(Deutsche Version unten)

Microsporidia are a group of parasites that infect a wide range of species, many of which play important roles in agriculture and human disease. At least 14 microsporidian species have been confirmed to cause dangerous infectious diseases in both immunocompromised and immunocompetent humans. Approximately 1,400 species of microsporidia have been described. Depending on their host and habitat they are being classified into three groups, the aquasporidia, the terresporidia and the marinosporidia.

Microsporidia were originally classified as fungi by Naegeli (1857). However, their lack of typical eukaryotic components – such as mitochondria, Golgi bodies or peroxisomes – suggested to place the microsporidia together with other amitochondriate protists within the Archezoa kingdom. This "microsporidia-early" hypothesis was further supported by some molecular phylogenies. Despite this evidence, the placement of microsporidia as an early branching eukaryote remains a topic for debate for two reasons: Firstly, the phylogeny of microsporidia is prone to suffer from biases in their reconstruction. The high evolutionary rate of microsporidian proteins tends to place these proteins together with other fast evolving lineages, a phenomenon known as long-branch attraction. Furthermore, the presence of the mitochondrial heat shock protein hsp70 suggests the ancestral presence of mitochondria in the microsporidia. This invalidates the main morphological argument for the "microsporidia-early" hypothesis. In 1996, the first molecular phylogenetic studies placed the microsporidia inside the fungi. Subsequently, several further studies located the microsporidia at different positions inside the fungal clade. Since then, microsporidia have been viewed as being members of the Ascomycota, Zygomycota, Crytomycota, or as being a sister group to the Ascomycota and Basidiomycota or even to all fungi.

These difficulties in determining the origin of microsporidia are not only caused by their lack of several cellular components but also by their reduced genomes and metabolism. Being obligate intracellular parasites, microsporidia successfully reduced their genome sizes, down to the range of bacteria. As the smallest eukaryotic genome described so far, the genome of *Encephalitozoon intestinalis* is just 2.3 Mbp, about half the size of the one of *Escherichia coli*. Due to their low number of protein coding genes (less than 4,000), microsporidia are thought to retain only genes essential for their survival and development. Furthermore, several key metabolic pathways are missing in the microsporidia, such as the citric acid cycle, oxidative phosphorylation, or the *de novo* biosynthesis of nucleotides. As a result they are in an obligatory dependence on many primary metabolites from the hosts. However, the functionally uncharacterized hsp70 protein suggests a more complex genome of the microsporidian ancestor. Consequently, the reduction of microsporidian genomes and metabolic pathways would be due to secondary loss processes.

Microsporidia are becoming model organisms for studying minimal eukaryotic genomes and obligate endoparasitic lifestyles in eukaryote. Moreover, a deeper understanding of the physiology of microsporidia is required to better manage these pathogens. Despite this, there is little knowledge regarding the evolution of the microsporidia and their metabolisms as different factors make their study challenging. Their obligate endoparasitic lifestyle hinders the study of microsporidia isolated from the host's cells. Furthermore, no techniques for genetic modifications are available for studying the development stage of microsporidia inside the host's cell. Comprehensive comparative analyses of the microsporidia are thus needed.

We investigated the evolutionary history of the contemporary microsporidia though the reconstruction and investigation of their last common ancestor (LCA). To enable the intuitive data analysis of the microsporidian phylogenetic profiles, we have developed PhyloProfile, a program that facilitates a dynamic visualization and exploration of the presence-absence patterns of microsporidian genes across several taxa in the tree of life, integrated with other additional information layers. PhyloProfile closes the methodological gap that existed between tools to generate large phylogenetic profiles to delineate the evolutionary history and the contemporary distribution of large – and ultimately complete – gene sets, and the more function-oriented analysis of individual protein. Additionally, we present HamFAS as a new approach for the reliable functional annotation of proteins. HamFAS integrates a targeted ortholog search based on the HaMStR algorithm with a weighted assessment of feature architecture similarities (FAS) between orthologs. In brief, for a seed protein we identify orthologs in reference species in which proteins have been functionally annotated based on manually curated assignments to KEGG Ortholog (KO) groups. The FAS scores between the orthologs and seed proteins are calculated. Subsequently, we compute pairwise FAS scores for all reference proteins within a KO group. A group's mean FAS score serves then as cutoff that must be exceeded to warrant transfer of its KO identifier to the seed. A benchmark using a manually curated yeast protein set showed that HamFAS yields the best precision (98.5%) when compared with two state-of-the-art annotation tools provided by KEGG, KAAS and BlastKOALA. Furthermore, the ability of inferring orthologs between distantly related species using hidden Markov model profiles allows HamFAS to achieve a higher sensitivity. On average HamFAS annotates 47.5% more proteins than KAAS or BlastKOALA.

