UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 2191

Resolving deep nodes of eukaryote phylogeny

CAESAR AL JEWARI

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2022

ISSN 1651-6214
ISBN 978-91-513-1599-7
URN urn:nbn:se:uu:diva-484580

Dissertation presented at Uppsala University to be publicly examined in Lindahlsalen, Evolutionsbiologiskt centrum, Norbyv. 18D, Uppsala, Wednesday, 2 November 2022 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Antonis Rokas (Department of Biological Sciences, Vanderbilt University).

Abstract
Al Jewari, C. 2022. Resolving deep nodes of eukaryote phylogeny. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 2191. 53 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-1599-7.

My thesis aims to solve deep nodes in the eukaryote tree of life (eToL), by developing new data sets and new approaches to analysing them. In **paper I**, I described a dataset of 76 universal eukaryotic proteins of bacterial descent (euBacs), in order to test the relations among the three main divisions of mitochondriate eukaryotes (Amorphea, Diaphretickes and Discoba). I developed two protocols to identify problematic data. The conJac protocol analyzes data by jackknifing to detect outlier sequences, while conWin uses a sliding window to find sequence fragments of potentially foreign origin. Phylogenetic analyses of the 76 euBacs, with and without conWin or conJac filtering place Discoba as the sister group to Amorphea and Diaphretickes. The results are largely consistent and highly supported under various evolutionary models except for highly complex CAT models. In **paper II**, I describe a dataset of 198 universal eukaryote proteins of archaeal ancestry (euArcs), which includes the remaining eukaryotes, informally referred to as amitochondriate excavate. These were excluded from the previous study because they lack euBacs. Phylogenetic analyses of the euArc dataset place the amitochondriate excavate as the first three branches of eToL, followed by Discoba, the only mitochondriate excavates, which appear as a sister group to the remaining eukaryotes. I also developed a protocol using predicted protein structures to increase the fitness of the model without inflating the parameter space, allowing me to conduct a series of control analyses and further support the multi-excavate root. In **Paper III,** I describe a new application of reciprocal-rooting using concatenated sequences, which I then use to test the euArc root. I also developed two sampling protocols unique to this kind of data. The protocols confirm the multi-excavate euArc root, which indicates that eukaryotes arose from an excavate ancestor. **Paper IV** describes a follow-up on the ConWin results from **Paper I**. These show moderate to strong support for mosaicism in 16 euBac proteins from diverse metabolic pathways and donor lineages. In summary, this thesis presents a novel root for the eukaryote tree of life. The new root requires revision of fundamental theories of eukaryote evolution including the source and timing of mitochondrial origins. The methods I have developed are applicable to many different kinds of phylogenetic studies, and the new protein structure model should make these analyses faster, more flexible, and more widely available.

Keywords: Eukaryote Tree of Life, Excavata, phylogenetics, phylogenomics, Mitochondria

Caesar Al Jewari, Department of Organismal Biology, Systematic Biology, Norbyv. 18 D, Uppsala University, SE-75236 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-513-1599-7
URN urn:nbn:se:uu:diva-484580 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-484580)

To you and me

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I. Al Jewari, A., Baldauf, S. L. (2022) Conflict over the eukaryote root resides in strong outliers, mosaics and missing data sensitivity of site-specific (CAT) mixture models. *Systematic Biology*, syac029

II. Al Jewari, A., Baldauf, S. L. (2022) An excavate root for the eukaryote tree of life. *Manuscript*

III. Al Jewari, A., Baldauf, S. L. (2022) Reciprocal rooting with concatenation supports the multi-excavate root. *Manuscript*

IV. Baldauf, S. L, Al Jewari, C. (2022) Widespread mosaicism in eukaryotic genes of bacterial origin. *Manuscript*

Reprints were made with permission from the respective publishers.

Contents

Abbreviations

eToL	Eukaryotic Tree of Life
LBA	Long Branch Attraction
HGT	Horizontal Gene Transfer
euBacs	Eukaryotic loci (genes/proteins) of bacterial ancestry
euArcs	Eukaryotic loci (genes/proteins) of archaeal ancestry
UB	unikont-bikont
AM	amitochondriate
LECA	Last Eukaryotic Common Ancestor
FECA	First Eukaryotic Common Ancestor
SGT	Single Gene Trees / Single Protein Trees
PIKs	Potentially Incongruent Kernels
mlBP	maximum likelihood bootstrap percentage
SAR	Stramenopila + Alveolata + Rhizaria

Introduction

Phylogenetic reconstruction

Phylogenetics is the attempt to reconstruct evolutionary history, for the most part using present-day data. After a rocky early start, phylogenetic analysis of molecular sequence data now dominates the field. The sub-discipline of organismal phylogeny is now dominated by analyses of multisequence data, usually concatenated into a single supermatrix and sometimes referred to as phylogenomics. Phylogenetic and phylogenomic analyses have radically revised our understanding of the relationships among organisms at all taxonomic levels, allowing us to recover what appears to be a reasonably accurate picture of the universal tree of life. This is especially the case for eukaryotes, whose evolution appears to have largely been by vertical descent rather than horizontal exchange between unrelated species.

Molecular phylogeny has also turned out to be surprisingly complex. This is particularly true for the deeper branches in the tree of life, where repeated over-writing of gene sequences obscures their true history. Various evolutionary models have been developed to try to more accurately capture this history. In addition, new methods are continually being devised to better identify the most misleading bits of data in order to remove them or to identify the most appropriate data to use for a particularly phylogenetic question. My thesis research is aimed at developing better methods for the detection of misleading data, better models to speed the analyses and allow more in-depth study of individual data sets, and a new approach to one of the most challenging phylogenetic problems, rooting the tree. I have then tested these methods on one of the central questions in eukaryote evolution, the root of the eukaryote tree of life.

The eukaryote tree of life (eToL)

Over the last twenty years, large phylogenomic (multi-gene) datasets have confidently and consistently identified the broad outlines of the eukaryote tree of life (eToL) (Baldauf et al., 2000; Bapteste et al., 2002; M. W. Brown et al., 2018; Burki et al., 2016, 2020; Strassert et al., 2019). As a result, a handful of major groups can be defined, which together encompass the bulk of known

eukaryotic diversity. This view comes with two caveats. One is that many of the major groups of eukaryotes, particularly those that consist largely or entirely of microbes, are underrepresented in these trees. Moreover, the bulk of microbial eukaryotic diversity remains to be discovered and described (Caron & Hu, 2019). Second, the confidence in eukaryote phylogenomics is based on the assumption that horizontal gene transfer (HGT) is rare (Martin, 2018; Sibbald et al., 2020). This is unlike bacteria, where HGT is widespread and no gene seems to be entirely immune (Kloesges et al., 2011). However, new studies are emerging suggesting that HGT is more common in eukaryotes than previously expected, and it is important to take HGT into account when reconstructing eukaryote phylogeny (He et al., 2016; Ke et al., 2000; Keeling & Palmer, 2008; Rice et al., 2013; Sibbald et al., 2020).

Current understanding of eToL places nearly all well characterized eukaryotes into three supergroups or "suprakingdoms": Amorphea, Diaphoretickes and Excavata (Figure 1). Amorphea includes animals (Holozoa), fungi (Holomycota or Nucletmycea) and amoebozoan amoebas (Amoebozoa), along with several less well-characterized lineages of uncertain affinity. Diaphoretickes includes Archaeplastida - red, green and glaucophyte algae plus land plants, Stramenopila – including chlorophyll a+c algae, oomycetes and diverse microbes, Alveolata – including mainly Ciliophora, Dinoflagellata and Apicomplexa, Rhizaria – including Foraminifera, Radiolaria and Cercozoa, cryptophyte (Cryptista) and haptophyte (Haptista) algae, plus a number of less-well characterized lineages within each of these groups. Finally, Excavata includes Discoba and Metamonada and possibly also Malawimonadida (Figure 1) (Adl et al., 2019).

The monophyly of Amorphea and Diaphoreticks is well established, as it is strongly and consistently supported by numerous phylogenomic studies with various gene-taxon sets and phylogenetic methods (Adl et al., 2019; Burki et al., 2007; Burki, 2014; Cavalier-Smith, 2018). However, the monophyly of Excavata is based only on a single, if complex morphological character, the presence of an excavated feeding groove (Simpson, 2003). Within Excavata only the monophyly of Discoba (Jakobida, Euglenozoa, Heterolobosea) is well-established (Figure 1) (Kamikawa et al., 2014). Unlike the other excavates, Discoba are largely free-living and have functional, if diverse mitochondria. In contrast, Metamonada are universally micro- or anaerobic, and the best-known species are mostly parasites. Consistent with this, Metamonads have highly-reduced mitochondria-like organelles (hydrogenosomes or mitosomes) that lack mitochondrial DNA and nearly all nuclear-encoded mitochondrial proteins. Since the monophyly of Metamonada is not well established, they will be referred to here by the informal designation of amitochondriate excavates (AM excavates, Figure 1). Malawimonads have an excavate morphology and aerobic mitochondria, but

may be just an odd branch of Amorphea (Figure 1) (Derelle et al., 2015; Heiss et al., 2018). This is all given the caveat that eukaryote diversity is still probably largely unknown. This is especially the case for eukaryotic microbes, which are the bulk of eukaryotes, and novel species continue to be described (e.g., Lax et al., 2018).

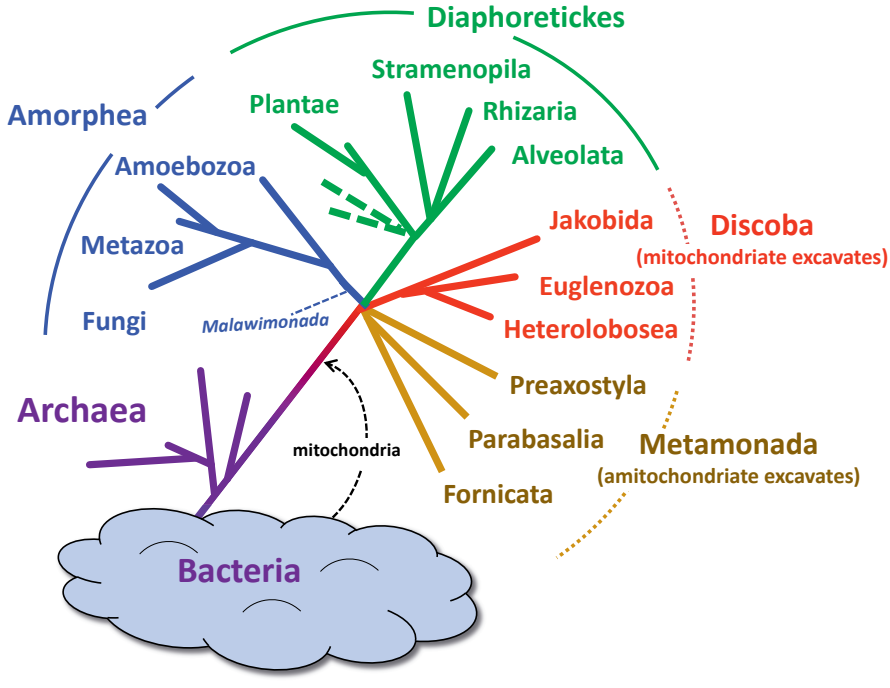

Figure 1. The eukaryote tree of life and taxa studied in this thesis.

Several scenarios have been proposed for relationships among the major groups of eukaryotes. One of the earliest was the Archaezoa hypothesis, based largely on rRNA phylogeny. These trees showed a collection of microbes with divergent sequences and simple cell structure as the deepest branches in the tree (Cavalier-Smith, 1989; Sogin et al., 1989). These cells appeared to lack many otherwise common eukaryote features such as golgi stacks and mitochondria. Thus, the Archaezoa hypothesis proposes that these deep branching taxa are remnants of early eukaryote evolution, pre-dating the advent of mitochondria (J. R. Brown & Doolittle, 1995). However, Cavalier-Smith's Archaezoa are also mostly parasites, which tend to have divergent gene sequences that are artefactually drawn toward the base of a tree. Subsequent protein sequence phylogenies showed that at least one group of Archaezoa, the obligate parasites of Microsporidia, are actually fungi (Lee et

al., 2008). Meanwhile, microscopic studies revealed the presence of mitochondrion-liked organelles (hydrogenosomes and mitosomes; Embley & Hirt, 1998) in all Archaeozoa. As a result, the archaezoan root was eventually rejected as a long-branch attraction artefact (Palmer & Delwiche, 1996; Philippe et al., 2000).

Since the rejection of the Archaezoa root, several alternative scenarios have been proposed. Predominant among these is the unikont-bikont (UB) root. This was originally based on the status of the genes for dihydrofolate reductase (DHFR) and thymidylate synthase (TS), which are separate in predominantly uni-flagellate eukaryotes (unikonts: mostly animals and fungi) but fused in eukaryotes with two or more flagella (bikonts: Diaphoretickes and Excavata) (Stechmann & Cavalier-Smith, 2002). Then, assuming that separate DHFR and TS genes and a single flagellum are the ancestral condition for eukaryotes, fused genes and the second flagellum were derived in a unique bikont ancestor. However, gene fusion and fission are not as rare as originally thought, including for DHFR and TS (Dohmen et al., 2020; Maguire et al., 2014). It is also likely that animals and fungi are ancestrally bikont, since the bi-flagellate Apusozoa now appear to be their sister taxon (M. W. Brown et al., 2018). Thus, it appears likely that the eukaryote last common ancestor (LECA) was bikont, and the taxonomic designation unikont is defunct with its former taxa now subsumed into a larger Amorphea (Opisthokonts: Holozoa + Holomycoa, Amoebozoa, Apusozoa, Breviata, Subulatomonas, and, possibly, Malawimonada) (Adl et al., 2012).

