

Genome-wide evolutionary analyses indicate a red, not a green, ancestry of the apicomplexan plastid.

Oliver Deusch¹, Geoffrey I. McFadden² and William Martin¹

¹Institute of Botany, University of Düsseldorf, 40225 Düsseldorf, Germany.
²Plant Cell Biology Research Centre, School of Botany, University of Melbourne, Australia.

The origin of the apicomplexan plastid is still debated, analyses of individual genes that would address the issue having produced conflicting results. However, analyses of individual proteins are subject to sampling errors because of the limited number of sites for comparison. Additionally, phylogenetic analyses of apicoplast proteins are biased by the high A+T content and convergent codon usage. Therefore we have investigated nuclear encoded *Plasmodium* proteins that have homologues in eight sequenced eukaryotic genomes, including representatives of the green and red primary plastid lineages. We examined both concatenated data sets and individual alignments in

excess of 200,000 amino acid site patterns from 600 different protein alignments. In individual comparisons we scored not only the strongest signal in each data set, as tree-building methods do, but also scored weaker, competing signals detectable with network methods. The signal supporting a red algal ancestry of the apicomplexan plastid is detected more frequently and is about twice as strong as the competing green signal, providing evidence in favour of a red algal ancestry for the secondary plastid of the apicomplexan lineage.

