eukaryotes. Interestingly, in some fungal species

such as *Neurospora crassa* and *Magnaporthe grisea*, the peroxisomal beta oxidation is active while AOX, the first enzyme catalyzing the peroxisomal beta oxidation spiral, is absent. We and others proposed that in these species one of the ACAD enzymes takes over the function of AOX (Wang, Soanes et al. 2007; Shen and Burger 2009). The only clue to identify which ACAD carries out this function is the localization information. We found that ACD11 is the most likely AOX substitute, because of two reasons: it is present in the genomes of the above mentioned fungi, and the only ACAD enzyme that bears the peroxisomal targeting signal. Therefore, ACD11 is the key to understand the mechanism of the noncanonical peroxisomal beta oxidation and its relationship with the other two forms of the pathway.

2. From protein localization to pathway localization

Knowing the location of proteins with established molecular function helps to infer where the corresponding biological process takes place, what the physiological role of this process is, and how the various processes are integrated within the cell. In my work, localization prediction helped to elucidate where beta oxidation occurs in the eukaryotic cell. As two pathways (a mitochondrial and a peroxisomal form) exist, it proved difficult to distinguish the components of the two forms due to sequence similarity, especially in species which are phylogenetically distant from the model species in which the enzymes have been

characterized. But we show that localization prediction allows to clearly identify the pathway form that a given component belongs to.

The accuracy of prediction is crucial for such analysis. In order to get reliable localization prediction, we designed the tool YimLoc, which integrates the strength of heterogeneous prediction methods built with different training data, computational methods, and sequence features. YimLoc, when tested with known data, showed high prediction accuracy. The application of this tool to available genomic data revealed a most complicated scenario of beta oxidation in fungi. Our findings overturned the previous assumption that only the peroxisomal form of beta oxidation is present in this taxa group, and showed that the dual localization of this pathway is predominant in fungi. The prevalence of mitochondrial and peroxisomal beta oxidation in both animals and fungi suggests that the two forms were present in the common ancestor of opisthokonts, with the loss of one or both forms in certain fungal lineages.

3. The power of cross-taxon comparison

In order to gain a deeper insight into the mitochondrial beta oxidation, we applied a cross-taxon comparison of its key enzyme—ACAD. Unlike its perosisomal counterpart AOX, ACAD has a much broader taxonomic distribution, consists of more subfamilies with fine-tuned substrate specificity, and participates in amino acid degradation in addition to fatty acid degradation. We combined ortholog detection, phylogenetic reconstruction,

and localization prediction to investigate ACAD in more than 200 species, and further built the subfamily distribution profiles of ACAD in archaea, bacteria, and eukaryotes. In animals, the ACAD function is carried out by a large number of subfamilies, and the enzyme of each subfamily has a quite narrow range of substrate specificity. Fungi adopted a different strategy, with less numbers of subfamilies, but a broader spectrum of substrates for some subfamilies. In plants, only the few subfamilies involved in amino acid degradation were identified, in addition to the function-unknown ACD11. Some bacterial groups, such as actinobacteria and α-proteobacteria, have a complex profile of ACAD subfamilies. But in other bacterial and archaea groups, only a few or even none of the subfamilies were found. Confrontation of our results with the literature suggests that the ACAD subfamily profile corresponds well to the spectrum of fatty acids utilized. Therefore, our *in silico* analysis provides valuable hints to the energy metabolism of species for which experimental data is sparse.

Distribution mapping combined with phylogenetic analysis of ACAD subfamilies suggests a complicated evolution of this enzyme family, likely starting from a few enzymes for amino acid and short-chain fatty acid dehydrogenation. During evolution, gene duplication, horizontal gene transfer, gene loss, and functional convergence have led to a diversified toolset tailored for efficient fatty acids and amino acids utilization depending on the energy demand of a given species.

Our studies of the beta oxidation and ACAD subfamilies exemplify that localization prediction, combined with other *in silico* analyses and experimental observations from

model organisms, is a powerful approach to investigate metabolic processes in a taxonomically comprehensive manner including species where experimental data are poor. Large-scale integrated analysis also permits exploring mitochondria diversity, reflected by the different pathways and proteins they host.