Mitochondria

Eukaryotic genomes are mosaics consisting of genes derived from both Archaea and Bacteria (Brueckner & Martin, 2020; Cotton & McInerney, 2010). Eukaryote genes of archaeal ancestry (euArcs) seem largely involved in essential housekeeping functions, particularly information processing, while their bacterial-derived genes (euBacs) are largely associated with mitochondrial and chloroplast function. These organelles were derived from endosymbiotic bacteria, which brought a large number of genes with them in order to function and replicate within their host. Of the several thousand genes presumably present in the original endosymbionts, it is estimated that roughly half were lost as they were no longer required in the intracellular environment, while most of the remaining genes were transferred to the host nucleus. As a result, most organellar proteins are encoded in the nucleus, synthesized on cytoplasmic ribosomes and post-translationally imported into the organelles. This is particularly true for mitochondria where roughly 95-98% of their proteins are nuclear-encoded versus ~90% for chloroplasts. In addition, a large number of eukaryote genes were adapted or invented to service the organelles

in their new environment, e.g., to perform endosymbiont unique functions such as post-translational protein import. As a result, roughly half of nuclear-encoded mitochondrial proteins are of eukaryotic, rather than bacterial, origin (Gray, 2015).

A large body of now widely accepted evidence suggests that all eukaryotes have, or have had, a mitochondrion at some point in their evolutionary history. However, eukaryotes living in anaerobic or micro-aerophilic habitats mostly lack respiratory-competent mitochondria and instead have DNA-free mitosomes or hydrogenosomes. These are interpreted as degenerate mitochondria and in some cases, there is clear evidence that they are (Embley, 2006; Leger et al., 2016; Makiuchi & Nozaki, 2014; Zimorski et al., 2019). This leads to the conclusion that mitochondria evolved before the last eukaryotic common ancestor (LECA), along with nearly all major eukaryote-defining features such as the nucleus, golgi apparatus and cytoskeleton.

The mosaic nature of eukaryote genomes means that eToL can be rooted using either archaeal or bacterial homologs as the outgroup. Mitochondrial proteins are particularly appealing candidates to use in investigating the eToL root. Since mitochondria almost certainly arose after the origin of eukaryotes (the first eukaryote common ancestor or FECA) but before LECA, this makes Bacteria a closer outgroup than Archaea, whose relationship with eukaryotes pre-dates FECA. Nonetheless, there are important caveats in using mitochondrial genes to root eToL. For one thing, these analyses exclude eukaryotes without mitochondria, including most of the former Archaezoa (e.g., Metamonada; Karnkowska et al., 2016). It is also important to note that, while the ancestor of mitochondria is assumed to have been an alpha-proteobacterium (αP-bacteria), only a small fraction of mitochondrial genes support αP-bacteria as the sister group to mitochondria (Gabaldón, 2018; Gray, 2015; Kurland & Andersson, 2000; Pittis & Gabaldón, 2016). Mitochondrial proteomes are also surprisingly host-specific, and only ~10% of the ~1000 mitochondrial proteins are universal among mitochondriate eukaryotes (Gray, 2012; Gray et al., 2004). This considerably limits the number of genes that can be used to investigate the eukaryote root and excludes potentially important taxa from the analysis.

The first use of mitochondrial traits to define the eukaryote root proposed Diaphoretickes as the earliest branch of eukaryotes. This was based on the ER-mitochondria encounter structure, which is found in all eukaryotes except Diaphoretickes (Wideman et al., 2013). However, no published phylogenomic data have been found to support this root. The first multi-gene mitochondrial protein (mitoP) phylogeny supported the unikont-bikont (UB) root using 42 α-proteobacterial-like mitochondrial proteins with an α-proteobacterial

outgroup. However, these analyses only recovered this root in analyses using the site-specific CAT-GTR mixture model (Derelle & Lang, 2012). He et al. (2014) showed that over half of the 42 protein alignments used by Derelle and Lang were not suitable for phylogeny, either because of extremely short alignments, sporadic taxon sampling, or deeply problematic single-protein control trees. Furthermore, two strongly opposed phylogenetic signals were found in the data by both automated congruence testing (Leigh et al., 2011) and taxon jackknifing (He et al., 2014). Instead, He et al. (2014) surveyed all universal eukaryotic proteins of bacterial ancestry (euBacs) and identified 37 that produced credible trees. Phylogenetic analyses of these data with a variety of methods and models placed Discoba as the sister group to the rest of Eukaryotes, referred to as the neozoan-excavate root. However, further analyses by Derelle et al. (2015) rejected the He et al. root, claiming that the latter data were corrupt and again showing strong support for a UB root but again only with CAT-GTR (Derelle et al., 2015).

It is striking that the results of these three studies are so different despite the fact that their approaches are generally similar, and the data used overlap considerably in terms of taxon sampling and genes. However, there are a number of important differences. These include phylogenetic model selection, inclusion/exclusion of malawimonads, outgroup composition, and inclusion/exclusion of fast-evolving alignment positions. These are potentially major differences. For example, poorly represented taxa with large amounts of missing data such as the malawimonads can introduce conflict into a supermatrix tree, because sequence orthology is hard to establish (Aberer et al., 2013; Jeffroy et al., 2006; Philippe et al., 2017; Young & Gillung, 2020). Model selection for phylogenetic analysis is important since, while parameter-rich models may give a better fit to the underlying data, they do not necessarily provide more accuracy (Kelchner & Thomas, 2007). This is due to the risk of overfitting and potential bias when training models on the data set to be tested. There are also contradictory views regarding the removal of fast-evolving sites, which is crucial to support the UB root (Derelle et al., 2015; Derelle & Lang, 2012). While some studies suggest that removing the faster evolving alignment positions can improve the tree (Wu et al., 2012; see also Kück & Wägele, 2016), others suggest that it tends to degrade tree accuracy (Tan et al., 2015). There is also a striking difference in the ingroup-outgroup distance, which is considerably shorter for the α-proteobacterial derived proteins used by Derelle et al. (2015) versus the combination of α - and γ-proteobacterial derived proteins used by He et al. (2014).

Testing congruence

Phylogenomics uses multiple genetic loci combined as a single concatenated alignment to construct evolutionary trees. The core premise of this approach is the assumption of congruence, *i.e.*, a single shared phylogenetic history for all of the concatenated loci (Huelsenbeck et al., 1996). The idea is that, with increasing data, the true phylogenetic signal should accumulate while random error (noise) will cancel out. Violation of the congruency assumption can give unsupported or positively misleading results (Kupczok et al., 2010; Philippe et al., 2011; Wägele et al., 2009). Thus, assessing congruence across a multi-gene dataset is a critical task. Incongruence is primarily evaluated by visual inspection of single-gene/protein trees (SGTs) reconstructed from nucleotide sequences or their conceptually translated proteins. Ideally, these trees should reproduce substantial amounts of the known underlying tree (canonical phylogeny) making it possible to assess the reliability of the data, or at least not produce any non-canonical phylogeny with strong statistical support (e.g., bootstrap support > 60-70%). However, SGTs invariably lack sufficient phylogenetic signal to resolve the deepest branches in the tree making it difficult to assess congruence at these levels. The varying length and conservation of different loci also make it challenging to apply a uniform standard evaluation process for SGTs. As a result, evaluating SGTs individually becomes a complicated process for large numbers of loci, as well as laborious and time-consuming. Various attempts to devise algorithmic solutions to these problems have been implemented with an uncertain degree of success (De Vienne et al., 2012; Leigh et al., 2008, 2011; Planet & Sarkar, 2005; Smith et al., 2020).

The primary source of incongruence among genetic loci is when paralogs or xenologs, sequences acquired by HGT, are mistakenly classified as orthologs. This problem is particularly challenging when genes or taxa evolve with substantially different patterns. Strategies for tackling multi-gene congruency can be classified into two categories: detecting outliers by measuring the deviation from a global reference point or clustering partitions into subsets based on a compatibility score. Each approach has its merits. In both cases, a metric is needed to measure the compatibility of deviation scores, whether between loci (genes or proteins) or between loci and a global reference point. This metric is usually based on tree dissimilarity (distance), tree likelihood, or a combination of the two. There are several metrics designed to measure the distance between trees. Three of the most basic and commonly used ones in multi-gene congruency tests are the Robinson-Foulds (RF) distance (Robinson & Foulds, 1981), path difference (Penny et al., 1982; W. T. Williams & Clifford, 1971) and Quartet metric (Day, 1986; Estabrook et al., 1985). There are many other methods to measure tree-to-tree distances, but these are harder to compute, especially for large trees, because they involve

searching for the minimum steps of change needed to transform one tree into another, for example, geodesic distance (Kupczok et al., 2008). Such methods are also computationally expensive for large trees.

There are two major problems with tree distance measurements – missing data and evolutionary pattern differences. Regardless of the metric, missing data is especially challenging for clustering approaches, because with missing taxa, congruency becomes a non-transitive relation. That is, given a set of three genes {A, B, C}, if A and C have taxon data that is missing in B, then it is possible for B to be congruent with both A and C, even if A and C are not congruent with each other (Figure 2). Misplacement of a few taxa can push the distance between two otherwise identical trees to the maximal limit, and, in some instances, one taxon is enough to induce such an effect. Clustering tools that use partition compatibility or RF distance tend to eliminate non-shared taxa, which helps in the calculations but does not address the transitivity problem. Outlier detection methods are not troubled by the transitivity conflict irrespective of the metric because each partition's taxon set is a subset of the global reference. Nevertheless, outlier detection methods still need to incorporate techniques to account for missing data in their implementation either through imputation or normalization.

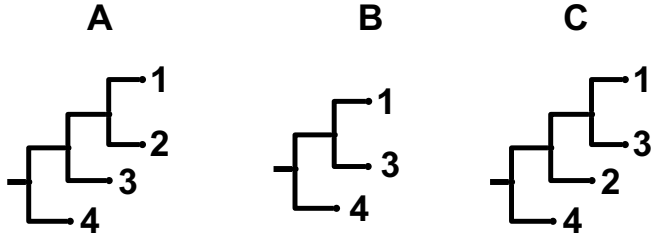

Figure 2 – Illustration of non-transitivity of congruency relations between three trees A, B, and C in which A and C are not congruent with each other but both are congruent to B because B is missing leaf node 2.

The second major problem is when different partitions have substantially different signal-to-noise ratios. Since the distance between trees is used as an indicator of incongruence, large amounts of random noise can cause two evolutionary congruent loci with only slightly different tree topologies to have a large distance between them. For instance, a gene with a low but true signal may produce a topology where one taxon is misplaced and seems incongruent. In contrast, another gene may produce a similar topology but with a strong signal (e.g., HGT). This problem is controlled for in visual SGT assessment when support values are used for guidance, albeit requiring a substantial amount of manual work. However, there is no definitive and reliable way to

automate the assessment process, especially for sorting out paralogs and xenologs.

Outlier and clustering approaches have different advantages over each other. Outlier tools are especially important for congruence analysis when examining deep nodes. This is because alternative topologies at the deepest nodes have very little impact on the overall tree distances. In other words, the topological distance between competing alternative topologies is smaller than the vast majority of topologies produced by single genes from the same data. On the other hand, a clustering approach is more useful if the competing hypotheses are relatively far apart and involve many alternative topologies, that is, when there is no approximate reference point.

Another interesting aspect of incongruence is the possibility that incongruence does not affect the entire gene. This could be due to genetic recombination between paralogous genes or between a host gene and a homolog acquired by HGT (xenolog). In either case, the result will be a mosaic gene consisting of fragments with different evolutionary histories. Both HGT and paralogous recombination can give very strong phylogenetic signals so that even a relatively small amount of non-orthologous sequence may make the entire sequence appear incongruent. Recombination between xenologs may not be uncommon because successful HGT includes incorporation of foreign DNA into the new host genome. This could be either random illegitimate recombination or recombination with a host homolog already present. One of the few clear examples of this is plant mitochondrial DNA, which carry foreign DNA recently acquired from distant relatives (Hao et al., 2010). Even if the xenologous fragments are small, they could still influence even multigene phylogeny. This could occur if, for example, there are multiple HGT events from the same or closely-related donors such as a symbiont, endosymbiont, parasite or common food source. Such contaminating signal could distort the species' phylogenetic signal or lead to a mosaic gene being flagged as incongruent and unnecessarily deleted in its entirety from a concatenated data set.

Various methods have been developed for automated testing of congruence in phylogenomic data. A relatively recent one is Conclustador (Leigh et al., 2011), a clustering tool that relies on tree distance. It uses a distribution of bootstrap or posterior trees for each locus and measures the distance between each pair of loci by counting the number of shared bi-partitions in their tree-set, excluding non-shared taxa. This is followed by loci clustering analyses using either k-means or spectral clustering. While the exclusion of non-shared taxa is necessary for calculating the distance between two loci, it ignores the transitivity of the relations between them. The transitivity problem seems more suitably addressed as a maximal clique problem (Day & Sankoff, 1986)

but using loci instead of characters. Conclustador also requires developer support which is no longer available.

The other more popular method is Phylo-MCOA (De Vienne et al., 2012). This one can be classified as an outlier detecting tool that relies on tree distance. It addresses the problem of missing data by imputation using mean substitution. This is possible because Phylo-MCOA first transforms the trees into distance matrices (either nodal or patristic), which allows filling of the missing data from the average values calculated from other matrices where the absent relation is present. Phylo-MCOA uses multiple co-inertia analysis (MCOA) to compare the topologies and identify outliers at two levels: complete outliers (either gene or species), and cell by cell outliers. The main drawback of Phylo-MCOA is taking one tree as the sole representative of each locus so that it cannot evaluate the significance of the observed topological differences between loci.

Various other tools have been developed for testing congruence, but these are either no longer supported or not used due to their limited ability to cope with the deep conflicts of phylogenomic data. Concaterpiller (Leigh et al., 2008) is a clustering tool that relies on likelihood ratio tests between constrained and relaxed topologies, a concept introduced by Huelsenbeck and Bull (1996). The program uses a hierarchical clustering procedure starting from the pair of loci that has the lowest likelihood ratio, combines them and then tests their significance of congruency using a user-defined α threshold and bootstrap analysis. The analysis stops when the p-value is less than the predefined α threshold. Smith et al. (2020) proposed a somewhat similar but more sophisticated method that incorporates RF distances and bipartition analysis of the single gene trees with reference to constrained topologies as well as heuristic strategies to find the most congruent subset of data. Shen et al. (2017) applied the same concept in their work to calculate the log-likelihood differences, site-wise and loci-wise, on full data between topologies of alternative hypotheses and then contrasted the distribution these differences across the data between the two topologies. As with Conclustador, these three methods only identify incongruent genes, and not which taxa are the source of incongruence. Planet and Sarkar (2005) implemented an automated process in their mILD tool, which identifies incongruence at the sequence level through sequential (one by one) jackknife until all partitions are congruent. The jackknife analysis is provided in mILD in addition to the main analysis, which is the pairwise Incongruence Length Difference (ILD) test. CADM (Campbell et al., 2009) is another congruency test that differs from the previous ones in that it assumes complete incongruency as the null hypothesis. It also uses distance matrices to represent trees, but it does not deal with missing data which makes it unsuitable for most phylogenomic datasets.

