Introduction

The Microsporidia

Microsporidia is a group of obligate intracellular parasites that infect a broad range of species from vertebrates to invertebrates. They are now classified as sister group of fungi. Microsporidia species have genome sizes from 2,3MB to 19,5MB. Most of these parasites have a very small number of genes, from 2000 to 4000 genes, which are thought to be very essential that cannot be lost or they are specific genes of Microsporidia. The understanding of the evolution of Microsporidia is still unclear and there are many question about the reducing process in their genomes, such as when they lost their genes, if the losing process happened once or many times, if those genes lost forever or just temporarily...

In order to find the answer for those questions, the proteins sets of the Microsporidia last common ancestor (LCA) will be inferred using hierarchical orthology inference approach. The distribution of these proteins will be then analyzed to find their origins and to get an overview how common they are in the living organisms. Moreover, we try to investigate the functionality of these Microsporidia LCA proteins and identify their roles to understand how they can work together with their hosts' metabolic pathways.

The estimation of the Microsporidia last common ancestor protein set

Introduction

Methods

Starting from 11 extant Microsporidia species (table A‑1), we used OrthoMCL (Li et al. 2003) to search for homologous proteins in those 11 Microsporidia species. OrthoMCL performed all-against-all BLASTP comparisons for all input data set and clustered homologous groups using Markov Cluster algorithm MCL (van Dongen 2000).

The initial homologous groups were then extended by using HaMStR (Ebersberger et al. 2009) to search for orthologs in other 24 search taxa (table A‑2). The Hidden Markov Model (HMM) profiles for the initial homologous groups (seed sequences) crated by HaMStR were used to search in the search taxa. The obtained hits were confirmed by re-BLAST search against the protein sets of seed sequences. We took into account here also the co-orthologs for re-BLAST and we limited the HMM hits up to only the best 10 hits.

We identified a core gene set, where we found orthologs in all taxa and each taxon has exactly one orthologous protein. Using the core gene set, we reconstructed a maximum likelihood species tree for those 35 taxa. Firstly, we created a super-alignment by concatenating the alignments of all core genes. To exclude the alignment columns that are uninformative, we removed columns that have at least 50% gaps. We used ProtTest (Abascal et al. 2005) to find the best fitting model for the tree reconstruction procedure. With the best model parameters obtained from ProtTest, we used RAxML (Stamatakis 2014) to build the maximum likelihood species tree with 100 bootstrap replicates.

Using the principle of minimum evolution (Edwards 1996), we filtered the orthologous group to obtain the final protein set representing the Microsporidia last common ancestor. Those final orthologous groups have to have either (1) at least one ortholog from N.parisii (the earliest branch of the Microsporidia clade), or (2) at least two orthologs from Microsporidia species different than N.parisii and one or more orthologs from non-Microsporidia taxa.

Results

OrthoMCL gave 2904 initial homologous groups for 11 Microsporidia protein sets. Out of 2904 extended groups, we found 80 groups, where all 11 Microsporidia and 24 non-Microsporidia taxa are present and each taxon has one representative ortholog. Those 80 groups serve as out core get set for the species tree reconstruction. The super-alignment after de-gapping has a length from 36.616 amino acids. The best model obtained from ProtTest was LG substitution model (Le & Gascuel 2008), GAMMA distribution G , including proportion of invariable sites estimation I & empirical base frequencies F. The input model parameter for RAxML was PROTGAMMAILGF. The reconstructed species tree with bootstrap support values is shown in the figure A.1 below.

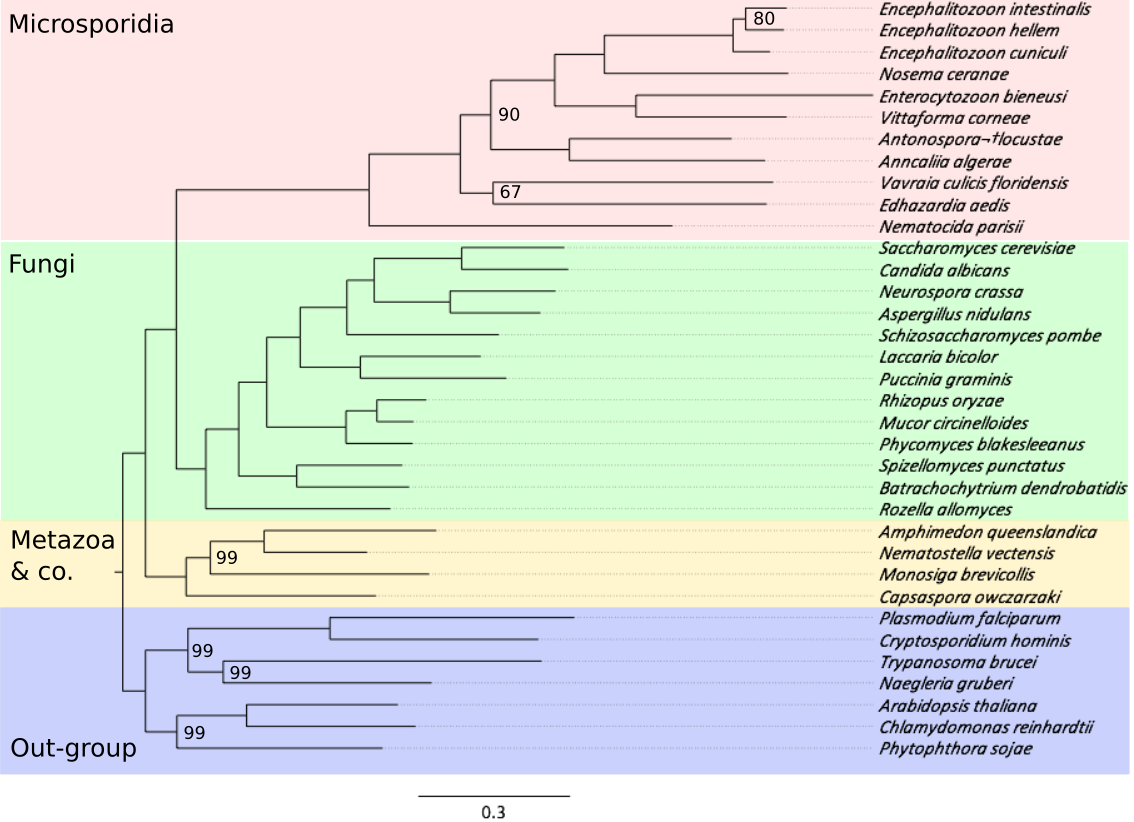
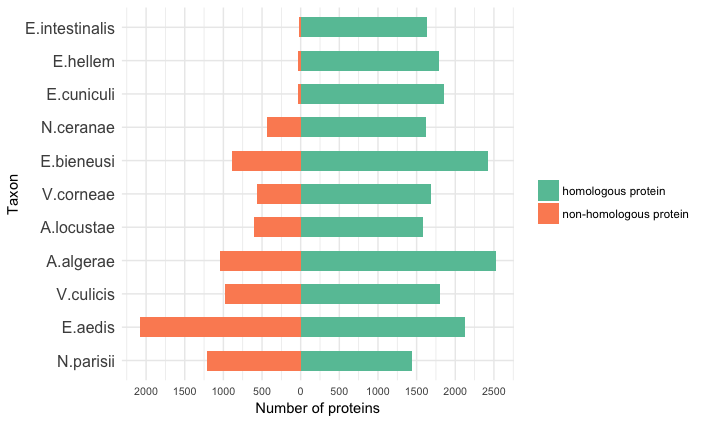