However, one caveat of such genome-wide protein screens is that they are based on the identification of orthologs. As mentioned in the introduction, the function of missing subfamilies could be present in the species, performed by a protein of unrecognizably low sequence similarity, or by another non-homologous protein. However, this problem can sometimes be alleviated by carefully investigating the taxonomic distribution of subfamilies. For example, the IBD and ACADS are absent in ascomycetes. But more detailed inspection indicated that ACDSB in these species also functions as IBD and ACADS (a hypothesis that can be easily tested experimentally). Therefore, the apparent absence of orthologs should be treated with caution, and may require additional analyses to confirm the absence of function.

4 Factors that influence localization prediction accuracy

Being fundamental to the analysis of organelle biology, localization prediction must be accurate. Yet, *in silico* localization prediction methods are sensitive to many factors, such as the training data, the choice of sequence features, and computational methods. All these factors need to be considered in order to develop an effective localization predictor.

4.1 The influence of training data

We observed that the scheme trained with full-length sequences does not perform well for short fragments of the same sequence such as ESTs. This problem raises concern, as not only the number of ESTs is growing exponentially, but the sequence reads get shorter, due to the new, massively parallel sequencing technology. Developing effective methods to analyze short sequence fragments has become more urgent than ever. And for that, fine-tuning of prediction schemes for different data sets is of prime importance.

ESTs data are available for many more species than are genome sequences. In the context of my thesis, I developed a new tool TESTLoc to infer mitochondrial proteins from ESTs. As a proof of principle, we show that localization prediction based on ESTs is feasible, but the prediction scheme must be well adjusted to the data.

4.2 The influence of sequence features

The choice of sequence features revealed to be crucial, especially for the localization prediction based on ESTs. Since we did not know in advance which sequence feature would perform well, we experimented with more than 40 different features, and observed that the accuracy of the corresponding prediction schemes vary drastically.

Experimenting with the features also led to the discovery that the prediction based on the frequency of four-amino-acid words (4-mers) performed best, suggesting that certain localization signals may be captured by such words. However, a preliminary survey did not

find overrepresented 4-mers for the proteins from different locations, so that the reason for the good performance is currently unknown.

4.3 The influence of computational methods

The choice of the computation method, here the machine learning approach, also affects accuracy to a certain degree, as each method has its intrinsic advantages and limitations. For example, SVM is the most frequently used approach in localization prediction, for its ability to deal efficiently with large number of features, the control of over-fitting, and its robustness against class imbalance. But it is a 'black-box' procedure, and the biological reasons behind the prediction are difficult to extract. In contrast, decision trees are more transparent and generate biological interpretable rules that can be assessed by biologists. But it suffers much from class imbalance and biases its prediction towards the larger classes while sacrificing the accuracy of smaller ones.

There is no globally optimal machine learning approach, and the choice of learning scheme depends on the practical situation. When the human readability of the prediction procedure is important, such as in the case of many medical diagnoses, decision trees and KNN are good choices, because their decision-making procedure is easily understandable. If the input data is high-dimensional discrete, and neither the training time nor the human readability of the result is important, then ANN is the approach to consider. Here, we chose decision trees in one project and SVM in another. When integrating available localization prediction tools, we used decision tree, because it allowed to see which tool was selected

and how the various tools were integrated. For localization prediction based on ESTs, we employed SVM to save the computation time due to the large quantity of data and number of sequence features to experiment with.

Perspectives

The methodology and results from our work provide the basis for future investigations in several directions.

One aspect to pursue further addresses subcellular localization prediction for EST-derived peptides. We have shown that our method works well for plant data. It would be worthwhile to build EST-based localization predictors for other defined taxonomical groups. Currently, this is possible for animals and fungi, where large-scale ESTs and protein data sets are available with well defined localization information.

Another interesting future project would be to combine localization prediction with other data, such as expression patterns deduced from microarrays or EST sequences. This would allow to define the localizome (localization of each protein in the proteome, Kumar, Agarwal et al. 2002) in the cell. It is well documented that the composition of the mitochondrial proteome is dynamic, varying in different cell types and under different physiological conditions or developmental stages (Mootha, Bunkenborg et al. 2003). Changes of the mitochondrial proteome are also associated with many human diseases, in particular cancer (Maximo, Lima et al. 2009). Localization predictions combined with gene

expression pattern would reveal such dynamics, which could be used to track the change of mitochondria caused by or leading to abnormal cell conditions.