Evolutionary models

Evolutionary models are an important part of any phylogenetic analysis. Models and methods have evolved in parallel with computer power, from simple unweighted parsimony and distance methods to complex parameter-rich models used in Bayesian inference and maximum likelihood. Since their introduction in 2004 (Lartillot & Philippe, 2004), site heterogeneous (mixture) models have been gaining in popularity, particularly the CAT-GTR model implemented in the program PhyloBayes (Lartillot et al., 2013) and the empirical profile mixture models implemented in IQTree (Le et al., 2008; Wang et al., 2018). These highly computationally demanding models are trained on the data and thus expected to be more flexible and better at modeling the underlying evolutionary history. This is thought to make tree reconstruction much more resistant to the highly problematic artefact of long-branch attraction, whereby unrelated taxa with high evolutionary rates tend to be drawn together in phylogenetic trees (Bergsten, 2005; Felsenstein, 1978). In fact, phylogenomic analyses with PhyloBayes CAT-GTR have been found to support several hypotheses undetected by other methods (Cannon et al., 2016; Feuda et al., 2017; Laumer et al., 2015; H. Li et al., 2015; Sharma et al., 2014; Simion et al., 2017; Song et al., 2016; Struck et al., 2014).

Parameter-rich evolutionary models are especially appealing as they should have the flexibility to cope with complex and varying evolutionary patterns that are especially problematic in "deep" phylogeny. However, this flexibility comes at a high computational cost, which hampers model evaluation and analytical testing of results such as evaluating alternative hypotheses or testing the influence of taxon or data subsets by jackknifing. Such complex models also run the risk of overfitting and, in the presence of incongruence, of estimating misleading parameters that fit to artefact rather than true signal (Kelchner & Thomas, 2007). The presence of non-orthologous sequences becomes more critical as more parameters are estimated from the data, making the removal of deviant data a critical step to prevent bias model parameters (Philippe et al., 2011, 2017). The problem of over-fitting comes from the trade-off between model complexity and parameter estimation error; every added parameter carries an estimation error cost. Thus, the simplest model that fits the data "reasonably" well is to be preferred (Abdo et al., 2005; Huelsenbeck et al., 2004; Minin et al., 2003; Ripplinger & Sullivan, 2008).

Although the potential problem of over-fitting is well-recognized, determining when increased model complexity is justified is not. The potential benefit of increased parameterization can be evaluated with measures such as the Akaike information criteria (AIC) or Bayesian information criteria (BIC). However, these grade the performance of the model on a given data set and do not test the reliability or the consistency of the model on unseen data as the resampling

methods try to simulate. Standard non-parametric bootstrapping (Efron, 1981) is the most widely used and accepted resampling method in phylogenetics, and bootstrap support values are generally expected to decrease with over parameterization. However, bootstrapping is often computationally infeasible with parameter-rich models like CAT-GTR and C60. Therefore, the analyses either rely on the posterior probabilities in the case of Bayesian Inference (Lartillot et al., 2013) or implement model approximation techniques to minimize the computational burden as in the case of posterior mean site frequency modeling (PMSF; Wang et al. 2018). However, posterior probabilities tend to be extremely high so there is little to distinguish between strong and weak hypotheses. Thus, posterior probabilities of below 0.95 have been found to be unreliable (Alfaro et al., 2003; Cummings et al., 2003; Douady et al., 2003; Erixon et al., 2003). Moreover, testing of this type is simply not feasible with computationally demanding rich-mixture models.

The development of new and more fit evolutionary models, especially for protein data, has been an ongoing quest since the first general substitution matrix (Dayhoff et al., 1978). Mixture models aim to capture more data patterns by incorporating more empirical substitution matrices (Le, Lartillot, et al., 2008) or frequency profiles (Le, Gascuel, et al., 2008). This means higher computational demands and likely higher risks of over-fitting. A safer approach is to make the substitution model more specific to the data, either to certain groups of organisms (Minh et al., 2021) or to the underlying protein structure (Le & Gascuel, 2010). For reconstructing eToL, the latter seems reasonable but requires the data to have a previously solved protein structure, which is likely unavailable for most of the data.

Main research aims

The primary aim of my doctoral thesis is to reconstruct major branches in the eukaryote tree of life (eToL), focusing on the root and relationships among the major groups. To reach this goal, these are the main objectives:

1. Assemble a data set of eukaryote genes of bacterial origin and use this to construct a phylogeny rooted with Bacteria (**Paper I**).
2. Assemble a data of eukaryote genes of archaeal origin in order to include amitochondriate excavates groups in a phylogeny rooted with Archaea (**Paper II**).
3. Develop new tools and pipelines to automate the process of curating phylogenomic datasets and minimizing their incongruency and to better understand how their constituent genes evolve (**Paper I**).
4. Development of a new model for fast and accurate phylogeny that incorporates predicted protein structures (**Paper II**).
5. Develop an alternative to outgroup rooting by combining reciprocal-rooting with phylogenomics (**Paper III**).
6. Investigate the significance of detected incongruent sequence fragments in protein data (**Paper IV**).

Paper Summaries

Paper I

Al Jewari C, Baldauf SL. 2022. Conflict over the eukaryote root resides in strong outliers, mosaics and missing data sensitivity of site-specific (CAT) mixture models. *Systematic Biology* 12:syac029. doi: 10.1093/sysbio/syac029.

Theoretical considerations suggest that phylogenetic rooting is most accurate with data from the closest outgroup (Kinene et al., 2016; Wheeler, 1990; T. A. Williams, 2014). The eukaryote tree can be rooted with either Bacteria or Archaea, since both contributed to the origin of eukaryotes. Bacteria have been favored over Archaea to root the eukaryote tree (eToL) because most of the bacteria-related eukaryote genes encode mitochondrial functions, and the acquisition of mitochondria is believed to post-date the original split of eukaryotes from Archaea, but pre-date the eukaryote last common ancestor. This is confirmed by the generally closer evolutionary distance between bacteria and eukaryotes in phylogenies based on eukaryotic genes of bacterial ancestry (euBacs) versus eukaryotic genes of archaeal ancestry (euArcs).

Three previous phylogenomic studies have attempted to root eToL using supermatrix phylogeny and a bacterial outgroup (Derelle et al., 2015; Derelle & Lang, 2012; He et al., 2014). Despite the similarity between the approaches and their assembled data, these studies found very different results regarding the earliest split in eToL. He et al. (2014) concluded that Excavata, represented in these data by Discoba, is a sister group to the other two major groups of mitochondriate eukaryotes, Amorphea and Diaphoretickes (He et al., 2014). However, studies by Derelle and Lang (2012) and Derelle et al. (2015) supported Amophorea as the sister group to Diaphoretickes and Discoba, similar to the older unikont-bikont hypothesis (Stechmann & Cavalier-Smith, 2002).

In order to try to resolve this conflict, I assembled a dataset combining all the data from the three previous studies into one larger supermatrix of 76 protein and systematically analysed the data for possible sources of artefact. To do this, I designed two protocols to investigate and remove the causes of

phylogenetic conflicts. While there have been several methods and tools to assess the congruency (agreement) among the different component sequences of multigene dataset, all of these methods do the assessment based on single/protein gene trees (SGTs). However, small but systematic artefacts can not necessarily be detected in SGTs but still can substantially influence the supermatrix phylogeny. Therefore, I developed a protocol to assess the consistency of individual loci in multiple random subsets of the full data, using a protocol I call ConJac. I also wanted to screen the alignments for mosaicism, which could result from partial horizontal gene transfer (HGT). HGT of whole genes is fairly easy to detect in SGTs, but detection of small exogenous fragments should be much harder. Given their small size, it is unlikely they could impact the supermatrix phylogeny unless there are many fragments derived from the same source, such as a parasite or symbiont. For this, I developed the sliding windows method ConWin. I then used these protocols to screen the euBac supermatrix and tested the root using phylogenetic analysis of the different versions of the data, including before and after screening, as well as the different evolutionary models used in the previous studies.

With ConJac, I created 1976 data subsets, each with 14 loci, where each locus is represented in 364 subsets. Trees are inferred for each subset, each tree is converted to a distance matrix and then the full set of matrices is analyzed collectively to identify which sequences appear in the most deviated subsets relative to a globally estimated mean. I then removed the top 10% most deviant points from the original supermatrix. I could then construct phylogenetic trees from the data before and after masking to test whether persistent outliers were affecting the results. For the ConWin analyses, I focused on assessing the Discoba since they are the least taxonomically sampled eukaryotic supergroup in the data set. In contrast, the other two groups, Amorphea and Diaphoretickes, are much better sampled and showed fewer outliers in the conJac results. These analyses identified 587 potentially mosaic fragments (potentially incongruent kernels or PIKs) in the 76 euBac genes of the Discoba. By analyzing the data before and after masking of these PIKs, I could test whether mosaicism was affecting the position of the Discoba in the resulting phylogenetic trees.

I then analysed the full 76 protein euBac supermatrix, with data from 178 taxa, using various versions of the data. This included before and after masking with ConWin and ConJak as well as data sets where the amount of missing data was severely reduced. Each of these different versions of the euBac data was also analysed using a variety of phylogenetic methods and models. The results indicate that the Discoba root is overall the most consistent hypothesis. In fact, it was only with unfiltered data, or data sets with large amounts of missing entries, that the alternative hypothesis was supported. Moreover, the

conflicting results are mostly associated with the most complex evolutionary models. Despite the current popularity of these models, my results suggest that they are overly sensitive to outliers and missing data, such that these models can support radically different outcomes depending on how or whether the data are curated.

Paper II

Caesar Al Jewari and Sandra L. Baldauf. An excavate root for the eukaryote tree of life (in review)

Excavata is a highly diverse collection of unicellular eukaryotes that was first recognized two decades ago based on the shared trait of a morphologically complex "excavated" feeding groove (Simpson & Patterson, 1999). The monophyly of these taxa is uncertain and has never been tested in a rooted multigene phylogeny. The Excavata is currently classified into three main groups, Discoba, Malawimonada and Metamonada. The first two have fairly normal mitochondria, while the metamonads live exclusively in anaerobic or low-oxygen environments and have only highly-reduced mitochondria-like organelles (hydrogenosomes or mitosomes). These organelles lack mitochondrial DNA, and thus metamonads also lack nearly all of the euBac proteins used in **Paper I**.

Discoba monophyly has been confirmed by several previous studies, and I showed in **Paper I** that they are likely to be the earliest branch in eToL, or at least the earliest branch of actively mitochondriate eukaryotes. However, this study was based on bacterial proteins and therefore it excluded the metamonads. Therefore, I developed a new data set to test the position of the various metamonads in the eukaryote tree of life. Metamonada actually consists of three very distinct groups - the Preaxostyla, Fornicata and Parabasalia. Since their monophyly is not well-established, I choose to refer to these taxa by the informal designation, amitochondriate excavates.

In order to construct a multigene data set including all eukaryotes and a well-defined outgroup, the only other option is to work with eukaryote proteins of archaeal ancestry (euArcs). To obtain the data, I screened a large comprehensive database of archaeal genomes, the arCOGs database (Makarova et al., 2015), to identify proteins that are universal among eukaryotes and widely distributed in Archaea. This led me to assemble a dataset of 198 euArc proteins for 178 taxa, including substantial taxon representation of all well-established major divisions of eukaryotes and archaea. The critical challenge was the amitochondriate excavates, for which

little fully-assembled public sequence data was available. However, I was able to find many publicly available raw transcriptome data files for the three metamonad groups, which I assembled and screened for usable sequences. I then screened the data extensively for evidence of HGT and paralog (multicopy genes). The latter was particularly challenging and critical, because early eukaryote evolution appears to have been rife with gene duplication.

Phylogenetic analyses of the 198-protein euArc data using a wide selection of evolutionary models all place the three groups of amitochondriate excavates as separate branches at the base of the first eukaryote tree. Each of the excavate branch points receives full maximum likelihood bootstrap (mlBP) support (100% mlBP) with all combinations of evolutionary models and implementations. The next higher branch of eukaryotes in all cases consists of the mitochondriate excavate group Discoba, again with 100% mlBP support. Thus, it appears that the mitochondriate eukaryotes are evolutionarily deeply embedded in a series of excavate lineages. This means that the eukaryote last common ancestor was an excavate, and that eukaryotes probably were exclusively excavate for much of their early history. This "multi-excavate" root also suggests that the advent of aerobic mitochondria was not, as some argue (Lane & Martin, 2010; Tria et al., 2021), the earliest defining event in eukaryote evolution.

Model selection has become something of a holy grail in the field of phylogenetics, and especially phylogenomics, which uses large supermatrices composed of 10's to 100's of sequences for each taxon in the data set. The general consensus in the field is that more complex models have a better fit to the data and therefore result in more accurate trees, however, this may not necessarily be the case (Spielman, 2020). Complex models also come at a very high cost in terms of speed and computer resources. Since a single analysis of a large data set using such models can take weeks or even months to complete, additional control analyses are unfeasible. Meanwhile, my results from **Paper I** and a few other studies (Y. Li et al., 2021; Whelan & Halanych, 2017) suggest that these complex models also have serious drawbacks in terms of over-fitting the data, which makes the models hyper-sensitive to artefacts such as outliers and missing data. This appears to be a result of the very large number of variables that must be estimated for complex models, since the uncertainty of these estimates is additive. Meanwhile, simpler models can often provide consistent and well-supported results. Nonetheless, simpler models generally display less fit to the data and may be more susceptible to artifacts such as long branch attraction effect.