We furthermore reconstruct a robust phylogeny of microsporidia to identify the placement of the microsporidia in the eukaryotic tree of life using the data available from eleven microsporidian genome sequencing projects at the Broad Institute and the Joint Genome Institute. We successfully identified a microsporidian core set comprising of 80 one-to-one orthologous groups, which were used as a basic for the phylogeny reconstruction. Our reconstructed phylogenetic tree from a diverse set of taxa, including 48 fungi representing the full fungal diversity as well as other representative species for both unikonts and bikonts, strongly supported the hypothesis that microsporidia form the early branch of fungal clade. The analyzed data explained this microsporidia-fungi relationship significantly better than any other hypotheses.

Our phylogenetic profile study clarified the dynamic evolutionary history of microsporidian genomes. Between 2% of the proteins in the compact microsporidia *Encephalitozoon intestinalis* and up to 49% of the proteins of *Edhazardia aedis* are private for individual microsporidian species. Those private proteins are useful to investigate adaptations of each microsporidium to their host. On the other hand, proteins that are found in all microsporidia are expected to be essential and evolutionary old. We found that 94% of the microsporidian LCA proteins could be tracked back to the last eukaryotic common ancestor and further 3% of the LCA proteins were shared in the common ancestor with fungi. Only 3% of the LCA proteins are microsporidia specific, which are potentially important for the obligate parasitic lifestyle of the microsporidia.

The functional annotation and metabolic pathway analysis of the microsporidian LCA protein set gave us more insight into the adaptation of the microsporidia to their parasitic lifestyle and the origin of the microsporidian genome reduction. The presence of E1 and E3 components of the pyruvate dehydrogenase complex and the mitochondrial hsp70 protein support an ancestral presence of mitochondria in the ancestral microsporidia. In addition, several ancient proteins that complement gapped metabolic pathways were found in the microsporidian LCA. They suggested a more complex genome and metabolism in the LCA. However, the microsporidian LCA still lacked many main pathways such as the TCA cycle for effective energy production, and key enzymes that are required for *in vivo* synthesis of critical metabolites like purines and pyrimidines. In summary, we find that the parasitic lifestyle and the genome reduction already occurred in the microsporidian LCA. This ancestral state was followed by further losses and gains during the evolution of each individual microsporidian lineage.

Mikrosporidien sind sporenbildende Parasiten, die verschiedene Organismen infizieren. Die Mikrosporidiose beeinträchtigt nicht nur die Agrarökonomie, sondern führt auch zu humanmedizinischen Krankheitsbildern. Es sind mindestens 14 Mikrosporidien bekannt, welche gefährliche infektiöse Krankheiten sowohl bei immunkompromittierten als auch bei immunkompetenten Menschen verursachen. Etwa 1.400 Mikrosporidien wurden bislang beschrieben. Nach ihren Wirten und Habitaten können sie in drei Gruppen eingeteilt werden, die Aquasporidien, Terresporidien und Marinosporidien.

Mikrosporidien wurden erstmals 1857 von Naegeli als Pilz klassifiziert. Wegen ihres Mangels vieler typischen eukaryotischen Komponenten – wie Mitochondrien, Golgi-Apparat oder Peroxisomen – wurden die Mikrosporidien später allerdings zusammen mit anderen amitochondrischen Protisten innerhalb des Archezoa-Reiches gruppiert. Diese "Microsporidia-early" Hypothese wurde darüber hinaus durch einige molekulare Phylogenien unterstützt. Trotz dieser Evidenzen wurde die phylogenetische Platzierung der Mikrosporidien in Frage gestellt. Zum einem werden die Phylogenie von Mikrosporidien durch die Rekonstruktionsartefakte verzerrt. Durch die hohe Evolutionsrate der mikrosporidischen Proteine gruppieren sie häufig zusammen mit anderen schnell evolvierenden Proteinen (long-branch-attraction). Darüber hinaus weist das mitochondriale Hitzeschockprotein hsp70 darauf hin, dass die frühen Mikrosporidien Mitochondrien besaßen. Dies wiederspricht dem morphologischen Hauptargument für die "Microsporidia-early" Hypothese. Im Jahr 1996 wurde die Verwandtschaft zwischen Pilzen und Mikrosporidien erstmals durch molekulare phylogenetische Studien unterstützt. Verschiedene Studien, basierend auf einzelnen und mehreren Genen, positionieren die Mikrosporidien unterschiedlich in dem Stammbaum der Pilze. Dabei werden die Mikrosporidien entweder innerhalb der Ascomycota, Zygomycota oder Crytomycota positioniert, oder als Schwestergruppe der Ascomycota und Basidiomycota, oder auch aller Pilze.