A third question that should be followed up is the taxonomic broad study of beta oxidation. Our study emphasizes the importance of investigating mitochondrial beta oxidation across a wider range of taxonomic groups. This will clarify the distribution of fatty acid degradation between mitochondria or peroxisomes, as well as track the carbon flux between the two organelles. Two groups of eukaryotes are of special interest: animal parasites and primitive eukaryotes. Studies in animal parasites, which often depend on fatty acids from their host, could lead to new treatment strategies and drug targets. Analyses of little derived eukaryotes will help to reconstruct the evolutionary history of this pathway.

Finally it would be worthwhile to study the enzymatics of ACAD proteins. The subfamily assignment and phylogenetic analysis presented in Chapter 3 have identified a pool of orthologs for each ACAD subfamily. This information is most valuable for inferring the highly conserved residues in each subfamily, which should be functionally and/or structurally important. Three-dimensional structures of several human ACAD enzymes are available. Mapping the conserved residues onto these structures will shed light on the enzymatic mechanism of ACAD proteins, and serve as a guide for the design of new experiments such as mutagenesis of these residues to see how the function or structure of the corresponding ACAD protein changes. Furthermore, comparisons of conserved residues will help to highlight the structural differences among subfamilies, which in turn will be instrumental to reveal the mechanism underlying the substrate preference.

In sum, *in silico* analysis of mitochondrial proteins, combined with the formulation of working hypotheses and guided experimental validation, is a powerful strategy for advancing our knowledge of mitochondrial biology.

References

Altenhoff, A. M. and C. Dessimoz (2009). "Phylogenetic and functional assessment of orthologs inference projects and methods." PLoS Comput Biol **5**(1): e1000262.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Amaral, S., P. J. Oliveira, et al. (2008). "Diabetes and the impairment of reproductive function: possible role of mitochondria and reactive oxygen species." Curr Diabetes Rev **4**(1): 46-54.

Andersson, S. G., A. Zomorodipour, et al. (1998). "The genome sequence of Rickettsia prowazekii and the origin of mitochondria." Nature **396**(6707): 133-40.

Andrade, M. A., S. I. O'Donoghue, et al. (1998). "Adaptation of protein surfaces to subcellular location." J Mol Biol **276**(2): 517-25.

Arvestad, L., A. C. Berglund, et al. (2003). "Bayesian gene/species tree reconciliation and orthology analysis using MCMC." Bioinformatics **19 Suppl 1**: i7-15.

Ashibe, B., T. Hirai, et al. (2007). "Dual subcellular localization in the endoplasmic reticulum and peroxisomes and a vital role in protecting against oxidative stress of fatty aldehyde dehydrogenase are achieved by alternative splicing." J Biol Chem **282**(28): 20763-73.

Assfalg, J., J. Gong, et al. (2009). "Supervised ensembles of prediction methods for subcellular localization." J Bioinform Comput Biol **7**(2): 269-85.

Bannai, H., Y. Tamada, et al. (2002). "Extensive feature detection of N-terminal protein sorting signals." Bioinformatics **18**(2): 298-305.

Barbe, L., E. Lundberg, et al. (2008). "Toward a confocal subcellular atlas of the human proteome." Mol Cell Proteomics **7**(3): 499-508.

Berg, J. M., J. L. Tymoczko, et al. (2002). Biochemistry. New York, W. H. Freeman and CO.

Bhasin, M. and G. P. Raghava (2004). "ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST." Nucleic Acids Res **32**(Web Server issue): W414-9.

Blum, T., S. Briesemeister, et al. (2009). "MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction." BMC Bioinformatics **10**(1): 274.

Boden, M. and J. Hawkins (2005). "Prediction of subcellular localization using sequence-biased recurrent networks." Bioinformatics **21**(10): 2279-86.

Bolender, N., A. Sickmann, et al. (2008). "Multiple pathways for sorting mitochondrial precursor proteins." EMBO Rep **9**(1): 42-9.

Boser, B. E., I. M. Guyon, et al. (1992). A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory, ACM Press.

Brown, J. R. (2003). "Ancient horizontal gene transfer." Nat Rev Genet **4**(2): 121-32.

Burger, G., M. W. Gray, et al. (2003). "Mitochondrial genomes: anything goes." Trends Genet **19**(12): 709-16.

Burges, C. J. C. (1998). "A Tutorial on Support Vector Machines for Pattern Recognition " Data Mining and Knowledge Discovery **2**(2): 121-167.

Casadio, R., P. L. Martelli, et al. (2008). "The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation." Brief Funct Genomic Proteomic **7**(1): 63-73.