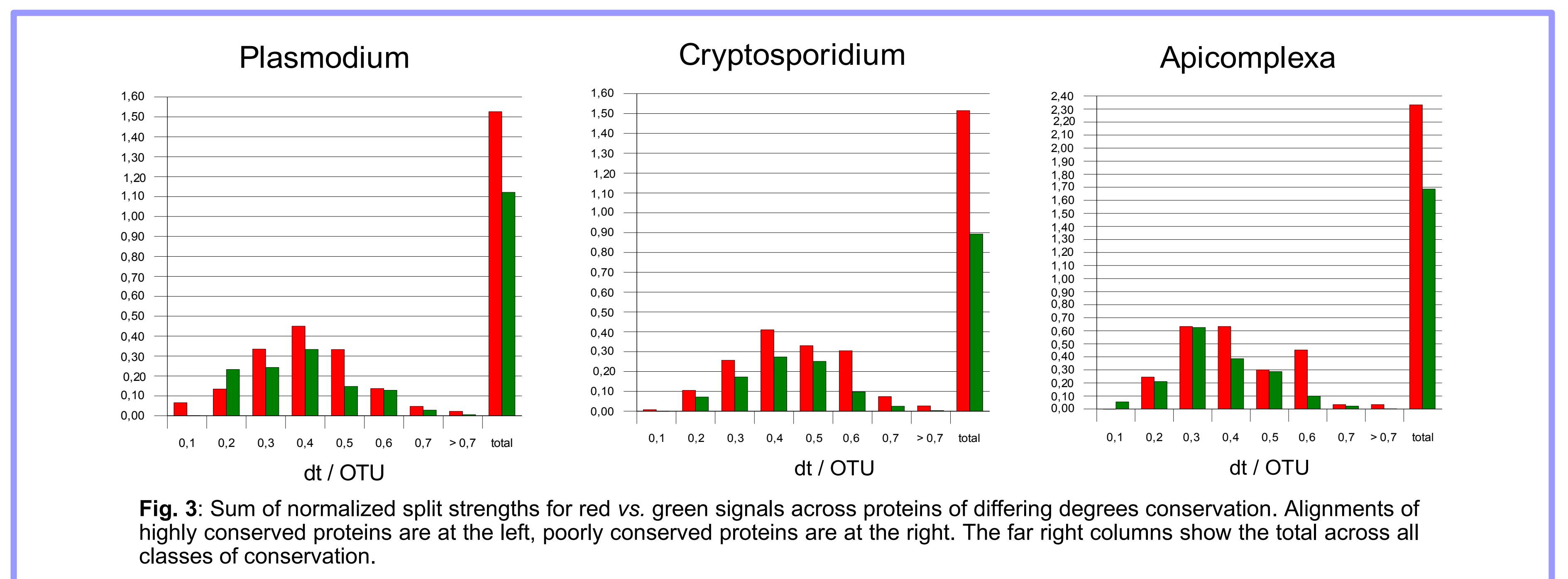
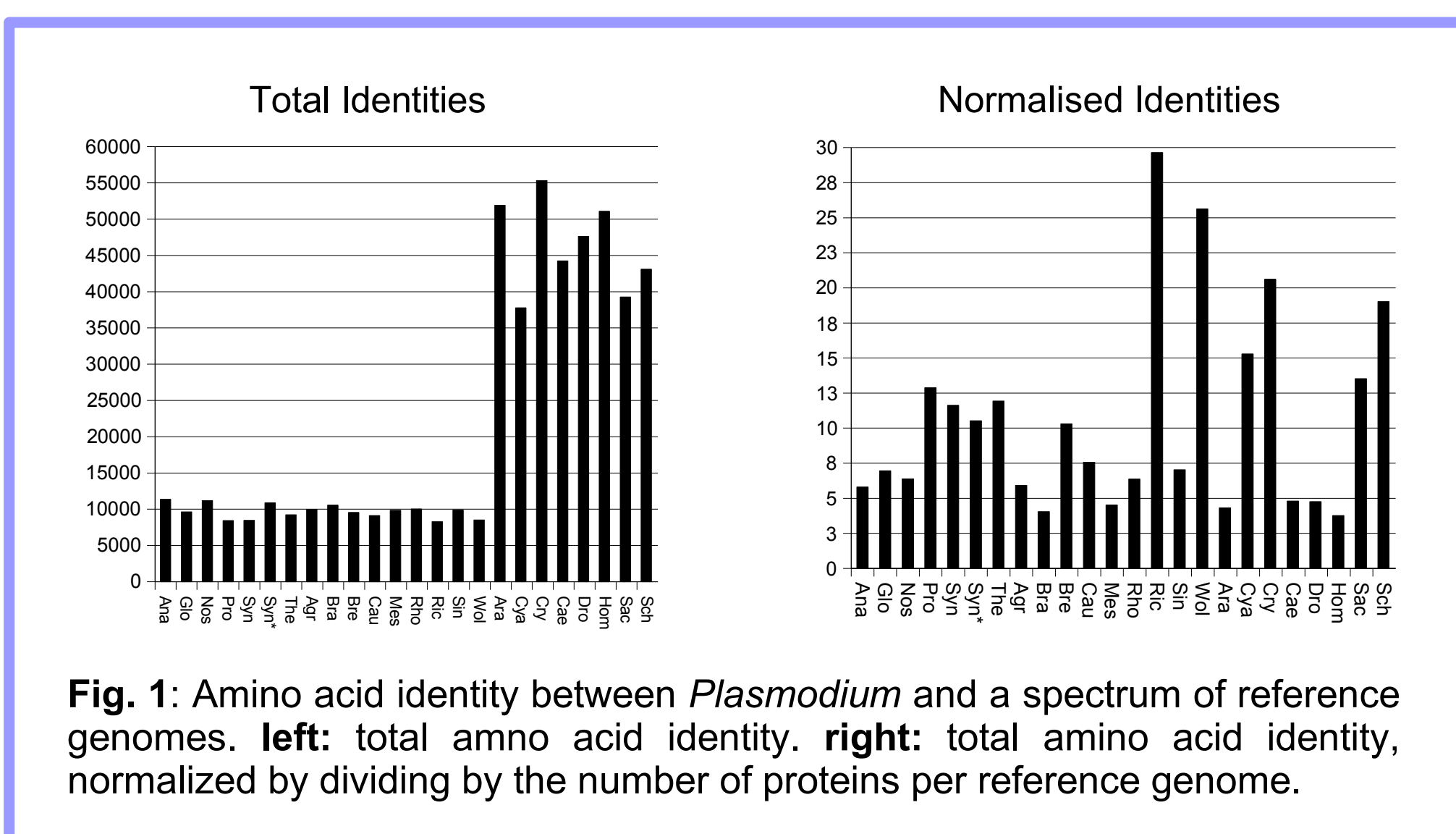


Table 1: Sum of normalized split strengths for nine taxa. For example, (*Plasmodium*, red) designates the split linking *Plasmodium* and the red representative to the exclusion of the other seven taxa.