In order to increase model fitness without increasing the number of categories or the number of free parameters in the model, I sought to incorporate protein

structural information into the phylogenetic inference. For this, I took advantage of a new deep learning approach to predicting protein solubility and secondary structure (Høie et al., 2022). This allowed me to quickly predict the structure of each protein for ten taxa from across the data set and use the consensus of these predictions to partition the alignment into six structure categories. Phylogenetic trees were then constructed using a corresponding set of six previously estimated structure-based protein substitution matrices (Le & Gascuel, 2010). I also ran a fitness benchmark analysis between various models and compared these to those of my predicted structure-based partition model. The results show that the new model has a fitness similar to the most complex models (Le et al., 2008), but with a fraction of the computational cost (Le & Gascuel, 2008). The reduced computational cost of the new model allowed me to conduct further control analyses, all of which produced the same phylogeny as the full data set analyses and with 100% mlBP support for the full multi-excavate root.

The multi-excavate root raises questions a number of interesting questions, including mode of the origin of mitochondria. If the mitochondrion was already present in the eukaryote last common ancestor, that would require that the organelle was lost three times independently, once near the origin of each of the three lineages of amitochondriate excavates. Therefore, we propose an alternative scenario where the mitochondrial-like organelles of the earliest eukaryotes were augmented by a second bacterial acquisition after the divergence of amitochondriate excavates but before the divergence of the mitochondriate excavates (Discoba) from the remainder of eukaryotes.

Paper III

Caesar Al Jewari and Sandra L. Baldauf. Reciprocal rooting with concatenating supports the multi-excavate root (manuscript).

The multi-excavate root for eukaryotes contradicts the currently dominant hypothesis of early eukaryote speciation, which places the Amorphea group, the supergroup that includes animals and fungi, as the first branch in the eukaryote tree (M. W. Brown et al., 2018; Cerón-Romero et al., 2022; Derelle et al., 2015; Katz et al., 2012). This makes it imperative to cross-check these results with an alternative method or data set. Reciprocal rooting is a viable alternative that obviates the need for an external outgroup and instead relies on paralogous loci to determine the root of the tree. The logic behind this is that the duplication node itself will be the root of the tree, since the duplication pre-dates all speciation events. In essence, the two paralogs serve as outgroups to each other. Since many duplication events predate the last common ancestor

of eukaryotes, it is possible to assemble a multi-locus data set to test the excavate root with a reciprocally rooted data set large enough to contain sufficient phylogenetic signal to resolve the deepest nodes. While this approach has been used before in a few cases, these cases have so far been limited to single locus trees.

In this study, I extended the reciprocal-rooting approach to multi-locus supermatrix phylogeny. I assembled a dataset of 35 pairs of orthologous proteins in which each pair consists of two orthologous sequence sets that are related via a duplication event that pre-dated the eukaryote last common ancestor. This means that both orthologous sequences for each paralog pair are found in all taxa. The data set was constructed by searching the 198 protein euArc data set from **Paper II** for clearly-distinct eukaryote-universal paralogs with the smallest branch distance between their component orthologs. Analyses of these data produce a phylogeny nearly identical to that derived by analyses of the full euArc supermatrix (**Paper II**). That is, the tree shows four separate branches of excavate taxa at the base of eukaryotes, with nearly all excavate branch points supported by 100% mlBP.

Unlike the outgroup-rooted supermatrix, the reciprocal supermatrix has the unique advantage of allowing various ways to concatenate the data. This is because the reciprocal supermatrix consists of two mirrored halves, referred to here as the upper and lower half of the matrix. This means that each paralogous pair of loci can be concatenated in two different ways: switching the component orthologs between the upper and lower half of the mirrored matrix. This leads to a vast number of potential random combinations that can be created allowing for control analyses similar to 50% sampling. Alternatively, all the loci can be concatenated in a scheme we are calling a 'supramatrix' where all the loci are combined in both possible orientations. This doubles the size of the matrix and maximizes the phylogenetic signal. In a third alternative, taxa can be deleted from either the upper or lower matrix, which makes it possible to conduct a jackknifing type of analysis to test the effects of individual taxa or groups of taxa on the resulting phylogeny. I used the latter approach to show that the multi-excavate was unaffected by the presence or absence of fast-evolving or phylogenetically unstable taxa, such as the supergroup SAR. In fact, all results of these analyses are compatible with the excavate root hypothesis placing the three amitochondriate excavate lineages as 2-3 separate branches at the base of the tree followed by Discoba as a sister clade to the remaining eukaryotes (Amorphea and Diaphoretickes). The tree node not definitively solved by these analyses is the branching order between the Parabasalia and Fornicata, which appear as the first and second major split in the full 198 protein analysis of **Paper II**. This suggests that the smaller 35 protein data set used in these analyses has insufficient phylogenetic signal to resolve the earliest split in the eukaryotes tree. However, even if

Parabasalia and Fornicata are sister taxa, rather than two separate branches, this still leaves a minimum of three major excavate branches at the base of the eukaryote tree.

Paper IV

Sandra L. Baldauf and Caesar Al Jewari, Widespread mosaicism in eukaryotic genes of bacterial ancestry (manuscript)

Horizontal gene transfer (HGT) refers to the non-vertical transfer of genetic material between unrelated organisms and is a major driving force in evolution, especially for Bacteria and Archaea. Sequences of HGT origin (xenologs) can have a strong impact on the accuracy of multi-locus phylogenies, with both supertree and supermatrix approaches. Identification of xenologs is a complex process and even more complex for partial transfer leading to mosaic genes. For **Paper I**, I developed a sliding window method to scan for possible mosaicism in multiple sequence alignments and applied the method to a multi-locus data set of euBac proteins. The method is designed to perform quick and rough scanning through sliding window trees targeting a specific taxon or taxon group of interest, in this case, the three main divisions of Discoba (Jakobida, Heterolobosea, and Euglenazoa). The goal of **Paper I** was to detect and eliminate any potential HGT fragments in order to reduce the error in the data. However, for individual mosaics, further analysis is needed to test their fidelity, identify their possible sources and detect any gene- or taxon-specific trends.

In this work, we investigate the top 18 euBac proteins predicted to be mosaics in one or more divisions of Discoba **in Paper I**. Proteins for the study were selected based on how deep the target taxon was found to be nested within a foreign group in the sliding window trees. The objective is to test the veracity of the ConWin results and investigate whether there are discernible patterns of transfer between identified donors and recipients. The latter could give some indication as to whether mosaics tend to arise from specific interactions such as symbioses or parasitism, or are more likely to be random acquisitions from the host's diet. We find that 10 out of the 18 predicted mosaics could be confirmed with bootstrap support above 70%. This shows that the ConWin protocol is sensitive enough to identify true mosaics. The most common donors are Amorphea, especially Amoebozoa, but overall, the mosaics appear to trace to taxonomically diverse sources. This suggests that mosaics are largely derived from food rather than specific species interactions. Most of the mosaics also appear to be old, affecting whole groups rather than individual taxa suggesting that mosaics arise rarely but can be long-lived once they do.

Popular science summary

Phylogenetic analyses aim to describe the kinship between different genes and organisms, mostly now by comparative analyses of their physical traits or sequences. Before the era of cheap molecular data, phylogenetics was mostly based on the examination of morphologically observable traits between life forms. However, this was not very useful for single-celled organisms (microbes), since they have little morphology to compare amongst them. Within the last few decades, the continuous advancement in cheap and fast sequencing techniques coupled with the ever-increasing power of computers has led to a massive increase in the availability of sequence data from across the tree of life, most of which is composed of microbes. This has revolutionized our understanding of the diversity of life and how it arose. Nowadays, almost all described species relationships are inferred using molecular data enabled by numerous and carefully optimized computational tools and statistical models. While a vast sampling of the diversity among living organisms has now been described and classified, some relationships are still unclear, especially the most ancient relationships in the tree of life. Indeed, evolution has proven to be a much more complex process than expected and therefore difficult to accurately model. As a result, numerous efforts have been made to improve the methods and the quality of data to reduce the amount of estimation error and provide a more accurate picture of the history of early life. Early speciation events are generally hard to estimate because the amount of the accumulated error increases in tandem with the evolutionary distance between organisms. This is furthered complicated by the fact that many of the evolutionary models we use do not generalize well to all organisms. However, data quality, evolutionary models, and sophisticated statistical approaches have been continuously improving, and we are gradually resolving more of such ancient relationships. This in turn gives us a much more accurate picture of how life first arose and how it has evolved through time.

The current classification of the tree of life into three main domains of life, Archaea, Bacteria, and Eukaryotes is now well established based on a very large body of data and scientific studies. Among, these the eukaryotic domain is particularly intriguing as it includes all complex-celled organisms, which encompass the vast majority of large, visible organisms such as Animals,

Plants, and Fungi. Our understanding of the evolutionary relationships among the major phyla of eukaryotes has been relentlessly revised as new species are described. However, some nodes in the tree are considerably tougher to resolve than others. Among the toughest questions are the composition of, and relationships among, the earliest branches of the eukaryote tree. Although these have received lots of attention due to their taxonomic and evolutionary importance, the problem has turned out to be one of the most difficult in phylogenetics and systematics. These are speciation events that date back to more than a billion years. While many theories have been proposed, proving any of them is a tremendous challenge, and so far, no consensus has emerged. Foremost among these questions is the identity of the root of the eukaryote tree of life. This defines the last common ancestor of eukaryotes, the organisms from which all extant eukaryote life evolved.

For my thesis research, I approached this problem from several angles. My goal was not only to identify the eukaryote root but to develop new methods to better understand the data and its strengths and weaknesses. In **Paper I**, I design two protocols to systematically detect, rank and remove outlier data points. The goal was to test how far such points could disturb the reconstructed phylogenies and which models of evolution are best suited to handle such data. Outliers are any data points that violate the assumptions under which the analysis is conducted. These include contaminated data or any conflicting data point that is likely not to follow the species tree and distort an already weak signal. I tested these protocols using a data set of eukaryote proteins derived from bacteria (euBacs) that I constructed in order to examine eukaryote relationships using the corresponding bacterial sequences as an external point of reference (outgroup) to root the tree. The euBac genes are especially for the function of mitochondria, which are referred to as the powerhouse of the eukaryote cell and arose early in eukaryote evolution from an endosymbiotic bacterium. I found that the data are highly complex and include many outlier sequences and fragments of foreign DNA, which have variable effects on the resulting tree depending on which evolutionary model is used to calculate it. This is especially a problem with the model most commonly used in these types of studies, referred to as the CAT model. However, by taking these inconsistencies into account, I was able to show that the earliest speciation event for the organisms included in these data is the Discoba, a type of eukaryotes referred to as excavate. Thus, I conclude that Discoba is the earliest branch of eukaryotes with active mitochondria.

Discoba is the only excavate group included in this study because Discoba are the only excavates that maintain functioning mitochondria, and therefore the only excavate with euBac proteins. This leaves open the question as to where the rest of the excavates fit in the eukaryote tree of life. To address this question, I needed to develop a new data set with a new outgroup. For this, I

took advantage of the fact that eukaryotes originally arose by the joining of cells from Archaea and Bacteria. The core set of eukaryote genes derived from bacteria, the euBacs, are mostly involved in mitochondrial function, and therefore found only in eukaryotes with active mitochondria. However, eukaryote genes of archaeal ancestry (euArcs) are universal and largely essential for all eukaryote life.

For **Paper II**, I constructed a data set of 198 euArc proteins with sequences from large and diverse sets of Archaea and eukaryotes. Most importantly these data include the three other major types of excavates, namely the Parabasalia, Fornicata and Preaxostyla. Data on these organisms are quite scarce, therefore I enriched these data by surveying and assembling protein sequences from large raw data files (transcriptomic SRA data). Assembly and quality control of these data was the most time-consuming part of the work. To ensure that the data were of the highest quality, I used a variety of methods of data assembly and extraction, followed by numerous rounds of phylogenetic analyses of individual genes to screen for various possible artefacts. In total, the data set consists of 198 sequences from 186 taxa resulting in 50226 alignment columns with which to build the trees and test the quality of the results.

I first analysed the full data set using various evolutionary models including the most commonly used ones and some newer less-well tested but potentially very powerful models. I also used the CAT model, since, although I and a few other researchers have now begun to question its validity, including the work described in my **Paper I**, it is still considered by many to be far superior to any other available model. All analyses of my data with all the different models yielded a single unambiguous conclusion, that the four excavate lineages represent the first four major branches of the eukaryote tree of life. The implications of this are profound. It means that for the first possibly one billion years of their history, the only eukaryotes were excavates, a complex morphology previously thought to be a unique aberration. Moreover, the first three excavate branches lack true mitochondria. While some people very strongly argue that the mitochondrion is what defines eukaryotes, my results imply that eukaryotes were true eukaryotes long before they had mitochondria.

Such novel results require numerous controls to rule out possible artefacts, particularly for a tree purporting to reconstruct ancient evolution. However, the CAT model is not only problematic, but is also slow and computationally demanding, requiring weeks or even months to complete a single analysis on a supercomputer array. Therefore, I also developed a new fast evolutionary model based on secondary protein structure and solubility, which have long been known to be two of the most important factors that control the way

protein sequences evolve. For this, I took advantage of a new deep-learning-based approach to predict protein structure, and then compared the speed and accuracy of my model with other commonly used models, including the CAT. The results show that my model has an accuracy comparable to CAT, but with a fraction of computational demand. Moreover, the simplicity of my model greatly reduces the risk of "over-fitting", which I showed in **Paper I** to be a major problem with the CAT. Having a fast and accurate model allowed me to run a series of controls for the most troublesome artifacts for such deep evolutionary trees. The results of these analyses showed that none of these problems affect the euArc results, and all analyses support the four-excavate root for eukaryotes.