Castresana, J. (2000). "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis." Mol Biol Evol **17**(4): 540-52.

Chacinska, A., S. Pfannschmidt, et al. (2004). "Essential role of Mia40 in import and assembly of mitochondrial intermembrane space proteins." EMBO J **23**(19): 3735-46.

Chen, J., X. Shi, et al. (2008). "Identification of novel modulators of mitochondrial function by a genome-wide RNAi screen in Drosophila melanogaster." Genome Res **18**(1): 123-36.

Chou, K. C. (2001). "Prediction of protein cellular attributes using pseudo-amino acid composition." Proteins **43**(3): 246-55.

Chou, K. C. and Y. D. Cai (2002). "Using functional domain composition and support vector machines for prediction of protein subcellular location." J Biol Chem **277**(48): 45765-9.

Claros, M. G. and P. Vincens (1996). "Computational method to predict mitochondrially imported proteins and their targeting sequences." Eur J Biochem **241**(3): 779-86.

Conway, D. J., C. Fanello, et al. (2000). "Origin of *Plasmodium falciparum* malaria is traced by mitochondrial DNA." Mol Biochem Parasitol **111**(1): 163-71.

Cotter, D., P. Guda, et al. (2004). "MitoProteome: mitochondrial protein sequence database and annotation system." Nucleic Acids Res **32**(Database issue): D463-7.

Crouch, P. J., K. Cimdins, et al. (2007). "Mitochondria in aging and Alzheimer's disease." Rejuvenation Res **10**(3): 349-57.

Deluca, T. F., I. H. Wu, et al. (2006). "Roundup: a multi-genome repository of orthologs and evolutionary distances." Bioinformatics **22**(16): 2044-6.

DeRisi, J. L., V. R. Iyer, et al. (1997). "Exploring the metabolic and genetic control of gene expression on a genomic scale." Science **278**(5338): 680-6.

Devlin, T. M. (1992). The Textbook of Biochemistry New York, Wiley-Liss Inc.

Dimmer, K. S., S. Fritz, et al. (2002). "Genetic basis of mitochondrial function and morphology in Saccharomyces cerevisiae." Mol Biol Cell **13**(3): 847-53.

Doolittle, W. F., Y. Boucher, et al. (2003). "How big is the iceberg of which organellar genes in nuclear genomes are but the tip?" Philos Trans R Soc Lond B Biol Sci **358**(1429): 39-57; discussion 57-8.

Drawid, A. and M. Gerstein (2000). "A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome." J Mol Biol **301**(4): 1059-75.

Duda, R. O., P. E. Hart, et al. (2001). Pattern classification. New York, Wiley.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-7.

Elstner, M., C. Andreoli, et al. (2009). "The mitochondrial proteome database: MitoP2." Methods Enzymol **457**: 3-20.

Emanuelsson, O., H. Nielsen, et al. (2000). "Predicting subcellular localization of proteins based on their N-terminal amino acid sequence." J Mol Biol **300**(4): 1005-16.

Endres, M., W. Neupert, et al. (1999). "Transport of the ADP/ATP carrier of mitochondria from the TOM complex to the TIM22.54 complex." EMBO J **18**(12): 3214-21.

Fan, H., C. Civalier, et al. (2006). "Detection of common disease-causing mutations in mitochondrial DNA (mitochondrial encephalomyopathy, lactic acidosis with stroke-like episodes MTTL1 3243 A>G and myoclonic epilepsy associated with ragged-red fibers MTTK 8344A>G) by real-time polymerase chain reaction." J Mol Diagn **8**(2): 277-81.

Fan, R. E., P. H. Chen, et al. (2005). "Working set selection using the second order information for training SVM." Journal of Machine Learning Research **6**: 1889-1918.

Forner, F., L. J. Foster, et al. (2006). "Quantitative proteomic comparison of rat mitochondria from muscle, heart, and liver." Mol Cell Proteomics **5**(4): 608-19.

Goldberg, A. V., S. Molik, et al. (2008). "Localization and functionality of microsporidian iron-sulphur cluster assembly proteins." Nature **452**(7187): 624-8.

Gray, M. W. (1998). "Rickettsia, typhus and the mitochondrial connection." Nature **396**(6707): 109-10.

Gregersen, N., P. Bross, et al. (2004). "Genetic defects in fatty acid beta-oxidation and acyl-CoA dehydrogenases. Molecular pathogenesis and genotype-phenotype relationships." Eur J Biochem **271**(3): 470-82.

