Phylogenomics reveal a robust fungal tree of life

Eiko E. Kuramae[1], Vincent Robert[1], Berend Snel[2], Michael Weiß[3] & Teun Boekhout[1]

[1]Yeast Research, Centraalbureau voor Schimmelcultures, Utrecht, The Netherlands; [2]Nijmegen Center for Molecular Life Sciences, University Medical Center St Radboud, pa CMBI, Nijmegen, The Netherlands; and [3]Spezielle Botanik und Mykologie, Universität Tübingen, Tübingen, Germany

Correspondence: Eiko E. Kuramae, Yeast Research, Centraalbureau voor Schimmelcultures, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands. Tel.: +31 30 2122600; fax: +33 302512097; e-mail: kuramae@cbs.knaw.nl

Abstract

Our understanding of the tree of life (TOL) is still fragmentary. Until recently, molecular phylogeneticists have built trees based on ribosomal RNA sequences and selected protein sequences, which, however, usually suffered from lack of support for the deeper branches and inconsistencies probably due to limited subsampling of the entire genome. Now, phylogenetic hypotheses can be based on the analysis of full genomes. We used available complete genome data as well as the eukaryote orthologous group (KOG) proteins to reconstruct with confidence basal branches of the fungal TOL. Phylogenetic analysis of a core of 531 KOGs shared among 21 fungal genomes, three animal genomes and one plant genome showed a single tree with high support resulting from four different methods of phylogenetic reconstruction. The single tree that we inferred from our dataset showed excellent nodal support for each branch, suggesting that it reflects the true phylogenetic relationships of the species involved.

Introduction

The amount of molecular data available has dramatically increased over the last few years. These data are very suitable for testing traditional phylogenetic hypotheses. Molecular phylogenetic inference has become the key technique for reconstructing the evolution of species. More recently, with the advantage of many genomes being sequenced and available, phylogenetic hypotheses can be tested using full genome information. Genome data have been successfully applied to reconstruct phylogenetic relationships among prokaryotes (Wolf *et al.*, 2001), and full genome analysis may also enable us to precisely reconstruct eukaryotic diversification.

The fungal kingdom is one of the main domains of eukaryotic diversification (Hawksworth, 2001). Fungi show wide variations in morphology, physiology, lifestyles, and ecology (McLaughlin *et al.*, 2001). Based on morphology, physiology and nuclear ribosomal DNA (rDNA) sequences, five main groups (*Chytridiomycota*, *Glomeromycota*, *Ascomycota*, *Basidiomycota*, *Zygomycota*) have been distinguished (Berbee & Taylor, 2001; Schüßler *et al.*, 2001; Lutzoni *et al.*, 2004). Fungi are one of the best studied groups of eukaryotes with respect to the number of species with fully sequenced genomes (Galagan *et al.*, 2005), and are

therefore ideally suited for the exploration of the use of genome information to test evolutionary hypotheses.

Several methods have been proposed to build genome trees, especially for prokaryotic genomes. Two of these are scoring the presence and absence of orthologous genes, or determining the order in which genes occur (Snel *et al.*, 1999; Wolf *et al.*, 2001), but both methods ignore the phylogenetic information contained in the genes themselves. Another method is the phylogenetic analysis of concatenated sequences of genes, as has been used to resolve incongruences in the phylogeny of bacteria and archaea (Wolf *et al.*, 2001), species of *Saccharomyces* (Rokas *et al.*, 2003), *Pseudocoelomata* and *Coelomata* (Wolf *et al.*, 2004), and plants (Sanderson *et al.*, 2003). Phylogenetic studies of a broader range of fungi based on one or a few genes suffer from lack of support and topological inconsistencies (Kurtzman & Robnett, 2003; Tehler *et al.*, 2003; Lutzoni *et al.*, 2004).

The aim of this article is to test current phylogenetic hypotheses of the fungal tree of life (TOL) using available complete fungal genome sequences. In summary, our analysis consisted first of assigning protein sequences to eukaryote orthologous groups (KOGs) and thereby identifying orthologous proteins shared by 25 eukaryotes. This was followed by phylogenetic analyses based on: (1) the presence

or absence of KOGs; and (2) sequence analysis using four different methods of phylogenetic reconstruction. We also determined the number of KOG families with similar evolutionary signals necessary to resolve different clades of the fungal phylogenetic tree.