The "multi-excavate" root essentially turns the eukaryote tree of life on its head. It will require major revision to current theories of how eukaryotes first evolved, and what are some of the fundamental forces shaping how cells work. Therefore, I felt it was important to continue to test the root. One of the biggest concerns in a "deep" evolutionary tree, such as the euArc tree, is the possibility that the outgroup (in this case Archaea) is too distantly related to the ingroup (the eukaryotes) to give an accurate root. Therefore, for **Paper III**, I used an alternative rooting technique that does not require an outgroup, a method referred to as reciprocal rooting. This uses gene duplications to root the tree. Early eukaryote evolution appears to have involved many gene duplications, which helps explain why most eukaryotes have much larger genomes than Bacteria or Archaea. Therefore, if we construct trees using genes that were duplicated before any speciation events in the tree, the sequences should result in two complete parallel phylogenies, since all organisms have both copies of each gene. Since the duplicated genes (paralogs) arose from an ancient duplication event, the root of the tree must lie between the two mirror trees. Essentially, each of the trees serves as each other's outgroup. This method was used previously with single pairs of duplicated genes. In fact, it was originally used to root the tree of life, showing that Archaea are the sister group to eukaryotes. However, the method has never before been applied to many genes at once. Since my 198 euArc protein data set includes many such gene pairs, I decided to try applying the reciprocal rooting approach to test the eukaryote root.

In all, I identified 35 pairs of duplicate genes in my original euArc data. I then combined these pairs of genes in all possible orientations and calculated a phylogenetic tree. This resulted in a tree with two mirror halves, as predicted, and these two halves had nearly identical topologies. Most importantly, both halves showed the four excavate lineages as the first four major branches of eukaryotes. Thus, these analyses confirmed the results of **Paper II**, i.e., a multi-excavate root for the eukaryote tree of life. I also experimented with different ways to combine the paired genes by randomly shuffling genes

between the two mirrored halves of the matrix. In addition, I tried deleting whole sets of taxa from the upper or lower half of the matrix. This allowed me to test whether some of the odder taxa in the tree were affecting the results and was able to show that they did not. This test also showed that the method should work even if the data are incomplete, which tends to be a problem in these types of studies. Thus, the method appears to be robust and potentially applicable to many problems in taxonomy, even when substantial chunks of the data are missing. Most importantly the results confirm that the earliest eukaryotes were excavates, from which all other eukaryotes have emerged.

In the final chapter of my thesis, **Paper IV**, I investigated some of the results from **Paper I**. One of the methods of data screening I developed for **Paper I** was a sliding-window method. The method moves along a sequence in steps, testing the gene for anomalies at each step. When I applied the method to the euBac genes of Discoba, I found evidence of foreign DNA fragments in many of the genes (mosaic genes). However, the method, which I named ConWin, was rapid but only approximate. This allowed me to screen a lot of data, but the mosaics identified were considered only potential. This was sufficient for the purpose of testing the impact of potential mosaics on the euBac phylogeny but does not definitely identify the mosaics or where they come from. **Paper IV** describes detailed analyses of 18 of the potentially mosaic euBac proteins. Surprisingly, all but two of these potentially mosaic proteins appear to be true mosaics. The foreign sequence fragments appear to have come from very different sources, suggesting that they probably originated from DNA leaked from the hosts' digestive food vacuoles rather than a specific symbiont or parasite. The genes also encode proteins involved in diverse functions, which further supports the idea that mosaics are random acquisitions rather than adaptation of specific pathways.

Throughout my work, I assembled large amounts of proteomic data from across the diversity of eukaryotes, Bacteria and Archaea. A major emphasis of my work was to rigorously examine these data to filter out instances of noise and erratic data points and to test the data under a variety of evolutionary models. In all, I have developed two new methods for screening large data sets for hidden artefacts, a new model for molecular phylogeny and a new approach to phylogenetic rooting. In the process, I have defined a new root for the eukaryote tree, with important implications for evolutionary theory. All the results of my work converge on a single major conclusion, that the earliest forms of eukaryotes bore a striking excavate morphology. Moreover, it seems likely that the current form of the mitochondrion arose at a much later stage in eukaryote evolution than previously thought, possibly involving more than one endosymbiotic event.

Svensk sammanfattning

Fylogenetiska analyser syftar till att beskriva släktskapet mellan olika gener och organismer, huvudsakligen nu genom jämförande analyser av deras egenskaper eller sekvenser. Innan upptäckten av molekylära data har fylogenetiken länge varit baserad på undersökning av morfologiskt observerbara egenskaper mellan livsformer. Detta var dock inte särskilt användbart för encelliga organismer (mikrober), eftersom de har liten morfologi att jämföra mellan dem. Under de senaste decennierna har den kontinuerliga utvecklingen av billiga och snabba sekvenseringstekniker i kombination med datorernas ständigt ökande kraft lett till en massiv ökning av tillgängligheten till sekvensdata från livets träd, varav de flesta består av mikrober. Detta har revolutionerat vår förståelse av livets mångfald och hur det uppstod. Nuförtiden antas nästan alla beskrivna artförhållanden med hjälp av molekylära data som möjliggörs av många och noggrant optimerade beräkningsverktyg och statistiska modeller. Medan ett stort urval av mångfalden bland levande organismer nu har beskrivits och klassificerats, är vissa samband fortfarande oklara, särskilt de äldsta relationerna i livets träd. Evolution har faktiskt visat sig vara en mycket mer komplex process än förväntat och därför svår att modellera exakt. Som ett resultat har många ansträngningar gjorts för att förbättra metoderna och kvaliteten på data för att minska mängden uppskattningsfel och ge en mer korrekt bild av historien om det tidiga livet. Tidiga artbildningshändelser är i allmänhet svåra att uppskatta eftersom mängden av det ackumulerade felet ökar i takt med det evolutionära avståndet mellan organismer. Detta kompliceras ytterligare av det faktum att många av de evolutionära modellerna vi använder inte generaliserar i stor utsträckning till alla organismer. Datakvaliteten, evolutionära modeller och sofistikerade statistiska tillvägagångssätt har dock kontinuerligt förbättrats, och vi löser gradvis fler av sådana uråldriga samband. Detta i sin tur ger oss en mycket mer korrekt bild av hur livet först uppstod och hur det har utvecklats genom tiden.

Den nuvarande klassificeringen av livets träd i tre huvuddomäner av livet, Archaea, Bakterier och Eukaryoter är nu väletablerad baserat på en mycket stor mängd data och vetenskapliga studier. Bland dessa är den eukaryota domänen särskilt intressant eftersom den inkluderar alla komplex-celliga organismer, som omfattar den stora majoriteten av stora, synliga organismer

som djur, växter och svampar. Vår förståelse av de evolutionära förhållandena mellan de viktigaste Phyla av eukaryoter har bearbetats obevekligt när nya arter beskrivs. Vissa noder i trädet är dock betydligt svårare att lösa än andra. Bland de svåraste frågorna är sammansättningen av, och relationerna mellan, de tidigaste grenarna av eukaryotträdet. Även om dessa har fått mycket uppmärksamhet på grund av sin taxonomiska och evolutionära betydelse, har problemet visat sig vara ett av de svåraste inom fylogenetiken och systematiken. Dessa är artbildningshändelser som går tillbaka till mer än en miljard år. Även om många teorier har föreslagits, är det en enorm utmaning att bevisa någon av dem, och hittills har ingen konsensus uppstått. Främst bland dessa frågor är identiteten för roten till livets eukaryota träd. Detta definierar den sista gemensamma förfadern till eukaryoter, organismerna från vilka allt existerande eukaryotliv utvecklades.

För min avhandlingsforskning närmade jag mig detta problem från flera håll. Mitt mål var inte bara att identifiera eukaryotroten utan att utveckla nya metoder för att bättre förstå data och dess styrkor och svagheter. I **uppsats I** designar jag två protokoll för att systematiskt upptäcka, rangordna och ta bort avvikande datapunkter, så kallade Outlier data. Målet var att testa hur långt sådana punkter kunde störa de rekonstruerade fylogenierna och vilka evolutionsmodeller som är bäst lämpade för att hantera sådana data. Outlier är alla datapunkter som bryter mot de antaganden under vilka analysen utförs. Dessa inkluderar förorenade data eller andra motstridiga datapunkter som sannolikt inte följer artträdet och förvränger en redan svag signal. Jag testade dessa protokoll med hjälp av en datamängd av eukaryota proteiner härledda från bakterier (euBacs) som jag konstruerade för att undersöka eukaryota relationer med hjälp av motsvarande bakteriesekvenser som en extern referenspunkt (utgrupp) för att finna roten av trädet. EuBac-generna är speciellt för funktionen av mitokondrier, som kallas kraftverket för den eukaryota cellen och uppstod tidigt i eukaryotevolutionen från en endosymbiotisk bakterie. Jag fann att data är mycket komplexa och inkluderar många avvikande sekvenser och fragment av främmande DNA, som har varierande effekter på det resulterande trädet beroende på vilken evolutionär modell som används för att beräkna det. Detta är särskilt ett problem med den modell som oftast används i dessa typer av studier, kallad CAT-modellen. Men genom att ta hänsyn till dessa inkonsekvenser kunde jag visa att den tidigaste artbildningshändelsen för de organismer som ingår i dessa data är Discoba, en medlem av en större grupp eukaryoter som kallas utgrävningar. Således drar jag slutsatsen att Discoba är den tidigaste grenen av eukaryoter med aktiva mitokondrier.

Discoba är den enda excavata-gruppen som ingår i denna studie eftersom Discoba är de enda utgrävningarna som upprätthåller fungerande mitokondrier och därför den enda utgrävningen med euBac-proteiner. Detta

lämnar frågan öppen om var resten av excavata får plats i livets eukaryota träd. För att ta itu med denna fråga behövde jag utveckla en ny datamängd med en ny utgrupp. För detta utnyttjade jag det faktum att eukaryoter ursprungligen uppstod genom sammanfogning av celler från Archaea och Bakterier. Kärnuppsättningen av eukaryota gener som erhållsfrån bakterier, euBacs, är mest involverade i mitokondriell funktion och finns därför endast i eukaryoter med aktiva mitokondrier. Emellertid är eukaryota gener av arkéal börd (euArcs) universella och till stor del väsentliga för allt eukaryotliv.

För **uppsats II** konstruerade jag en datamängd av 198 euArc-proteiner med sekvenser från stora och olika uppsättningar av Archaea och eukaryoter. Viktigast av allt inkluderar dessa data de tre andra huvudtyperna av utgrävningar, nämligen Parabasalia, Fornicata och Preaxostyla. Data om dessa organismer är ganska otillräckliga, därför berikade jag dessa data genom att kartlägga och sammanställa proteinsekvenser från stora rådatafiler (transkriptomiska SRA-data). Montering och kvalitetskontroll av dessa data var den mest tidskrävande delen av arbetet. För att säkerställa att data var av högsta kvalitet använde jag en mängd olika metoder för datasammansättning och extraktion, följt av många omgångar av fylogenetiska analyser av individuella gener för att screena för olika möjliga artefakter. Totalt består datamängden av 198 sekvenser från 186 taxa vilket resulterar i 50226 justeringskolumner för att bygga träden och testa kvaliteten på resultaten.

Jag analyserade först hela datamängden med hjälp av olika evolutionära modeller inklusive de mest använda och några nyare mindre väl testade men potentiellt mycket kraftfulla modeller. Jag använde också CAT-modellen, eftersom även om jag och några andra forskare nu har börjat ifrågasätta dess giltighet, inklusive det arbete som beskrivs i min artikel I, anses den fortfarande av många vara vida överlägsen alla andra tillgängliga modeller. Alla analyser av mina data med alla olika modeller gav en enda entydig slutsats, att de fyra excavata-linjerna representerar de fyra första stora grenarna av livets eukaryota träd. Konsekvenserna av detta är djupgående. Det betyder att under de möjligtvis första miljarder åren av deras historia var Excavata de enda eukaryoterna, en komplex morfologi som tidigare ansågs vara en unik abnormitet. Dessutom saknar de tre första excavata grenarna äkta mitokondrier. Medan vissa människor mycket starkt hävdar att mitokondrien är det som definierar eukaryoter, menar mina resultat att eukaryoter var riktiga eukaryoter långt innan de hade mitokondrier.

Sådana nya resultat kräver många kontroller för att utesluta möjliga artefakter, särskilt för ett träd som påstår sig rekonstruera forntida evolution. CAT-modellen är dock inte bara problematisk, utan den är också långsam och beräkningskrävande, som kräver veckor eller till och med månader för att slutföra en enda analys på en superdator-array. Därför utvecklade jag också

en ny snabb evolutionär modell baserad på sekundär proteinstruktur och löslighet, som länge har varit kända för att vara två av de viktigaste faktorerna som styr hur proteinsekvenser utvecklas. För detta utnyttjade jag en ny djupinlärningsbaserad metod för att förutsäga proteinstruktur och jämförde sedan hastigheten och noggrannheten hos min modell med andra vanliga modeller, inklusive CAT. Resultaten visar att min modell har en noggrannhet jämförbar med CAT, men med en bråkdel av beräkningsbehovet. Dessutom minskar min modells enkelhet avsevärt risken för "överanpassning", vilket jag visade i **uppsats I** vara ett stort problem med CAT. Att ha en snabb och exakt modell gjorde att jag kunde köra en serie kontroller för de mest besvärliga artefakterna för sådana djupa evolutionära träd. Resultaten av dessa analyser visade att inget av dessa problem påverkar euArc-resultaten, och alla analyser stöder fyra-excavata-roten för eukaryoter.

Den "multi-excavata" roten vänder i huvudsak livets eukaryota träd på huvudet. Det kommer att kräva stor översyn av nuvarande teorier om hur eukaryoter först utvecklades, och vilka är några av de grundläggande krafterna som formar hur celler fungerar. Därför kände jag att det var viktigt att fortsätta testa roten. Ett av de största problemen i ett "djupt" evolutionärt träd, som euArc-trädet, är möjligheten att utgruppen (i det här fallet Archaea) är för långt besläktade med ingruppen (eukaryoterna) för att ge en korrekt rot. Därför använde jag för **uppsats III** en alternativ teknik till att finna roten som inte kräver en utgrupp, en metod som kallas reciprok rotning. Detta använder genduplicering för att rota trädet. Tidig eukaryotevolution verkar ha involverat många genduplicering, vilket hjälper till att förklara varför de flesta eukaryoter har mycket större genom än Bakterier eller Archaea. Följaktligen, om vi konstruerar träd med gener som duplicerades före eventuella artbildningshändelser i trädet, borde sekvenserna resultera i två fullständiga parallella fylogenier, eftersom alla organismer fick båda kopiorna av varje gen. Eftersom de duplicerade generna (paralogerna) uppstod från en uråldrig dupliceringshändelse måste trädets rot ligga mellan de två spegelträden. I huvudsak fungerar vart och ett av träden som varandras utgrupp. Denna metod användes tidigare med enstaka par av dubblerade gener. Faktum är att det ursprungligen användes för att rota livets träd, vilket visar att Archaea är systergruppen till eukaryoter. Metoden har dock aldrig tidigare tillämpats på många gener samtidigt. Eftersom min 198 euArc-proteindatauppsättning innehåller många sådana genpar, bestämde jag mig för att försöka tillämpa den ömsesidiga metoden till att finna roten för att testa eukaryotroten.