Guda, C., E. Fahy, et al. (2004). "MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins." Bioinformatics **20**(11): 1785-94.

Guda, C. and S. Subramaniam (2005). "pTARGET [corrected] a new method for predicting protein subcellular localization in eukaryotes." Bioinformatics **21**(21): 3963-9.

Hatzigeorgiou, A. G., P. Fiziev, et al. (2001). "DIANA-EST: a statistical analysis." Bioinformatics **17**(10): 913-9.

Heazlewood, J. L., K. A. Howell, et al. (2003). "Towards an analysis of the rice mitochondrial proteome." Plant Physiol **132**(1): 230-42.

Heazlewood, J. L. and A. H. Millar (2005). "AMPDB: the Arabidopsis Mitochondrial Protein Database." Nucleic Acids Res **33**(Database issue): D605-10.

Heazlewood, J. L., J. S. Tonti-Filippini, et al. (2004). "Experimental analysis of the *Arabidopsis* mitochondrial proteome highlights signaling and regulatory components, provides assessment of targeting prediction programs, and indicates plant-specific mitochondrial proteins." Plant Cell **16**(1): 241-56.

Henze, K. and W. Martin (2003). "Evolutionary biology: essence of mitochondria." Nature **426**(6963): 127-8.

Hoglund, A., P. Donnes, et al. (2006). "MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition." Bioinformatics **22**(10): 1158-65.

Horton, P. and K. Nakai (1997). "Better prediction of protein cellular localization sites with the k nearest neighbors classifier." Proc Int Conf Intell Syst Mol Biol **5**: 147-52.
Horton, P., K. J. Park, et al. (2007). "WoLF PSORT: protein localization predictor." Nucleic Acids Res **35**(Web Server issue): W585-7.
Hua, S. and Z. Sun (2001). "Support vector machine approach for protein subcellular localization prediction." Bioinformatics **17**(8): 721-8.
Huang, W. L., C. W. Tung, et al. (2008). "ProLoc-GO: utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization." BMC Bioinformatics **9**: 80.
Huang, Y. and Y. Li (2004). "Prediction of protein subcellular locations using fuzzy k-NN method." Bioinformatics **20**(1): 21-8.
Huh, W. K., J. V. Falvo, et al. (2003). "Global analysis of protein localization in budding yeast." Nature **425**(6959): 686-91.
Hunt, M. C., S. Greene, et al. (2007). "Alternative exon usage selectively determines both tissue distribution and subcellular localization of the acyl-CoA thioesterase 7 gene products." Cell Mol Life Sci **64**(12): 1558-70.
Ichishita, R., K. Tanaka, et al. (2008). "An RNAi screen for mitochondrial proteins required to maintain the morphology of the organelle in *Caenorhabditis elegans*." J Biochem **143**(4): 449-54.
Iseli, C., C. V. Jongeneel, et al. (1999). "ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences." Proc Int Conf Intell Syst Mol Biol: 138-48.
Johnson, D. T., R. A. Harris, et al. (2007). "Tissue heterogeneity of the mammalian mitochondrial proteome." Am J Physiol Cell Physiol **292**(2): C689-97.
Karev, G. P., Y. I. Wolf, et al. (2003). "Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve?" Bioinformatics **19**(15): 1889-900.
Kennedy, E. P. and A. L. Lehninger (1949). "Oxidation of fatty acids and tricarboxylic acid cycle intermediates by isolated rat liver mitochondria." J Biol Chem **179**(2): 957-72.
Koene, S. and J. Smeitink (2009). "Mitochondrial medicine: entering the era of treatment." J Intern Med **265**(2): 193-209.
Kompare, M. and W. B. Rizzo (2008). "Mitochondrial fatty-acid oxidation disorders." Semin Pediatr Neurol **15**(3): 140-9.
Kondrashov, F. A. and E. V. Koonin (2003). "Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences." Trends Genet **19**(3): 115-9.
Koski, L. B., M. W. Gray, et al. (2005). "AutoFACT: an automatic functional annotation and classification tool." BMC Bioinformatics **6**: 151.
Krogh, A. (2008). "What are artificial neural networks?" Nat Biotechnol **26**(2): 195-7.
Kumar, A., S. Agarwal, et al. (2002). "Subcellular localization of the yeast proteome." Genes Dev **16**(6): 707-19.