So bleibt die exakte Position der Mikrosporidien im Speziesbaum der Pilze immer noch ungelöst. Dabei wird die Bestimmung des Ursprungs der Mikrosporidien durch deren reduzierten Genome weiter erschwert. Als obligate intrazelluläre Parasiten verminderten die Mikrosporidien ihre Genome soweit das deren Größe im Bereich bakterieller Genome liegt. Das kleinste beschriebene eukaryotische Genom von *Encephalitozoon intestinalis* ist mit 2,3 Mbp etwa halb so groß wie das von *Escherichia coli*. Die geringe Anzahl von protein-kodierenden Genen (weniger als 4.000) deutet darauf hin, dass die Genome der Mikrosporidien nur Gene enthalten die für ihr Überleben und ihre Entwicklung essentiell sind. Darüber hinaus fehlen den Mikrosporidien mehrere Stoffwechselwege, wie der Zitronensäurezyklus, die oxidative Phosphorylierung oder die *de novo* Biosynthese von Nukleotiden. Dies führt zu einer obligaten Abhängigkeit vom Wirt für viele primäre Metabolite. Das uncharakterisierte hsp70-Protein setzt jedoch ein komplexeres Genom des mikrosporidischen Vorfahren voraus. Folglich ist die Reduktion der mikrosporidischen Genome und Stoffwechselwege auf Sekundärprozesse zurückzuführen.

Mikrosporidien können als Modellorganismen für die Untersuchung eines minimalen eukaryotischen Genoms und obligater endoparasitischen Lebensweisen in Eukaryoten dienen. Darüber hinaus ist ein tieferes Verständnis der mikrosporidischen Physiologie erforderlich, um es als Pathogen besser zu bewältigen. Dennoch gibt es bislang nur wenig Studien über die mikrosporidische Evolution und ihren Metabolismus. Zum einen behindert die obligate endoparasitische Lebensweise die Untersuchung von Mikrosporidien. Darüber hinaus steht für Mikrosporidien keine Technik zur genetischen Modifikation zur Verfügung. Deshalb bedarf es einer umfassenden vergleichenden Analyse der Mikrosporidien.

Wir untersuchten daher die Entwicklungsgeschichte der heutigen Mikrosporidien durch die Rekonstruktion und Untersuchung ihres letzten gemeinsamen Vorfahrens (LCA). Für eine intuitive Datenanalyse der phylogenetischen Profile der Mikrosporidien entwickelten wir PhyloProfile, ein Programm für dynamische Visualisierung und Untersuchung von phylogenetischen Profilen. Durch die Verbindung mit anderen, zusätzlichen Informationsebenen – die in das An- und Abwesenheitsmuster der mikrosporidischen Genen integriert werden – erlaubt PhyloProfile eine Analyse auch in großen Stammbäumen. PhyloProfile schließt damit eine methodologische Lücke um die evolutionäre Geschichte und die gegenwärtige Verteilung großer – auch vollständiger – Gensets und die funktionalere Analyse einzelner Proteine zu beschreiben.