Materials and methods

Genomes

We analyzed the following fungal genomes (Table 1): *Ashbya gossypii* (Ago), *Aspergillus nidulans* (Ani), *Candida albicans* (Cal), *Candida glabrata* (Cgl), *Cryptococcus neoformans* (Cne), *Debaryomyces hansenii* (Dha), *Fusarium graminearum* (Fgr), *Kluyveromyces lactis* (Kla), *Magnaporthe grisea* (Mgr), *Neurospora crassa* (Ncr), *Phanerochaete chrysosporium* (Pch), *Saccharomyces bayanus* (Sba), *Saccharomyces castellii* (Sca), *Saccharomyces cerevisiae* (Sce), *Saccharomyces kluyveri* (Skl), *Saccharomyces kudriavzevii* (Sku), *Saccharomyces mikatae* (Smi), *Saccharomyces paradoxus* (Spa), *Schizosaccharomyces pombe* (Spo), *Ustilago maydis* (Uma), *Yarrowia lipolytica* (Yli), and other representative eukaryotic genomes, i.e. *Arabidopsis thaliana* (Ath), *Caenorhabditis elegans* (Cel), *Homo sapiens* (Hsp) and *Drosophila melanogaster* (Dme).

Assessment of orthology

The group orthology framework presented in the KOG database (Tatusov *et al.*, 2003) was the basis of our analyses. The KOGs of Ath, Cel, Dme, Hsp, Sce and Spo were obtained from the KOG database ftp://ftp.ncbi.nih.gov/pub/COG/KOG/. Nineteen proteomes (Table 1; Ani, Cal, Cne, Fgr, Mgr, Ncr, Sca, Sku, Spa, Sba, Skl, Smi, Uma, Ago, Cgl, Dha, Kla, Yli and Pch) were assigned for orthologies using the STRING program (Snel *et al.*, 2000). The STRING procedure basically follows the cognitor that is available at the NCBI COG website. It works by making a BLAST search of each gene of each genome against a database of protein sequences whose KOG and KOG domain assignment are known from the KOG database. Subsequently it is determined for each region of the query protein to which KOG it is most similar and if the similarity is significant. The analysis is on a per region basis to account for fusion proteins.

Phylogenetic analysis

Rooting the fungal phylogenetic tree

To root our phylogenetic trees, we selected eight KOGs shared by eukaryotes and one archaea and one bacterium

Table 1. Genome sources, genome size (Mb), number and percentage of KOGs used in the study

Genome	Strain	Number of KOGs	Genome size (mb)	% KOG used in this study	Location
Arabidopsis thaliana		3286	125	18.30	NCBI
Ashbya gossypii	ATCC 10895	2592	9.2	23.30	EBI
Aspergillus nidulans	FGSC A4	2982	31	20.15	Whitehead
Caenorhabditis elegans		4235	100	14.19	Sanger
Candida albicans	SC5314	2636	15	22.80	Stanford
Candida glabrata	CBS138	2505	13	24.00	Genolevures
Cryptococcus neoformans	JEC21	2856	24	21.02	TIGR
Debaryomyces hansenii	CBS767	2760	12–13	21.78	Genolevures
Drosophila melanogaster		4352	120	13.81	NCBI
Fusarium graminearum	PH-1 (NRRL 31084)	3063	36	19.62	Whitehead
Homo sapiens		4597	3200	13.07	NCBI
Kluyveromyces lactis	CLIB210	2596	11.4	23.15	Genolevures
Magnaporthe grisea	70-15	2917	40	20.60	Whitehead
Neurospora crassa	N-150	2962	40	20.29	Whitehead
Phanerochaete chrysosporium	RP78	2945	29.9–30	20.41	JGI
Saccharomyces bayanus	MCYC623	2560	12	23.48	Stanford
Saccharomyces castellii	NRRL Y-12630	2390	10.2	25.15	Stanford
Saccharomyces kluyveri	NRRL Y-12651	1747	10.2	30.40	Stanford
Saccharomyces kudriavzevii	IFO 1802	1855	10.6	32.40	Stanford
Saccharomyces mikatae	IFO1815	2557	12	23.50	Stanford
Saccharomyces paradoxus	NRRLY-17217	2592	12	23.19	Stanford
Saccharomyces cerevisiae	S288C	2668	13	22.53	Stanford
Schizosaccharomyces pombe	Urs Leupold 972 h⁻	2762	14	21.76	Sanger
Ustilago maydis	521	2850	20	21.09	Whitehead
Yarrowia lipolytica	CLIB99	2699	20–21	22.27	Genolevures

Table 2. Eight common KOGs among plants (*Arabidopsis thaliana*), animals (*Drosophila melanogaster, Homo sapiens, Caenorhabditis elegans*), fungi (*Saccharomyces cerevisiae, Neurospora crassa, Ashbya gossypii, Cryptococcus neoformans, Encephalitozoon cuniculi*), archaea (*Thermoplasma volcanium*) and bacteria (*Deinococcus radiodurans*)