Lang, B. F., G. Burger, et al. (1997). "An ancestral mitochondrial DNA resembling a eubacterial genome in miniature." Nature **387**(6632): 493-7.

Lang, B. F., M. W. Gray, et al. (1999). "Mitochondrial genome evolution and the origin of eukaryotes." Annu Rev Genet **33**: 351-97.

Lartillot, N. and H. Philippe (2004). "A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process." Mol Biol Evol **21**(6): 1095-109.

Lascaris, R., H. J. Bussemaker, et al. (2003). "Hap4p overexpression in glucose-grown Saccharomyces cerevisiae induces cells to enter a novel metabolic state." Genome Biol **4**(1): R3.

Leipe, D. D., E. V. Koonin, et al. (2003). "Evolution and classification of P-loop kinases and related proteins." J Mol Biol **333**(4): 781-815.

Li, J., T. Cai, et al. (2009). "Proteomic analysis of mitochondria from *Caenorhabditis elegans*." Proteomics.

Li, L., C. J. Stoeckert, Jr., et al. (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." Genome Res **13**(9): 2178-89.

Lill, R. and U. Muhlenhoff (2005). "Iron-sulfur-protein biogenesis in eukaryotes." Trends Biochem Sci **30**(3): 133-41.

Liu, J., S. Kang, et al. (2007). "Meta-prediction of protein subcellular localization with reduced voting." Nucleic Acids Res **35**(15): e96.

Lu, Z. and L. Hunter (2005). "Go molecular function terms are predictive of subcellular localization." Pac Symp Biocomput: 151-61.

Lu, Z., D. Szafron, et al. (2004). "Predicting subcellular localization of proteins using machine-learned classifiers." Bioinformatics **20**(4): 547-56.

MacKay, D. J. C. (2003). Information theory, inference, and learning algorithms. Cambridge, U.K. ; New York, Cambridge University Press.

Maggio-Hall, L. A. and N. P. Keller (2004). "Mitochondrial beta-oxidation in *Aspergillus nidulans*." Mol Microbiol **54**(5): 1173-85.

Makarova, K. S. and E. V. Koonin (2003). "Filling a gap in the central metabolism of archaea: prediction of a novel aconitase by comparative-genomic analysis." FEMS Microbiol Lett **227**(1): 17-23.

Martin-Kleiner, I., J. Gabrilovac, et al. (2006). "Leber's hereditary optic neuroretinopathy (LHON) associated with mitochondrial DNA point mutation G11778A in two Croatian families." Coll Antropol **30**(1): 171-4.

Maximo, V., J. Lima, et al. (2009). "Mitochondria and cancer." Virchows Arch **454**(5): 481-95.

McBride, H. M., M. Neuspiel, et al. (2006). "Mitochondria: more than just a powerhouse." Curr Biol **16**(14): R551-60.

Meisinger, C., A. Sickmann, et al. (2008). "The mitochondrial proteome: from inventory to function." Cell **134**(1): 22-4.

Mootha, V. K., J. Bunkenborg, et al. (2003). "Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria." Cell **115**(5): 629-40.

Nair, R. and B. Rost (2002). "Inferring sub-cellular localization through automated lexical analysis." Bioinformatics **18 Suppl 1**: S78-86.