Sammanlagt identifierade jag 35 par dubbletter av gener i mina ursprungliga euArc-data. Jag kombinerade sedan dessa genpar i alla möjliga orienteringar och beräknade ett fylogenetiskt träd. Detta resulterade i ett träd med två spegelhalvor, som förutspått, och dessa två halvor hade nästan identiska

topologier. Viktigast av allt, båda halvorna visade de fyra excavata-linjerna som de första fyra stora grenarna av eukaryoter. Sålunda bekräftade dessa analyser resultaten av **uppsats II**, det vill säga en multi-excavata rot för livets eukaryota träd. Jag experimenterade också med olika sätt att kombinera de parade generna genom att slumpmässigt blanda gener mellan de två speglade halvorna av matrisen. Dessutom försökte jag ta bort hela uppsättningar av taxa från den övre eller nedre halvan av matrisen. Detta gjorde att jag kunde testa om några av ovanliga taxa i trädet påverkade resultaten och kunde visa att de inte gjorde det. Detta test visade också att metoden borde fungera även om data är ofullständig, vilket tenderar att vara ett problem i den här typen av studier. Således verkar metoden vara robust och potentiellt tillämpningsbar i många problem inom taxonomi, även när betydande bitar av data saknas. Viktigast av allt bekräftar resultaten att de tidigaste eukaryoterna är Excavata, från vilka alla andra eukaryoter har dykt upp.

I det sista kapitlet av min avhandling, **uppsats IV**, undersökte jag några av resultaten från **uppsats I**. En av metoderna för datascreening som jag utvecklade för **uppsats I** var en sliding window metoden rör sig längs en sekvens i steg och testar genen för anomalier i varje steg. När jag tillämpade metoden på Discobas euBac-gener hittade jag bevis på främmande DNA-fragment i många av generna (mosaikgener). Metoden, som jag döpte till ConWin, var dock snabb men bara ungefärlig. Detta gjorde det möjligt för mig att screena en hel del data, men de identifierade mosaikerna ansågs bara vara potentiella. Detta var tillräckligt för att testa effekten av potentiella mosaiker på euBac-fylogenin, men identifierar inte definitivt mosaikerna eller var de kommer ifrån. Artikel IV beskriver detaljerade analyser av 18 av de potentiellt mosaik-euBac-proteinerna. Överraskande nog verkar alla utom två av dessa potentiellt mosaikproteiner vara sanna mosaiker. De främmande sekvensfragmenten verkar ha kommit från mycket olika källor, vilket tyder på att de troligen härstammar från DNA som läckt från värdarnas matsmältningsvakuoler snarare än en specifik symbiont eller parasit. Generna kodar också för proteiner involverade i olika funktioner, vilket ytterligare stöder tanken att mosaikproteiner representerar slumpmässiga förvärv snarare än anpassning av specifika vägar.

Under hela mitt arbete samlade jag ihop stora mängder proteomisk data från mångfalden av eukaryoter, bakterier och archaea. En stor huvudpunkt i mitt arbete var att noggrant undersöka dessa data för att filtrera bort förekomster av brus och oregelbundna datapunkter och att testa data under en mängd olika evolutionära modeller. Sammantaget har jag utvecklat två nya metoder för att screena stora datamängder för dolda artefakter, en ny modell för molekylär fylogeni och ett nytt förhållningssätt till fylogenetisk rotbildning. I processen har jag definierat en ny rot i eukaryotträdet, med viktiga implikationer för evolutionsteorin. Alla resultat av mitt arbete sammanfaller med en enda viktig

slutsats, att de tidigaste formerna av eukaryoter bar en slående excavata-morfologi. Dessutom verkar det troligt att den nuvarande formen av mitokondrien uppstod i ett mycket senare skede i eukaryotevolutionen än man tidigare trott, möjligen involverat mer än en endosymbiotisk händelse.

Concluding remarks and future perspectives

I came to this work with a background in computer science, bioinformatics and statistics. I started working on eukaryote phylogeny as a course project in 2016, during my MSc in Bioinformatics. This was the first I knew about the controversy over the eukaryote root, and I found I wanted to be involved in this work. Therefore, I choose my master's project to try and resolve the euBac root of eukaryotes. Unsurprisingly, the problem turned out to be much more complex, and I concluded that the data carry mixed signals. I started my doctoral studies in 2018 and the main goal was to work on the root node and solve it. However, I found much of my time was spent devising ways to work reliably with complex data. After almost six years of working with eukaryotes phylogenies, I realize that carefully designed phylogenetics will usually result in useful answers. However, like all evolutionary biology, we can never be 100% sure about what we find. The truth also reveals itself in increments. The conclusions I reach here appear to be correct, and they are certainly highly supported, but new facts may emerge. For instance, so much of eukaryote diversity is undiscovered, and there may be yet unknown organisms out there that represent even earlier branches of eukaryotes and with very different morphologies and organelles. Therefore, there will always be some doubt and I hope other researchers will continue to dispute these conclusions.

I have also learned some basic principles, most importantly that data quality is paramount. No method or model, no matter how sophisticated, can recover the past if the past is not accurately reflected in the data. Garbage in garbage out. Unfortunately, the large size of modern supermatrices can give a fast sense of confidence, under the assumption that, if there is enough data, any problems will sort themselves out and the true signal will emerge. The complexity of resolving deep nodes in any phylogenetic tree with multi-locus data can be narrowed down to three main factors: reliability of the data, model selection, and controls. The way these factors affect each other is what really makes the problem hard. First, it is impossible to make sure that all loci are pure orthologs because there is rarely enough phylogenetic signal with single loci, especially for the deeper nodes in a tree. Second, it is very difficult without external evidence to tell whether a contentious node is the result of a lack of phylogenetic signal, which may lead to LBA artefact, or due to hidden orthology violations. Both LBA and orthology violations often lead to highly

supported erratic resolutions. More complex models are undoubtedly better at handling weaker phylogenetic signals, therefore better suited at handling LBA, but they are also conceivably worse at handling orthology errors. Since LBA nodes result from weak signal, then such nodes will also be at the mercy of orthology violations if such signals reach them. In order to test for both LBA and orthology violations, the analyses should be conducted with both simple and complex models, with sufficient resampling controls. However, it is mostly unfeasible, and in many cases practically impossible, to run sufficiently reliable test controls with locus or site-wise resampling under the most complex models. Breaking the long branches is an alternative remedy for LBA in which more taxa closely related to the long branches are incorporated. However, such an approach would increase the computation demand, and also the added taxa may increase the risks of orthology violations, especially if these taxa are newly sequenced and/or their data are loosely vetted.

For this, I aim to focus in future work on two points. First, continue improving the structure-partition-based model I have presented in **Paper II,** to make the model easier to use, better fit to diverse data, and more robust to the two types of errors I described, orthology violations and LBA. The concomitant reduction in computation demand should help open the field to less well-resourced labs, especially if I can design a user-friendly implementation of the model. Second, I hope to continue a paused project on methodically simulating and benchmarking the interactive effects of these errors on synthetic multi-locus data. The motivation of this project is to see how these errors interact and how different models handle them. LBA is often seen as some type of invidious artefact – hard to test for, hard to avoid, but potentially devastating in effect.

In terms of the eukaryote tree, many taxa were not included in my analyses. This was mostly because either they lacked sufficient data at the time, or the available data seemed to be problematic making the taxa phylogenetically unstable. Excluding these taxa allowed me to focus on very specific questions, *i.e.,* the relationships among the known major divisions of eukaryotes. These relationships will remain stable to the addition of new taxa, but new taxa may add new dimensions to the data by extending major groups, defining new ones, and possibly revealing even deeper branches in the tree. However, I have built a resilient data set and a strong tree that will make it easier to accurately place new taxa. It will also be easier to expand to other important outstanding questions, which should be easier to solve using the stable framework I provide. The ConJak and ConWin methods may also be useful in these studies, as extremely outliers and mosaicism are possible explanations for the instability of some taxa in the tree, particularly within the supergroup Diaphoretickes. Finally, the new structure-partition model should ease the

burden that the complex models have placed on the field of phylogenomics, which has limited the types of analyses that can be run and the people who can afford to run them. This should help make the field more accessible in general and allow more complete and flexible analyses of the data. This will be important as the size of the data sets and the complexity of the data and taxa involved will only increase as we try to reconstruct the full eukaryote tree of life.

Acknowledgment

First and foremost, I would like to thank my supervisor, Sandie. Every part of this work is done with your direct help and guidance. I am grateful for the opportunity you gave me. I always felt equipped with your support and encouragement which have enabled me to reach this point. My interest in phylogenetics started when I first took your course back in 2016, followed by my master thesis project in your Lab. I have been working with you for almost six years now, and I am eternally indebted to you for being professional, understanding, truthful and such a great mentor. Your impact on me far exceeds this work.

Fabien, you were not only my co-supervisor, but I always felt like we are coworkers' friends, and above that, you welcomed me to your lab meetings from which I benefitted a lot. I would like to extend my appreciation to all your lab members. Teaching with you in the phylogenetic course has been one of the greatest experiences I had during the past 4 years. I am fortunate to have been supervised by you and Sandie, great two scientists whose contributions have outlined a large portion of the tree of life.

Sanea, thank you for your sincere advice and guidance. Even after you moved to Malmö you continued to be supportive sharing your wisdom with such a cheering spirit and approachable personality.

Thank you, Hanna, Martin and Peter for all the scientific discussions, advice, follow-up, and support.

Special thanks to Karin and the administration team. It would have been much harder without your support.

I am thankful to everyone who has interacted with me during this journey.

Thank you Rokas. Knowing that you accepted to be my opponent has given a much-needed boost to my productivity in the last four months of work.

Finally, thank you Noor for helping me write the Swedish summary.

References

Abdo, Z., Minin, V. N., Joyce, P., & Sullivan, J. (2005). Accounting for Uncertainty in the Tree Topology Has Little Effect on the Decision-Theoretic Approach to Model Selection in Phylogeny Estimation. *Molecular Biology and Evolution*, *22*(3), 691–703. https://doi.org/10.1093/molbev/msi050

Aberer, A. J., Krompass, D., & Stamatakis, A. (2013). Pruning rogue taxa improves phylogenetic accuracy: An efficient algorithm and webservice. *Systematic Biology*, *62*(1), 162–166. https://doi.org/10.1093/sysbio/sys078

Adl, S. M., Bass, D., Lane, C. E., Lukeš, J., Schoch, C. L., Smirnov, A., Agatha, S., Berney, C., Brown, M. W., Burki, F., Cárdenas, P., Čepička, I., Chistyakova, L., del Campo, J., Dunthorn, M., Edvardsen, B., Eglit, Y., Guillou, L., Hampl, V., Heiss, A. A., Hoppenrath, M., James, T. Y., Karnkowska, A., Karpov, S., Kim, E., Kolisko, M., Kudryavtsev, A., Lahr, D. J. G., Lara, E., Le Gall, L., Lynn, D. H., Mann, D. G., Massana, R., Mitchell, E. A. D., Morrow, C., Park, J. S., Pawlowski, J. W., Powell, M. J., Richter, D. J., Rueckert, S., Shadwick, L., Shimano, S., Spiegel, F. W., Torruella, G., Youssef, N., Zlatogursky, V., & Zhang, Q. (2019). Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *Journal of Eukaryotic Microbiology*, *66*(1), 4–119. https://doi.org/10.1111/jeu.12691

Adl, S. M., Simpson, A. G. B., Lane, C. E., Lukeš, J., Bass, D., Bowser, S. S., Brown, M. W., Burki, F., Dunthorn, M., Hampl, V., Heiss, A., Hoppenrath, M., Lara, E., Gall, L. L., Lynn, D. H., McManus, H., Mitchell, E. A. D., Mozley-Stanridge, S. E., Parfrey, L. W., Pawlowski, J., Rueckert, S., Shadwick, L., Schoch, C. L., Smirnov, A., & Spiegel, F. W. (2012). The revised classification of eukaryotes. *Journal of Eukaryotic Microbiology*, *59*(5), 429–514. https://doi.org/10.1111/j.1550-7408.2012.00644.x

Al Jewari, C., & Baldauf, S. L. (2022). Conflict over the eukaryote root resides in strong outliers, mosaics and missing data sensitivity of site-specific (CAT) mixture models. *Systematic Biology*, syac029. https://doi.org/10.1093/sysbio/syac029

Alfaro, M. E., Zoller, S., & Lutzoni, F. (2003). Bayes or Bootstrap? A Simulation Study Comparing the Performance of Bayesian Markov Chain Monte Carlo Sampling and Bootstrapping in Assessing Phylogenetic Confidence. *Molecular Biology and Evolution*, *20*(2), 255–266. https://doi.org/10.1093/molbev/msg028

Baldauf, S. L., Roger, A. J., Wenk-Siefert, I., & Doolittle, W. F. (2000). A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science*, *290*(5493), 972–977. https://doi.org/10.1126/science.290.5493.972

Bapteste, E., Brinkmann, H., Lee, J. A., Moore, D. V., Sensen, C. W., Gordon, P., Duruflé, L., Gaasterland, T., Lopez, P., Müller, M., & Philippe, H. (2002).