Außerdem präsentieren wir HamFAS, einen neuen Ansatz zur zuverlässigen funktionellen Annotation von Proteinen. HamFAS integriert eine gezielte Orthologen-Suche basierend auf dem HaMStR-Algorithmus mit einer gewichteten Bewertung von Feature Architektur Ähnlichkeiten (FAS) zwischen Orthologen. Für ein Seed-Protein identifizieren wir Orthologe in Referenzspezies, deren Proteine bereits durch eine manuelle Annotation in KEGG-Ortholog (KO)-Gruppen eingeordnet wurden. Zwischen den Orthologen und den Seed-Proteinen werden die FAS-Werte berechnet. Anschließend berechnen wir paarweise FAS-Werte für alle Referenzproteine innerhalb einer KO-Gruppe. Der mittlere FAS-Wert einer Gruppe dient dann als Cutoff, der überschritten werden muss, um die Übertragung seines KO-Identifikation an den Seed zu rechtfertigen. Wir benchmarkten die Performance von HamFAS mit einem manuell kuratierten, KO-annotierten, Hefeprotein-Set. HamFAS erzielte die beste Genauigkeit (98,5%) im Vergleich zu zwei State-of-the-Art Annotationstools von KEGG, KAAS und BlastKOALA. Darüber hinaus zeigte HamFAS eine höhere Sensitivität, da es dank der Verwendung von Hidden Markov Modellen auch Orthologe zwischen weiter entfernt verwandten Arten bestimmen konnte. Hier findet HamFAS im Durchschnitt 47,5% mehr Proteine von HamFAS als KAAS oder BlastKOALA.

Um die Mikrosporidien besser im eukaryotischen Stammbaum zu platzieren erstellten wir eine robuste Phylogenie der Mikrosporidien. Um dies zu erreichen, verwendeten wir die Daten von elf Mikrosporidien-Genom-Sequenzierungsprojekten des Broad Institute und des Joint Genome Institute. Wir identifizierten einen mikrosporidisches Geneset, welches aus 80 eins-zu-eins-Orthologen Gruppen besteht. Diese Gruppen dienten als Grundlage für die Phylogenie-Rekonstruktion. Unser rekonstruierter phylogenetischer Baum beinhaltet 48 Pilze, die die gesamte Pilzdiversität abbilden, sowie andere repräsentative Arten der Unikonten und Bikonten. Der resultierende Baum unterstützte die Hypothese, dass Mikrosporidien den frühesten Zweig der Pilzevolution bilden. Die analysierten Daten erklärten diese Mikrosporidien-Pilz-Verwandtschaft signifikant besser als alle anderen Hypothesen.

Unsere Studie der phylogenetischen Profile zeigte die dynamische Evolutionsgeschichte der mikrosporidischen Genomen. Je nach Mikrosporidium finden wir das zwischen 2% der Proteine in der kompakten Mikrosporidie *Encephalitozoon intestinalis* bis hin zu 49% der Proteine im Fall von *Edhazardia adis* nur in einer Art gefunden werden. Diese privaten Proteine deuten darauf hin, dass sie für die Wirtsanpassung benötigt werden. Im Gegensatz dazu stehen jene Proteine die in anderen Mikrosporidien gefunden werden und als essentiell und evolutionär alt angesehen werden. Wir finden, dass 94% der Proteinen des mikrosporidischen LCAs auf den letzten gemeinsamen Vorfahren aller Eukaryoten zurückverfolgt werden können und dass 3% der Proteine, mit dem gemeinsamen Vorfahren der Pilze geteilt werden. Nur 3% LCA-Proteine sind Mikrosporidien-spezifisch. Diese sind potentiell für die obligatorische parasitäre Lebensweise von Mikrosporidien zentral.

Die funktionelle Annotation und die Analyse der Stoffwechselwege des mikrosporidischen LCAs ermöglichte ein besseres Verständnis der Anpassung von Mikrosporidien an ihre parasitäre Lebensweise und den Ursprung ihrer Genom-Reduktion. Die Anwesenheit von E1, E3-Komponenten des Pyruvat-Dehydrogenase-Komplexes und des mitochondrialen hsp70-Proteins deuten darauf hin, dass die anzestralen Mikrosporidien Mitochondrien besaßen. Zusätzlich wurden mehrere neue Proteine im mikrosporidischen LCA gefunden, die einige Lücken Stoffwechselwege schließen können. Dies deutet auf ein komplexeres Genom und einen aufwändigeren Metabolismus im LCA hin als bislang vermutet. Dem mikrosporidischen LCA fehlen jedoch weiterhin viele primäre Stoffwechselwege, wie der Citratzyklus, oder Schlüsselenzyme, die für die in vivo Synthese von kritischen Metaboliten wie Purinen und Pyrimidinen benötigt werden. Zusammenfassend nehmen wir an, dass die parasitische Lebensweise bereits in der mikrosporidischen LCA vorkam. Die reduzierten Genome sind damit der anzestrale Zustand für die Mikrosporidien, welchem weitere Genverluste und Genzuwächse auf einzelnen mikrosporidischen Linien folgte.