COG	KOG	Predicted function	Main KOG functional category	Functional category
COG0156	KOG1358	Serine palmitoyltransferase	Cellular process and signaling	Posttranslational modification, protein turnover, chaperones
COG0476	KOG2015	NEDD8-activating complex, catalytic component UBA3	Cellular process and signaling	Posttranslational modification, protein turnover, chaperones
COG0533	KOG2708	Predicted metalloprotease with chaperone activity (RNAse H/HSP70-fold)	Cellular process and signaling	Posttranslational modification, protein turnover, chaperones
COG1236	KOG1135	mRNA cleavage and polyadenylation factor II complex, subunit CFT2 (CPSF subunit)	Information storage and processing	RNA processing and modification
COG0164	KOG1299	Vacuolar sorting protein VPS45/Stt10 (Sec1 family)	Cellular process and signaling	Intracellular trafficking, secretion, and vesicular transport
COG0101	KOG2554	Pseudouridylate synthase	Information storage and processing	Translation, ribosomal structure and biogenesis
COG0016	KOG2784	Phenylalanyl-tRNA synthetase, β-subunit	Information storage and processing	Translation, ribosomal structure and biogenesis
COG0697	KOG1443	Predicted integral membrane protein	Poorly characterized	Function unknown

COG, clusters of orthologous groups of proteins of prokaryote genomes.

genome (Table 2), and built a phylogenetic tree by concatenating the eight KOGs using maximum likelihood quartet puzzling as described below.

Presence or absence of KOGs

For the presence–absence of cluster of orthologous proteins, we used the method developed for prokaryotes as described by Snel *et al.* (2000).

Concatenation of 531 shared proteins

Multiple divergent copies of protein sequences of a single genome resulting from duplications of various age were assigned to the same KOG in several cases. To circumvent this problem, we aligned all the proteins available for the same KOG in a given genome against the proteins available for the other genomes and kept the most similar copy only.

For the phylogenetic analysis, the sequences of each protein KOG family from different species were aligned using CLUSTAL X (Thompson *et al.*, 1997). The GBlocks program (Castresana, 2000) was used to eliminate poorly alignable regions. The multiple alignments of each protein commonly present in all genomes were concatenated.

The concatenated alignment was analyzed using maximum parsimony (MP), neighbor joining (NJ), quartet puzzling using maximum likelihood (QP) and Bayesian inference (BI) using Markov chain Monte Carlo. MP and NJ analyses were done using PROTPARS (heuristic search with characters equally weighted) and PROTDIST (Kimura formula) from Phylip (Felsenstein, 1996), respectively. Nonparametric bootstrap support for MP and NJ was calculated from 100 resampling rounds. QP trees were constructed

with the TREE-PUZZLE program (Schmidt *et al.*, 2002) using the Whelan & Goldman (2001) model of amino acid substitution and 1000 puzzling rounds. For BI, we used MrBayes 3.0b4 (Huelsenbeck & Ronquist 2001), with four incrementally heated simultaneous Monte Carlo Markov chains over 50 000 generations using random starting trees and a Poisson model of amino acid substitution. Trees were sampled every 10 generations, resulting in an overall sampling of 5000 trees, of which the first 2000 were discarded. The remaining 3000 trees were used to estimate posterior probabilities (i.e. probabilities that groups of taxa are monophyletic, given the data) by computation of a 50% majority-rule consensus tree. Branch lengths were averaged over the sampled trees, again discarding the first 2000 trees. Stationarity of the process was controlled using Tracer software, version 1.0 (Rambaut & Drummond, 2003). The Bayesian MCMC phylogenetic analysis was repeated, again using random starting trees, to test the independence of the results from topological priors.

Number of concatenated proteins needed to resolve the fungal phylogenetic tree branches

The next approach was to examine the number of concatenated proteins needed to resolve the tree branches with full support (100%). First, we calculated each distance matrix of the total of 531 proteins using CLUSTAL X. Second, we computed the Pearson correlation between all 531 single protein distance matrices and the distance matrix obtained from the concatenation of all 531 proteins. We ordered the proteins according to their correlation values and, subsequently, concatenated groups of 90, 180, 270, 360, 450 and

531 proteins, with decreasing correlation values. Each concatenated group of proteins (i.e. 90, 180, 270, 360, 450 and 531) was phylogenetically analyzed by MP in order to estimate the number of proteins required for well-supported trees and branches.