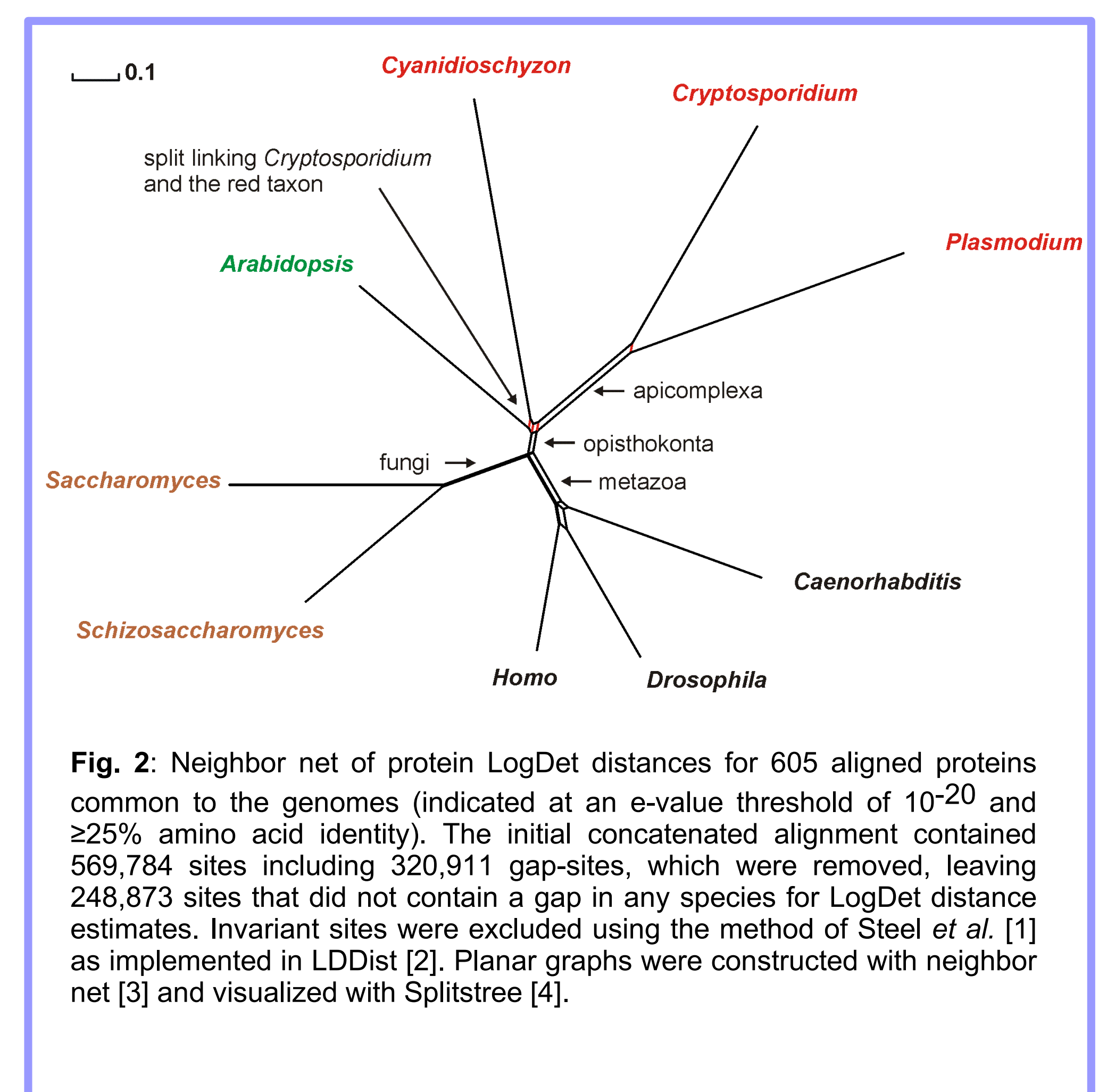
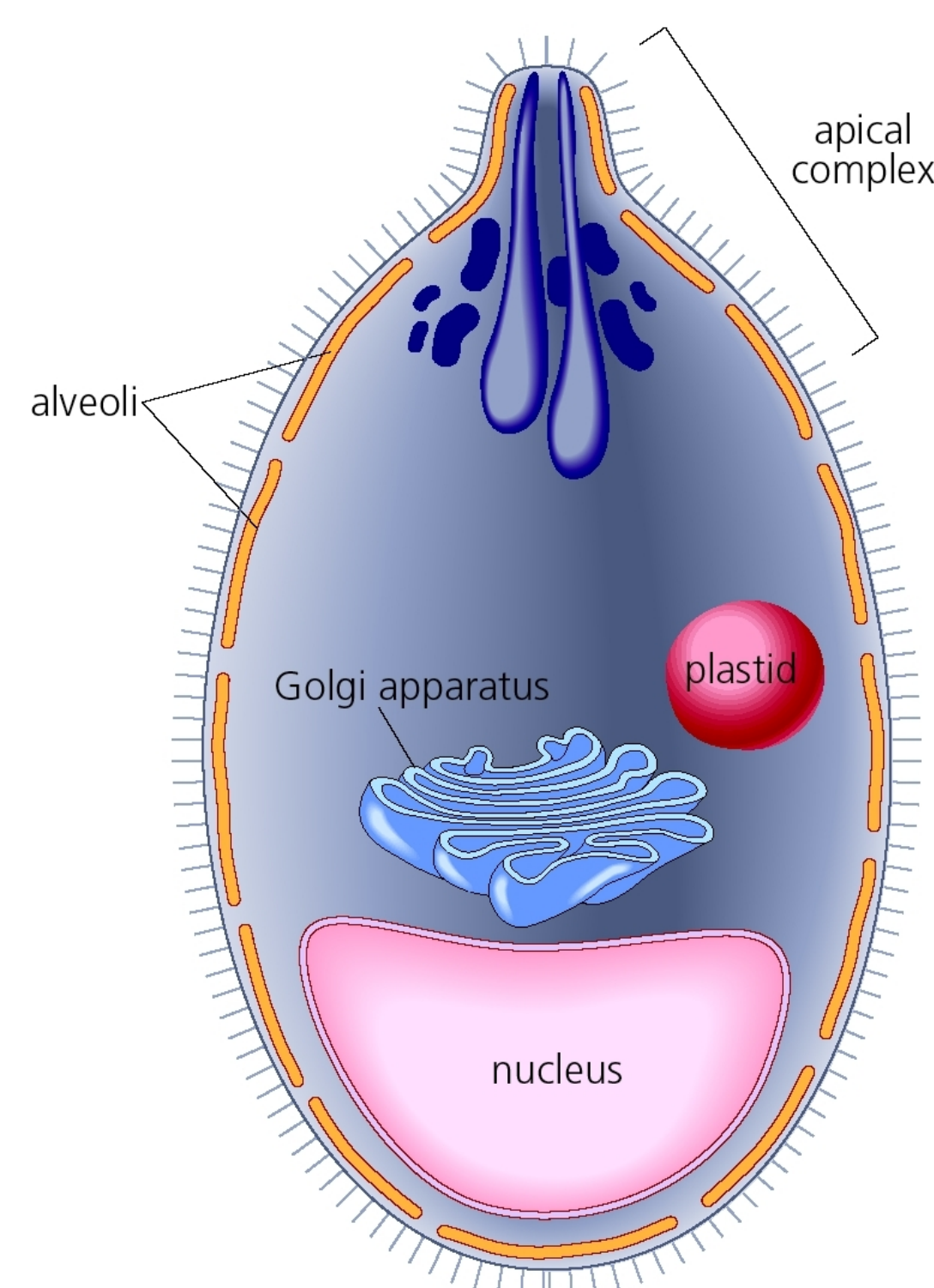
Split	sum of split strengths	number of occurrences ¹
(<i>Plasmodium</i> , red)	1,53	107
(<i>Plasmodium</i> , green)	1,12	105
(<i>Cryptosporidium</i> , red)	1,52	109
(<i>Cryptosporidium</i> , green)	0,89	92
(apicomplexa, red)	2,33	174
(apicomplexa, green)	1,69	125
two fungi	16,69	488
three animals	10,04	406
two algae ²	3,12	198
apicomplexa	19,15	481
opisthokonta	2,7	173
(animals, plants) ³	0,93	94
(fungi, plants) ³	0,59	40