The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(3), 1414–1419. https://doi.org/10.1073/pnas.032662799

Bergsten, J. (2005). A review of long-branch attraction. *Cladistics*, *21*(2), 163–193. https://doi.org/10.1111/j.1096-0031.2005.00059.x

Brown, J. R., & Doolittle, W. F. (1995). Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(7), 2441–2445. https://doi.org/10.1073/pnas.92.7.2441

Brown, M. W., Heiss, A. A., Kamikawa, R., Inagaki, Y., Yabuki, A., Tice, A. K., Shiratori, T., Ishida, K.-I., Hashimoto, T., Simpson, A. G. B., & Roger, A. J. (2018). Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group. *Genome Biology and Evolution*, *10*(2), 427–433. https://doi.org/10.1093/gbe/evy014

Brueckner, J., & Martin, W. F. (2020). Bacterial Genes Outnumber Archaeal Genes in Eukaryotic Genomes. *Genome Biology and Evolution*, *12*(4), 282–292. https://doi.org/10.1093/gbe/evaa047

Burki, F. (2014). The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harbor Perspectives in Biology*, *6*(5), a016147–a016147. https://doi.org/10.1101/cshperspect.a016147

Burki, F., Kaplan, M., Tikhonenkov, D. V., Zlatogursky, V., Minh, B. Q., Radaykina, L. V., Smirnov, A., Mylnikov, A. P., & Keeling, P. J. (2016). Untangling the early diversification of eukaryotes: A phylogenomic study of the evolutionary origins of centrohelida, haptophyta and cryptista. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1823), 20152802. https://doi.org/10.1098/rspb.2015.2802

Burki, F., Roger, A. J., Brown, M. W., & Simpson, A. G. B. (2020). The New Tree of Eukaryotes. *Trends in Ecology and Evolution*, *35*(1), 43–55. https://doi.org/10.1016/j.tree.2019.08.008

Burki, F., Shalchian-Tabrizi, K., Minge, M., Skjæveland, Å., Nikolaev, S. I., Jakobsen, K. S., & Pawlowski, J. (2007). Phylogenomics reshuffles the eukaryotic supergroups. *PLoS ONE*, *2*(8). https://doi.org/10.1371/journal.pone.0000790

Campbell, V., Legendre, P., & Lapointe, F. J. (2009). Assessing congruence among ultrametric distance matrices. *Journal of Classification*, *26*(1), 103–117. https://doi.org/10.1007/s00357-009-9028-x

Cannon, J. T., Vellutini, B. C., Smith, J., Ronquist, F., Jondelius, U., & Hejnol, A. (2016). Xenacoelomorpha is the sister group to Nephrozoa. *Nature*, *530*(7588), 89–93. https://doi.org/10.1038/nature16520

Caron, D. A., & Hu, S. K. (2019). Are We Overestimating Protistan Diversity in Nature? *Trends in Microbiology*, *27*(3), 197–205. https://doi.org/10.1016/j.tim.2018.10.009

Cavalier-Smith, T. (1989). Archaebacteria and Archezoa. *Nature*, *339*(6220), 100–101. https://doi.org/10.1038/339100a0

Cavalier-Smith, T. (2018). Kingdom Chromista and its eight phyla: a new synthesis emphasising periplastid protein targeting, cytoskeletal and periplastid evolution, and ancient divergences. *Protoplasma*, *255*(1), 297–357. https://doi.org/10.1007/s00709-017-1147-3

Cerón-Romero, M. A., Fonseca, M. M., de Oliveira Martins, L., Posada, D., & Katz, L. A. (2022). Phylogenomic Analyses of 2,786 Genes in 158 Lineages Support a Root of the Eukaryotic Tree of Life between Opisthokonts and All Other Lineages. *Genome Biology and Evolution*, *14*(8), evac119. https://doi.org/10.1093/gbe/evac119

Cotton, J. A., & McInerney, J. O. (2010). Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(40), 17252–17255. https://doi.org/10.1073/pnas.1000265107

Cummings, M. P., Handley, S. A., Myers, D. S., Reed, D. L., Rokas, A., & Winka, K. (2003). Comparing Bootstrap and Posterior Probability Values in the Four-Taxon Case. *Systematic Biology*, *52*(4), 477–487. https://doi.org/10.1080/10635150390218213

Day, W. H. E. (1986). Analysis of Quartet Dissimilarity Measures Between Undirected Phylogenetic Trees. *Systematic Zoology*, *35*(3), 325–333. JSTOR. https://doi.org/10.2307/2413385

Day, W. H. E., & Sankoff, D. (1986). Computational Complexity of Inferring Phylogenies by Compatibility. *Systematic Biology*, *35*(2), 224–229. https://doi.org/10.1093/sysbio/35.2.224

Dayhoff, M., Schwartz, R., & Orcutt, B. (1978). 22 a model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, *5*, 345–352.

De Vienne, D. M., Ollier, S., & Aguileta, G. (2012). Phylo-MCOA: A fast and efficient method to detect outlier genes and species in phylogenomics using multiple co-inertia analysis. *Molecular Biology and Evolution*, *29*(6), 1587–1598. https://doi.org/10.1093/molbev/msr317

Derelle, R., & Lang, B. F. (2012). Rooting the eukaryotic tree with mitochondrial and bacterial proteins. *Molecular Biology and Evolution*, *29*(4), 1277–1289. https://doi.org/10.1093/molbev/msr295

Derelle, R., Torruella, G., Klimeš, V., Brinkmann, H., Kim, E., Vlček, Č., Lang, B. F., & Eliáš, M. (2015). Bacterial proteins pinpoint a single eukaryotic root. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(7), E693–E699. https://doi.org/10.1073/pnas.1420657112

Dohmen, E., Klasberg, S., Bornberg-Bauer, E., Perrey, S., & Kemena, C. (2020). The modular nature of protein evolution: Domain rearrangement rates across eukaryotic life. *BMC Evolutionary Biology*, *20*(1). https://doi.org/10.1186/s12862-020-1591-0

Douady, C. J., Delsuc, F., Boucher, Y., Doolittle, W. F., & Douzery, E. J. P. (2003). Comparison of Bayesian and Maximum Likelihood Bootstrap Measures of Phylogenetic Reliability. *Molecular Biology and Evolution*, *20*(2), 248–254. https://doi.org/10.1093/molbev/msg042

Efron, B. (1981). Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods. *Biometrika*, *68*(3), 589. https://doi.org/10.2307/2335441

Embley, T. M. (2006). Multiple secondary origins of the anaerobic lifestyle in eukaryotes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1470), 1055–1067. https://doi.org/10.1098/rstb.2006.1844

Embley, T. M., & Hirt, R. P. (1998). Early branching eukaryotes? *Current Opinion in Genetics & Development*, *8*(6), 624–629. https://doi.org/10.1016/S0959-437X(98)80029-4

Erixon, P., Svennblad, B., Britton, T., & Oxelman, B. (2003). Reliability of Bayesian Posterior Probabilities and Bootstrap Frequencies in Phylogenetics. *Systematic Biology*, *52*(5), 665–673. https://doi.org/10.1080/10635150390235485

Estabrook, G. F., Mc Morris, F. R., & Meacham, C. A. (1985). Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic Zoology*, *34*(2), 193–200. https://doi.org/10.2307/sysbio/34.2.193

Felsenstein, J. (1978). Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. *Systematic Zoology*, *27*(4), 401–410. JSTOR. https://doi.org/10.2307/2412923

Feuda, R., Dohrmann, M., Pett, W., Philippe, H., Rota-Stabelli, O., Lartillot, N., Wörheide, G., & Pisani, D. (2017). Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals. *Current Biology*, *27*(24), 3864-3870.e4. https://doi.org/10.1016/j.cub.2017.11.008

Gabaldón, T. (2018). Relative timing of mitochondrial endosymbiosis and the "pre-mitochondrial symbioses" hypothesis: RELATIVE TIMING OF MITOCHONDRIAL SYMBIOSIS. *IUBMB Life*, *70*(12), 1188–1196. https://doi.org/10.1002/iub.1950

Gray, M. W. (2012). Mitochondrial Evolution. *Cold Spring Harbor Perspectives in Biology*, *4*(9), a011403. https://doi.org/10.1101/cshperspect.a011403

Gray, M. W. (2015). Mosaic nature of the mitochondrial proteome: Implications for the origin and evolution of mitochondria. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(33), 10133–10138. https://doi.org/10.1073/pnas.1421379112

Gray, M. W., Lang, B. F., & Burger, G. (2004). Mitochondria of Protists. *Annual Review of Genetics*, *38*(1), 477–524. https://doi.org/10.1146/annurev.genet.37.110801.142526

Hao, W., Richardson, A. O., Zheng, Y., & Palmer, J. D. (2010). Gorgeous mosaic of mitochondrial genes created by horizontal transfer and gene conversion. *Proceedings of the National Academy of Sciences*, *107*(50), 21576–21581. https://doi.org/10.1073/pnas.1016295107

He, D., Fiz-Palacios, O., Fu, C. J., Tsai, C. C., & Baldauf, S. L. (2014). An alternative root for the eukaryote tree of life. *Current Biology*, *24*(4), 465–470. https://doi.org/10.1016/j.cub.2014.01.036

He, D., Fu, C. J., & Baldauf, S. L. (2016). Multiple Origins of Eukaryotic cox15 Suggest Horizontal Gene Transfer from Bacteria to Jakobid Mitochondrial

DNA. *Molecular Biology and Evolution*, *33*(1), 122–133. https://doi.org/10.1093/molbev/msv201

Heiss, A. A., Kolisko, M., Ekelund, F., Brown, M. W., Roger, A. J., & Simpson, A. G. B. (2018). Combined morphological and phylogenomic re-examination of malawimonads, a critical taxon for inferring the evolutionary history of eukaryotes. *Royal Society Open Science*, *5*(4). https://doi.org/10.1098/rsos.171707

Høie, M. H., Kiehl, E. N., Petersen, B., Nielsen, M., Winther, O., Nielsen, H., Hallgren, J., & Marcatili, P. (2022). NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. *Nucleic Acids Research*, *50*(W1), W510–W515. https://doi.org/10.1093/nar/gkac439

Huelsenbeck, J. P., & Bull, J. J. (1996). A likelihood ratio test to detect conflicting phylogenetic signal. *Systematic Biology*, *45*(1), 92–98. https://doi.org/10.1093/sysbio/45.1.92

Huelsenbeck, J. P., Bull, J. J., & Cunningham, C. W. (1996). Combining data in phylogenetic analysis. *Trends in Ecology and Evolution*, *11*(4), 152–158. https://doi.org/10.1016/0169-5347(96)10006-9

Huelsenbeck, J. P., Larget, B., & Alfaro, M. E. (2004). Bayesian Phylogenetic Model Selection Using Reversible Jump Markov Chain Monte Carlo. *Molecular Biology and Evolution*, *21*(6), 1123–1133. https://doi.org/10.1093/molbev/msh123

Jeffroy, O., Brinkmann, H., Delsuc, F., & Philippe, H. (2006). Phylogenomics: the beginning of incongruence? *Trends in Genetics*, *22*(4), 225–231. https://doi.org/10.1016/j.tig.2006.02.003

Kamikawa, R., Kolisko, M., Nishimura, Y., Yabuki, A., Brown, M. W., Ishikawa, S. A., Ishida, K. I., Roger, A. J., Hashimoto, T., & Inagaki, Y. (2014). Gene content evolution in discobid mitochondria deduced from the phylogenetic position and complete mitochondrial genome of Tsukubamonas globosa. *Genome Biology and Evolution*, *6*(2), 306–315. https://doi.org/10.1093/gbe/evu015

Karnkowska, A., Vacek, V., Zubáčová, Z., Treitli, S. C., Petrželková, R., Eme, L., Novák, L., Žárský, V., Barlow, L. D., Herman, E. K., Soukal, P., Hroudová, M., Doležal, P., Stairs, C. W., Roger, A. J., Eliáš, M., Dacks, J. B., Vlček, Č., & Hampl, V. (2016). A eukaryote without a mitochondrial organelle. *Current Biology*, *26*(10), 1274–1284. https://doi.org/10.1016/j.cub.2016.03.053

Katz, L. A., Grant, J. R., Parfrey, L. W., & Burleigh, J. G. (2012). Turning the crown upside down: Gene tree parsimony roots the eukaryotic tree of life. *Systematic Biology*, *61*(4), 653–660. https://doi.org/10.1093/sysbio/sys026

Ke, D., Boissinot, M., Huletsky, A., Picard, F. J., Frenette, J., Ouellette, M., Roy, P. H., & Bergeron, M. G. (2000). Evidence for horizontal gene transfer in evolution of elongation factor Tu in enterococci. *Journal of Bacteriology*, *182*(24), 6913–6920. https://doi.org/10.1128/JB.182.24.6913-6920.2000

Keeling, P. J., & Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, *9*(8), 605–618. https://doi.org/10.1038/nrg2386

Kelchner, S. A., & Thomas, M. A. (2007). Model use in phylogenetics: nine key questions. *Trends in Ecology and Evolution*, *22*(2), 87–94. https://doi.org/10.1016/j.tree.2006.10.004

Kinene, T., Wainaina, J., Maina, S., & Boykin, L. M. (2016). Rooting Trees, Methods for. In R. M. Kliman (Ed.), *Encyclopedia of Evolutionary Biology* (pp. 489–493). Academic Press. https://doi.org/10.1016/B978-0-12-800049-6.00215-8

Kloesges, T., Popa, O., Martin, W., & Dagan, T. (2011). Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. *Molecular Biology and Evolution*, *28*(2), 1057–1074. https://doi.org/10.1093/molbev/msq297

Kück, P., & Wägele, J. W. (2016). Plesiomorphic character states cause systematic errors in molecular phylogenetic analyses: a simulation study. *Cladistics*, *32*(4), 461–478. https://doi.org/10.1111/cla.12132

Kupczok, A., Haeseler, A. V., & Klaere, S. (2008). An exact algorithm for the geodesic distance between phylogenetic trees. *Journal of Computational Biology*, *15*(6), 577–591. https://doi.org/10.1089/cmb.2008.0068

Kupczok, A., Schmidt, H. A., & von Haeseler, A. (2010). Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. *Algorithms for Molecular Biology*, *5*(1), 37. https://doi.org/10.1186/1748-7188-5-37

Kurland, C. G., & Andersson, S. G. E. (2000). Origin and Evolution of the Mitochondrial Proteome. *Microbiology and Molecular Biology Reviews*, *64*(4), 786–820. https://doi.org/10.1128/mmbr.64.4.786-820.2000

Lane, N., & Martin, W. (2010). The energetics of genome complexity. *Nature*, *467*(7318), 929–934. https://doi.org/10.1038/nature09486