Nair, R. and B. Rost (2005). "Mimicking cellular sorting improves prediction of subcellular localization." J Mol Biol **348**(1): 85-100.
Neuberger, G., S. Maurer-Stroh, et al. (2003). "Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence." J Mol Biol **328**(3): 581-92.
O'Brien, E. A., L. B. Koski, et al. (2007). "TBestDB: a taxonomically broad database of expressed sequence tags (ESTs)." Nucleic Acids Res **35**(Database issue): D445-51.
O'Rourke, T. W., N. A. Doudican, et al. (2005). "Differential involvement of the related DNA helicases Pif1p and Rrm3p in mtDNA point mutagenesis and stability." Gene **354**: 86-92.
Pagliarini, D. J., S. E. Calvo, et al. (2008). "A mitochondrial protein compendium elucidates complex I disease biology." Cell **134**(1): 112-23.
Park, K. J. and M. Kanehisa (2003). "Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs." Bioinformatics **19**(13): 1656-63.
Parkinson, J. and M. Blaxter (2009). "Expressed sequence tags: an overview." Methods Mol Biol **533**: 1-12.
Pearson, W. R. (1990). "Rapid and sensitive sequence comparison with FASTP and FASTA." Methods Enzymol **183**: 63-98.
Pfanner, N., N. Wiedemann, et al. (2004). "Assembling the mitochondrial outer membrane." Nat Struct Mol Biol **11**(11): 1044-8.
Pierleoni, A., P. L. Martelli, et al. (2006). "BaCelLo: a balanced subcellular localization predictor." Bioinformatics **22**(14): e408-16.
Poirier, Y., V. D. Antonenkov, et al. (2006). "Peroxisomal beta-oxidation--a metabolic pathway with multiple functions." Biochim Biophys Acta **1763**(12): 1413-26.
Priddy, K. L. and P. E. Keller (2005). Artificial neural networks : an introduction. Bellingham, Wash. , SPIE Press.
Priller, J., C. R. Scherzer, et al. (1997). "Frataxin gene of Friedreich's ataxia is targeted to mitochondria." Ann Neurol **42**(2): 265-9.
Prokisch, H., C. Scharfe, et al. (2004). "Integrative analysis of the mitochondrial proteome in yeast." PLoS Biol **2**(6): e160.
Quinlan, J. R. (1993). C4.5 : programs for machine learning. San Mateo, Calif., Morgan Kaufmann Publishers.
Ralph, S. J. and J. Neuzil (2009). "Mitochondria as targets for cancer therapy." Mol Nutr Food Res **53**(1): 9-28.
Rehling, P., K. Model, et al. (2003). "Protein insertion into the mitochondrial inner membrane by a twin-pore translocase." Science **299**(5613): 1747-51.
Reinders, J. and A. Sickmann (2007). "Proteomics of yeast mitochondria." Methods Mol Biol **372**: 543-57.
Reinhardt, A. and T. Hubbard (1998). "Using neural networks for prediction of the subcellular location of proteins." Nucleic Acids Res **26**(9): 2230-6.
Remm, M., C. E. Storm, et al. (2001). "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons." J Mol Biol **314**(5): 1041-52.

Richly, E., P. F. Chinnery, et al. (2003). "Evolutionary diversification of mitochondrial proteomes: implications for human disease." Trends Genet **19**(7): 356-62.
Ritov, V. B., E. V. Menshikova, et al. (2005). "Deficiency of subsarcolemmal mitochondria in obesity and type 2 diabetes." Diabetes **54**(1): 8-14.
Rodriguez-Ezpeleta, N., H. Brinkmann, et al. (2007). "Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans." Curr Biol **17**(16): 1420-5.
Rogozin, I. B., L. Aravind, et al. (2003). "Differential action of natural selection on the N and C-terminal domains of 2'-5' oligoadenylate synthetases and the potential nuclease function of the C-terminal domain." J Mol Biol **326**(5): 1449-61.
Sarda, D., G. H. Chua, et al. (2005). "pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties." BMC Bioinformatics **6**: 152.
Sardiello, M., F. Licciulli, et al. (2003). "MitoDrome: a database of *Drosophila melanogaster* nuclear genes encoding proteins targeted to the mitochondrion." Nucleic Acids Res **31**(1): 322-4.
Schaefer, A. M., R. W. Taylor, et al. (2004). "The epidemiology of mitochondrial disorders--past, present and future." Biochim Biophys Acta **1659**(2-3): 115-20.
Schapira, A. H. (2008). "Mitochondria in the aetiology and pathogenesis of Parkinson's disease." Lancet Neurol **7**(1): 97-109.
Scheffler, I. E. (2008). Mitochondria. Hoboken, N.J., Wiley-Liss.
Scott, M. S., D. Y. Thomas, et al. (2004). "Predicting subcellular localization via protein motif co-occurrence." Genome Res **14**(10A): 1957-66.
Seibel, N. M., J. Eljouni, et al. (2007). "Nuclear localization of enhanced green fluorescent protein homomultimers." Anal Biochem **368**(1): 95-9.
Shakhnarovich, G., T. Darrell, et al. (2005). Nearest-neighbor methods in learning and vision, MIT Press.
Shatkay, H., A. Hoglund, et al. (2007). "SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data." Bioinformatics.
Shen, Y. Q. and G. Burger (2007). "'Unite and conquer': enhanced prediction of protein subcellular localization by integrating multiple specialized tools." BMC Bioinformatics **8**: 420.
Shen, Y. Q. and G. Burger (2009). "Plasticity of a key metabolic pathway in fungi." Funct Integr Genomics **9**(2): 145-51.
Sickmann, A., J. Reinders, et al. (2003). "The proteome of *Saccharomyces cerevisiae* mitochondria." Proc Natl Acad Sci U S A **100**(23): 13207-12.
Small, I., N. Peeters, et al. (2004). "Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences." Proteomics **4**(6): 1581-90.
Smeitink, J., L. van den Heuvel, et al. (2001). "The genetics and pathology of oxidative phosphorylation." Nat Rev Genet **2**(5): 342-52.
Smith, D. G., R. M. Gawryluk, et al. (2007). "Exploring the mitochondrial proteome of the ciliate protozoon *Tetrahymena thermophila*: direct analysis by tandem mass spectrometry." J Mol Biol **374**(3): 837-63.