Figure A.1. Maximum likelihood tree over 35 species. The 11 Microsporidia taxa are highlighted in red. Other non-Microsporidia taxa include 13 Fungi (green), 2 Metazoa and M.brevicollis, C.owczarzaki (yellow) and 7 out-group species (purple). Node labels denote the bootstrap support and only values <100 are shown. The tree topology is congruent with the known species tree.

Filtered the HaMStR result that did not match the parsimony criteria, we got at the end 1605 final orthologous groups. They present the set of Microsporidia last common ancestor proteins.

Discussion



The orthology assignment result agreed with other studies about the fraction of microsporidia only proteins.

The reconstructed species tree shows a stable signal for placing the Microsporidia group as the earliest clade of the fungi, which are supported by the recent phylogenetic study of the taxonomy of Microsporidia that hypothesis that Microsporidia is the sister clade of fungi. It can therefore serve as a fundamental phylogenetic background for filtering the orthology assignment and estimating the last common ancestor protein set of Microsporidia.

Conclusion

Distribution analysis of the Microsporidia last common ancestor proteins

Introduction

Methods

Results

Discussion

Conclusion

Functional annotation

Introduction

Methods

Results

Discussion

Conclusion

Metabolic pathway analysis

Introduction

Methods

Results

Discussion

Conclusion

# Appendix

Tables

Table A‑1: Eleven extant Microsporidia species

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Name | Source | Strain | Number of proteins |
| 1 | *Encephalitozoon hellem* | JGI | ATCC 50504 | 1827 |
| 2 | *Encephalitozoon intestinalis* | Broad Inst | ATCC 50506 | 1657 |
| 3 | *Encephalitozoon cuniculi* | Broad Inst | GB-M1 | 1896 |
| 4 | *Nosema ceranae* | Broad Inst | BRL01 | 2057 |
| 5 | *Enterocytozoon bieneusi* | JGI | H348 | 3312 |
| 6 | *Vittaforma corneae* | Broad Inst | ATCC 50505 | 2243 |
| 7 | *Anncaliia algerae* | Broad Inst | PRA339 | 3576 |
| 8 | *Antonospora locustae* | JGI | HM-2013 | 2191 |
| 9 | *Edhazardia aedis* | Broad Inst | USNM 41457 | 4208 |
| 10 | *Vavraia culicis* subsp. floridensis | Broad Inst |  | 2775 |
| 11 | *Nematocida parisii* | Broad Inst | ERTm1 | 2659 |

Table A‑2. 24 taxa used for extent the initial homologous groups

|  |  |  |  |
| --- | --- | --- | --- |
| No. | Name | No. | Name |
| 1 | *S.pombe* | 14 | *M.brevicollis* |
| 2 | *C.albicans* | 15 | *N.vectensis* |
| 3 | *S.cerevisiae* | 16 | *A.queenslandica* |
| 4 | *N.crassa* | 17 | *C.owczarzaki* |
| 5 | *A.nidulans* | 18 | *T.brucei* |
| 6 | *L.bicolor* | 19 | *N.gruberi* |
| 7 | *P.graminis* | 20 | *A.thaliana* |
| 8 | *M.circinelloides* | 21 | *C.reinharditii* |
| 9 | *R.oryzae* | 22 | *P.sojae* |
| 10 | *P.blakesleeanus* | 23 | *C.hominis* |
| 11 | *B.dendrobatidis* | 24 | *P.falciparum* |
| 12 | *S.punctatus* |  |  |
| 13 | *R.allomyces* |  |  |