Lartillot, N., & Philippe, H. (2004). A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Molecular Biology and Evolution*, *21*(6), 1095–1109. https://doi.org/10.1093/molbev/msh112

Lartillot, N., Rodrigue, N., Stubbs, D., & Richer, J. (2013). PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. *Systematic Biology*, *62*(4), 611–615. https://doi.org/10.1093/sysbio/syt022

Laumer, C. E., Bekkouche, N., Kerbl, A., Goetz, F., Neves, R. C., Sørensen, M. V., Kristensen, R. M., Hejnol, A., Dunn, C. W., Giribet, G., & Worsaae, K. (2015). Spiralian Phylogeny Informs the Evolution of Microscopic Lineages. *Current Biology*, *25*(15), 2000–2006. https://doi.org/10.1016/j.cub.2015.06.068

Lax, G., Eglit, Y., Eme, L., Bertrand, E. M., Roger, A. J., & Simpson, A. G. B. (2018). Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. *Nature*, *564*(7736), 410–414. https://doi.org/10.1038/s41586-018-0708-8

Le, S. Q., & Gascuel, O. (2008). An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution*, *25*(7), 1307–1320. https://doi.org/10.1093/molbev/msn067

Le, S. Q., & Gascuel, O. (2010). Accounting for Solvent Accessibility and Secondary Structure in Protein Phylogenetics Is Clearly Beneficial. *Systematic Biology*, *59*(3), 277–287. https://doi.org/10.1093/sysbio/syq002

Le, S. Q., Gascuel, O., & Lartillot, N. (2008). Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*, *24*(20), 2317–2323. https://doi.org/10.1093/bioinformatics/btn445

Le, S. Q., Lartillot, N., & Gascuel, O. (2008). Phylogenetic mixture models for proteins. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1512), 3965–3976. https://doi.org/10.1098/rstb.2008.0180

Lee, S. C., Corradi, N., Byrnes, E. J., Torres-Martinez, S., Dietrich, F. S., Keeling, P. J., & Heitman, J. (2008). Microsporidia Evolved from Ancestral Sexual Fungi. *Current Biology*, *18*(21), 1675–1679. https://doi.org/10.1016/j.cub.2008.09.030

Leger, M. M., Eme, L., Hug, L. A., & Roger, A. J. (2016). Novel Hydrogenosomes in the Microaerophilic Jakobid Stygiella incarcerata. *Molecular Biology and Evolution*, *33*(9), 2318–2336. https://doi.org/10.1093/molbev/msw103

Leigh, J. W., Schliep, K., Lopez, P., & Bapteste, E. (2011). Let Them Fall Where They May: Congruence Analysis in Massive Phylogenetically Messy Data Sets. *Molecular Biology and Evolution*, *28*(10), 2773–2785. https://doi.org/10.1093/molbev/msr110

Leigh, J. W., Susko, E., Baumgartner, M., & Roger, A. J. (2008). Testing congruence in phylogenomic analysis. *Systematic Biology*, *57*(1), 104–115. https://doi.org/10.1080/10635150801910436

Li, H., Shao, R., Song, N., Song, F., Jiang, P., Li, Z., & Cai, W. (2015). Higher-level phylogeny of paraneopteran insects inferred from mitochondrial genome sequences. *Scientific Reports*, *5*(1), 8527. https://doi.org/10.1038/srep08527

Li, Y., Shen, X.-X., Evans, B., Dunn, C. W., & Rokas, A. (2021). Rooting the Animal Tree of Life. *Molecular Biology and Evolution*, *38*(10), 4322–4333. https://doi.org/10.1093/molbev/msab170

Maguire, F., Henriquez, F. L., Leonard, G., Dacks, J. B., Brown, M. W., & Richards, T. A. (2014). Complex patterns of gene fission in the eukaryotic folate biosynthesis pathway. *Genome Biology and Evolution*, *6*(10), 2709–2720. https://doi.org/10.1093/gbe/evu213

Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2015). Archaeal Clusters of Orthologous Genes (arCOGs): An Update and Application for Analysis of Shared Features between Thermococcales, Methanococcales, and Methanobacteriales. *Life (Basel, Switzerland)*, *5*(1), 818–840. https://doi.org/10.3390/life5010818

Makiuchi, T., & Nozaki, T. (2014). Highly divergent mitochondrion-related organelles in anaerobic parasitic protozoa. *Biochimie*, *100*(1), 3–17. https://doi.org/10.1016/j.biochi.2013.11.018

Martin, W. F. (2018). Eukaryote lateral gene transfer is Lamarckian. *Nature Ecology and Evolution*, *2*(5), 754. https://doi.org/10.1038/s41559-018-0521-7

Minh, B. Q., Dang, C. C., Vinh, L. S., & Lanfear, R. (2021). QMaker: Fast and Accurate Method to Estimate Empirical Models of Protein Evolution. Systematic Biology, 70(5), 1046–1060. https://doi.org/10.1093/sysbio/syab010

Minin, V., Abdo, Z., Joyce, P., & Sullivan, J. (2003). Performance-Based Selection of Likelihood Models for Phylogeny Estimation. *Systematic Biology*, *52*(5), 674–683. https://doi.org/10.1080/10635150390235494

Palmer, J. D., & Delwiche, C. F. (1996). Second-hand chloroplasts and the case of the disappearing nucleus. *Proceedings of the National Academy of Sciences*, *93*(15), 7432–7435. https://doi.org/10.1073/pnas.93.15.7432

Penny, D., Foulds, L. R., & Hendy, M. D. (1982). Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. In *Nature* (Vol. 297, Issue 5863, pp. 197–200). https://doi.org/10.1038/297197a0

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., & Baurain, D. (2011). Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biology*, *9*(3), e1000602. https://doi.org/10.1371/journal.pbio.1000602

Philippe, H., De Vienne, D. M., Ranwez, V., Roure, B., Baurain, D., & Delsuc, F. (2017). Pitfalls in supermatrix phylogenomics. *European Journal of Taxonomy*, *2017*(283), 1–25. https://doi.org/10.5852/ejt.2017.283

Philippe, H., Germot, A., & Moreira, D. (2000). The new phylogeny of eukaryotes. *Current Opinion in Genetics and Development*, *10*(6), 596–601. https://doi.org/10.1016/S0959-437X(00)00137-4

Pittis, A. A., & Gabaldón, T. (2016). Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature*, *531*(7592), 101–104. https://doi.org/10.1038/nature16941

Planet, P. J., & Sarkar, I. N. (2005). mILD: A tool for constructing and analyzing matrices of pairwise phylogenetic character incongruence tests. *Bioinformatics*, *21*(24), 4423–4424. https://doi.org/10.1093/bioinformatics/bti744

Rice, D. W., Alverson, A. J., Richardson, A. O., Young, G. J., Sanchez-Puerta, M. V., Munzinger, J., Barry, K., Boore, J. L., Zhang, Y., DePamphilis, C. W., Knox, E. B., & Palmer, J. D. (2013). Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm Amborella. *Science*, *342*(6165), 1468–1473. https://doi.org/10.1126/science.1246275

Ripplinger, J., & Sullivan, J. (2008). Does Choice in Model Selection Affect Maximum Likelihood Analysis? *Systematic Biology*, *57*(1), 76–85. https://doi.org/10.1080/10635150801898920

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, *53*(1–2), 131–147. https://doi.org/10.1016/0025-5564(81)90043-2

Sharma, P. P., Kaluziak, S. T., Pérez-Porro, A. R., González, V. L., Hormiga, G., Wheeler, W. C., & Giribet, G. (2014). Phylogenomic Interrogation of Arachnida Reveals Systemic Conflicts in Phylogenetic Signal. *Molecular*

Biology and Evolution, *31*(11), 2963–2984. https://doi.org/10.1093/molbev/msu235

Shen, X. X., Hittinger, C. T., & Rokas, A. (2017). Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology and Evolution*, *1*(5), 1–10. https://doi.org/10.1038/s41559-017-0126

Sibbald, S. J., Eme, L., Archibald, J. M., & Roger, A. J. (2020). Lateral Gene Transfer Mechanisms and Pan-genomes in Eukaryotes. *Trends in Parasitology*, *36*(11), 927–941. https://doi.org/10.1016/j.pt.2020.07.014

Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., Roure, B., Satoh, N., Quéinnec, É., Ereskovsky, A., Lapébie, P., Corre, E., Delsuc, F., King, N., Wörheide, G., & Manuel, M. (2017). A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. *Current Biology*, *27*(7), 958–967. https://doi.org/10.1016/j.cub.2017.02.031

Simpson, A. G. B. (2003). Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota). *International Journal of Systematic and Evolutionary Microbiology, 53*(6), 1759–1777. https://doi.org/10.1099/ijs.0.02578-0

Simpson, A. G. B., & Patterson, D. J. (1999). The ultrastructure of Carpediemonas membranifera (Eukaryota) with reference to the "excavate hypothesis." *European Journal of Protistology*, *35*(4), 353–370. https://doi.org/10.1016/S0932-4739(99)80044-3

Smith, S. A., Walker-Hale, N., Walker, J. F., & Brown, J. W. (2020). Phylogenetic Conflicts, Combinability, and Deep Phylogenomics in Plants. *Systematic Biology*, *69*(3), 579–592. https://doi.org/10.1093/sysbio/syz078

Sogin, M., Gunderson, J., Elwood, H., Alonso, R., & Peattie, D. (1989). Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from Giardia lamblia. *Science*, *243*(4887), 75–77. https://doi.org/10.1126/science.2911720

Song, F., Li, H., Jiang, P., Zhou, X., Liu, J., Sun, C., Vogler, A. P., & Cai, W. (2016). Capturing the Phylogeny of Holometabola with Mitochondrial Genome Data and Bayesian Site-Heterogeneous Mixture Models. *Genome Biology and Evolution*, *8*(5), 1411–1426. https://doi.org/10.1093/gbe/evw086

Spielman, S. J. (2020). Relative Model Fit Does Not Predict Topological Accuracy in Single-Gene Protein Phylogenetics. *Molecular Biology and Evolution*, *37*(7), 2110–2123. https://doi.org/10.1093/molbev/msaa075

Stechmann, A., & Cavalier-Smith, T. (2002). Rooting the eukaryote tree by using a derived gene fusion. *Science*, *297*(5578), 89–91. https://doi.org/10.1126/science.1071196

Strassert, J. F. H., Jamy, M., Mylnikov, A. P., Tikhonenkov, D. V., & Burki, F. (2019). New phylogenomic analysis of the enigmatic phylum Telonemia further resolves the eukaryote tree of life. *Molecular Biology and Evolution*, *36*(4), 757–765. https://doi.org/10.1093/molbev/msz012

Struck, T. H., Wey-Fabrizius, A. R., Golombek, A., Hering, L., Weigert, A., Bleidorn, C., Klebow, S., Iakovenko, N., Hausdorf, B., Petersen, M., Kück, P., Herlyn, H., & Hankeln, T. (2014). Platyzoan Paraphyly Based on

Phylogenomic Data Supports a Noncoelomate Ancestry of Spiralia. *Molecular Biology and Evolution*, *31*(7), 1833–1849. https://doi.org/10.1093/molbev/msu143

Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M., & Dessimoz, C. (2015). Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. *Systematic Biology*, *64*(5), 778–791. https://doi.org/10.1093/sysbio/syv033

Tria, F. D. K., Brueckner, J., Skejo, J., Xavier, J. C., Kapust, N., Knopp, M., Wimmer, J. L. E., Nagies, F. S. P., Zimorski, V., Gould, S. B., Garg, S. G., & Martin, W. F. (2021). Gene Duplications Trace Mitochondria to the Onset of Eukaryote Complexity. *Genome Biology and Evolution*, *13*(5), evab055. https://doi.org/10.1093/gbe/evab055

Wägele, J. W., Letsch, H., Klussmann-Kolb, A., Mayer, C., Misof, B., & Wägele, H. (2009). Phylogenetic support values are not necessarily informative: The case of the Serialia hypothesis (a mollusk phylogeny). *Frontiers in Zoology*, *6*(1), 12. https://doi.org/10.1186/1742-9994-6-12

Wang, H. C., Minh, B. Q., Susko, E., & Roger, A. J. (2018). Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Systematic Biology*, *67*(2), 216–235. https://doi.org/10.1093/sysbio/syx068

Wheeler, W. C. (1990). Nucleic Acid Sequence Phylogeny and Random Outgroups. *Cladistics*, *6*(4), 363–367. https://doi.org/10.1111/j.1096-0031.1990.tb00550.x

Whelan, N. V., & Halanych, K. M. (2017). Who Let the CAT Out of the Bag? Accurately Dealing with Substitutional Heterogeneity in Phylogenomic Analyses. *Systematic Biology*, *66*(2), 232–255. https://doi.org/10.1093/sysbio/syw084

Wideman, J. G., Gawryluk, R. M. R., Gray, M. W., & Dacks, J. B. (2013). The ancient and widespread nature of the ER-mitochondria encounter structure. *Molecular Biology and Evolution*, *30*(9), 2044–2049. https://doi.org/10.1093/molbev/mst120

Williams, T. A. (2014). Evolution: Rooting the eukaryotic tree of life. *Current Biology*, *24*(4), R151–R152. https://doi.org/10.1016/j.cub.2014.01.026

Williams, W. T., & Clifford, H. T. (1971). on the Comparison of Two Classifications of the Same Set of Elements. *Taxon*, *20*(4), 519–522. https://doi.org/10.2307/1218253

Wu, M., Chatterji, S., & Eisen, J. A. (2012). Accounting for alignment uncertainty in phylogenomics. *PLoS ONE*, *7*(1), e30288. https://doi.org/10.1371/journal.pone.0030288

Young, A. D., & Gillung, J. P. (2020). Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics. *Systematic Entomology*, *45*(2), 225–247. https://doi.org/10.1111/syen.12406

Zimorski, V., Mentel, M., Tielens, A. G. M., & Martin, W. F. (2019). Energy metabolism in anaerobic eukaryotes and Earth's late oxygenation. *Free Radical Biology and Medicine*, *140*, 279–294. https://doi.org/10.1016/j.freeradbiomed.2019.03.030

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 2191

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-484580

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2022