Stamatakis, A. (2006). "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models." Bioinformatics **22**(21): 2688-90.

Steinmetz, L. M., C. Scharfe, et al. (2002). "Systematic screen for human disease genes in yeast." Nat Genet **31**(4): 400-4.

Storm, C. E. and E. L. Sonnhammer (2002). "Automated ortholog inference from phylogenetic trees and calculation of orthology reliability." Bioinformatics **18**(1): 92-9.

Taylor, S. W., E. Fahy, et al. (2003). "Characterization of the human heart mitochondrial proteome." Nat Biotechnol **21**(3): 281-6.

Ueyama, T., K. Lekstrom, et al. (2007). "Subcellular localization and function of alternatively spliced Noxo1 isoforms." Free Radic Biol Med **42**(2): 180-90.

Uusimaa, J., J. S. Moilanen, et al. (2007). "Prevalence, segregation, and phenotype of the mitochondrial DNA 3243A>G mutation in children." Ann Neurol **62**(3): 278-87.

Valentini, G. and F. Masulli (2002). Ensembles of learning machines. Italian workshop on neural nets No13, Vietri sul Mare , ITALIE, Springer, Berlin, ALLEMAGNE

Wanders, R. J., P. Vreken, et al. (2001). "Peroxisomal fatty acid alpha- and beta-oxidation in humans: enzymology, peroxisomal metabolite transporters and peroxisomal diseases." Biochem Soc Trans **29**(Pt 2): 250-67.

Wang, Z. Y., D. M. Soanes, et al. (2007). "Functional analysis of lipid metabolism in *Magnaporthe grisea* reveals a requirement for peroxisomal fatty acid beta-oxidation during appressorium-mediated plant infection." Mol Plant Microbe Interact **20**(5): 475-91.

Wicker, N., G. R. Perrin, et al. (2001). "Secator: a program for inferring protein subfamilies from phylogenetic trees." Mol Biol Evol **18**(8): 1435-41.

Wiedemann, N., V. Kozjak, et al. (2003). "Machinery for protein sorting and assembly in the mitochondrial outer membrane." Nature **424**(6948): 565-71.

Wiedemann, N., N. Pfanner, et al. (2006). "Chaperoning through the mitochondrial intermembrane space." Mol Cell **21**(2): 145-8.

Xie, D., A. Li, et al. (2005). "LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST." Nucleic Acids Res **33**(Web Server issue): W105-10.

Yu, C. S., Y. C. Chen, et al. (2006). "Prediction of protein subcellular localization." Proteins.

Supplementary information

Other publications

Shen YQ, O'Brien E, Koski L, Lang BF, Burger G (2009). Chapter "EST databases and Web tools for EST projects.", in Expressed sequence tags - generation and analysis, Methods in Molecular Biology Vol. 533 (J. Parkinson, ed.), Humana Press, Totowa, USA.

Ma LJ, Ibrahim AS, Skory C, Grabherr MG, Idnurm A, Corrochano LM, Elias M, Burger G, Lang BF, Abe A, Butler M, Calvo S, Engels R, Fu J, Hansberg W, Kim JM, Kodira CD, Koehrsen MJ, Liu B, Miranda-Saavedra D, Rodriguez-Romero J, O'Leary S, Ortiz-Castellanos L, Poulter R, Ruiz-Herrera J, **Shen YQ** et al (2009). "Genomic analysis of the basal lineage fungus *Rhizopus oryzae* reveals a whole-genome duplication." PLoS Genetics, 5(7): e1000549.