Notes:

¹number of times that the split was sufficiently strong to be detected by Nnet in individual analyses of 590 proteins common to the 9 OTU data set. Although 605 proteins were common to all nine taxa, LDDist would not compute a distance for 15 data sets in individual analyses.

²the split linking the red taxon to *Arabidopsis*.

³splits competing with the opisthokont split.



Results

Homologous proteins to 5,267 *Plasmodium* proteins were identified using the BLASTP algorithm. The search set comprised the proteomes of eight eukaryotes including *Arabidopsis thaliana* as representative of the green and *Cyanidioschyzon merolae* as representative of the red primary plastid lineage. The proteome of a second apicomplexan taxon, *Cryptosporidium parvum*, is available, so it was included in our search.

The highest total amino acid identity was found between *Plasmodium* and *Cryptosporidium* (Fig 1a). *Plasmodium* showed 37 percent more identity with *Arabidopsis* than *Cyanidioschyzon*, but genome sizes of *Cyanidioschyzon* and *Arabidopsis* are very different (4,772 vs. 28,860 proteins). When the total amino acid identity was normalised by proteome size, the red representative showed three times more identity than the green one (Fig 1b). However, this normalisation strongly favours small genomes.

A set of 605 *Plasmodium* proteins was found to have homologues in all eight eukaryotic databases. Each protein was aligned with its counterparts. The alignments were concatenated and a phylogenetic network analysis was carried out (Fig. 2). All major groups could be resolved indicating that this approach detected biologically reasonable results. A strong split linked *Cryptosporidium* and *Cyanidioschyzon*. However, there was no signal linking *Plasmodium* or both apicomplexa to the red or green primary plastid lineages. Therefore individual analyses were carried out (Table 1). Association of the apicomplexa with the red lineage was 39 percent more frequent, with 38 percent stronger splits, than with the green lineage. Split strengths were sorted in intervals representing classes of protein conservation (Fig. 3). The red signal dominated the green one in nearly all classes, indicating that signals are not biased due to conservation.

Summary / Conclusions

Whole genomes of nine eukaryotes were analysed to elucidate the origin of the apicoplast, providing the largest dataset used to date to address this question. Phylogenetic analyses of concatenated and individual protein datasets favour a red algal apicoplast progenitor. When it comes to overall amino acid identity the picture is not that clear. In total, the 5,267 *Plasmodium* proteins show more identity with the proteins from *Arabidopsis* than with those from *Cyanidioschyzon*. Normalised datasets linked *Plasmodium* to the red plastid lineage. However, this procedure strongly favours small genomes and therefore is biased in favour of the red plastid lineage. When a green algal nuclear genome similar in size to *Cyanidioschyzon* becomes available this analysis should be repeated.

Taken together, our analyses support a red algal apicoplast ancestry as proposed by Cavalier-Smith in his chromalveolate hypothesis [5], which posits that a single endosymbiosis between a phagotroph biciliate and a red alga gave rise to a broad range of organisms dubbed chromalveolates.

References

- Steel M, Huson D, Lockhart PJ. (2000). Invariable sites models and their use in phylogeny reconstruction. *Syst Biol*, 49:225-232.
- Thollessen M. (2004). LDDist: a Perl module for calculating LogDet pair-wise distances for protein and nucleotide sequences. *Bioinformatics*, 20:416-418.
- Bryant D, Moulton V. (2002). NeighborNet: an agglomerative method for the construction of planar phylogenetic networks. *Proceedings of WABI*.
- Huson DH. (1998). SplitsTree: Analyzing and visualizing evolutionary data. *Bioinformatics*, 14(1):68-73.
- Cavalier-Smith T. (1999). Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J Eukaryot Microbiol*, 46 347-366.