Results and discussion

Rooting the fungal phylogenetic tree

To root our phylogenetic trees, we selected eight common KOGs among one plant genome, three animal genomes, five fungal genomes, one archaea genome and one bacterial genome (Table 2), and built a phylogenetic tree using maximum likelihood QP by concatenating the eight KOGs. According to our analysis, fungi and animals are sister groups, with *Arabidopsis thaliana* being more basal (Fig. 1). This is in agreement with current hypotheses about the evolution of the main groups of eukaryotes (Baldauf, 2003). On the basis of this result, we rooted the phylogenetic trees derived from the 25 eukaryotic genomes with *Arabidopsis thaliana*.
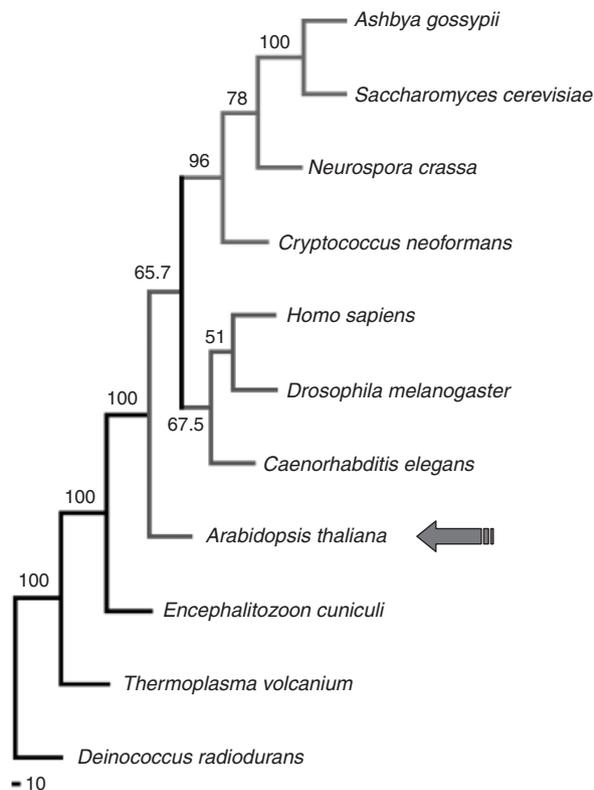

Fig. 1. Quartet puzzling tree based on concatenation of eight common KOGs among eukaryotes and prokaryotes listed in Table 2 to root the fungal phylogenetic tree. Numbers on branches are percentages of maximum likelihood quartet puzzling support values.

Fungal phylogenetic tree based on the presence or absence of KOGs

Analysis of genome content by comparing the presence or absence of KOGs is a straightforward strategy for genome comparison, because the more closely the genomes are related, the more genes will be shared between them. To our knowledge, a phylogenetic analysis based on the presence or absence of KOGs has not been performed for fungi. This method may be better suited for the analysis of eukaryote genomes than for those of prokaryotes, because the latter may be more heavily affected by horizontal gene transfer, which seems to be less important in eukaryotes (Rubin *et al.*, 2000). The main disadvantage of this approach is that the information contained in the protein sequences is not considered, and one assumes indirectly that those orthologs not present were lost in the same evolutionary way. The phylogenetic tree based on the presence or absence of proteins inferred from 4852 KOG families was found to be somewhat different from the tree inferred from the concatenation of all 531 shared proteins (Fig. 2; see discussion below). The main difference was related to the position of the fission yeast, *Sch. pombe*. In contrast to the 531 shared-protein tree, this species clustered within the *Hemi-ascomycetes* lineage, basal to all species, except *Y. lipolytica*, and not as a basal branch to all *Ascomycota* (data not shown). The KOGs are excellent for finding ubiquitous proteins and assigning these to functional categories. It seems that a gene content tree, probably because of the limited species sampling in the original KOG database, is less reliable for phylogenetic inferences.

Fungal phylogenetic tree based on concatenation of 531 shared proteins

Out of a total of 4852 KOG families, a substantial subset of 601 KOG proteins was shared by all 25 genomes. However, only 531 KOGs (Table S1) had informative amino acid sequences based on GBlocks analysis. This means that we retained a fraction of 10.9% from the total number of KOGs for phylogenetic analyses. These 531 common KOGs represent almost all classes of functional KOG categories: (1) information storage and processing (28.9%); (2) cellular processes and signaling (33.7%); (3) metabolism (23.2%); and (4) poorly characterized (14.2%) (Fig. 3). This is different from the results of bacterial studies, where only the information storage and processing category was found to be strongly represented and phylogenetically informative (Daubin *et al.*, 2002). This suggests that eukaryotes have a larger core set of phylogenetically informative proteins.

The concatenated alignment of the shared proteins had 67 101 amino acid positions. All phylogenetic methods applied, i.e. MP, NJ, QP and BI, resulted in a single tree
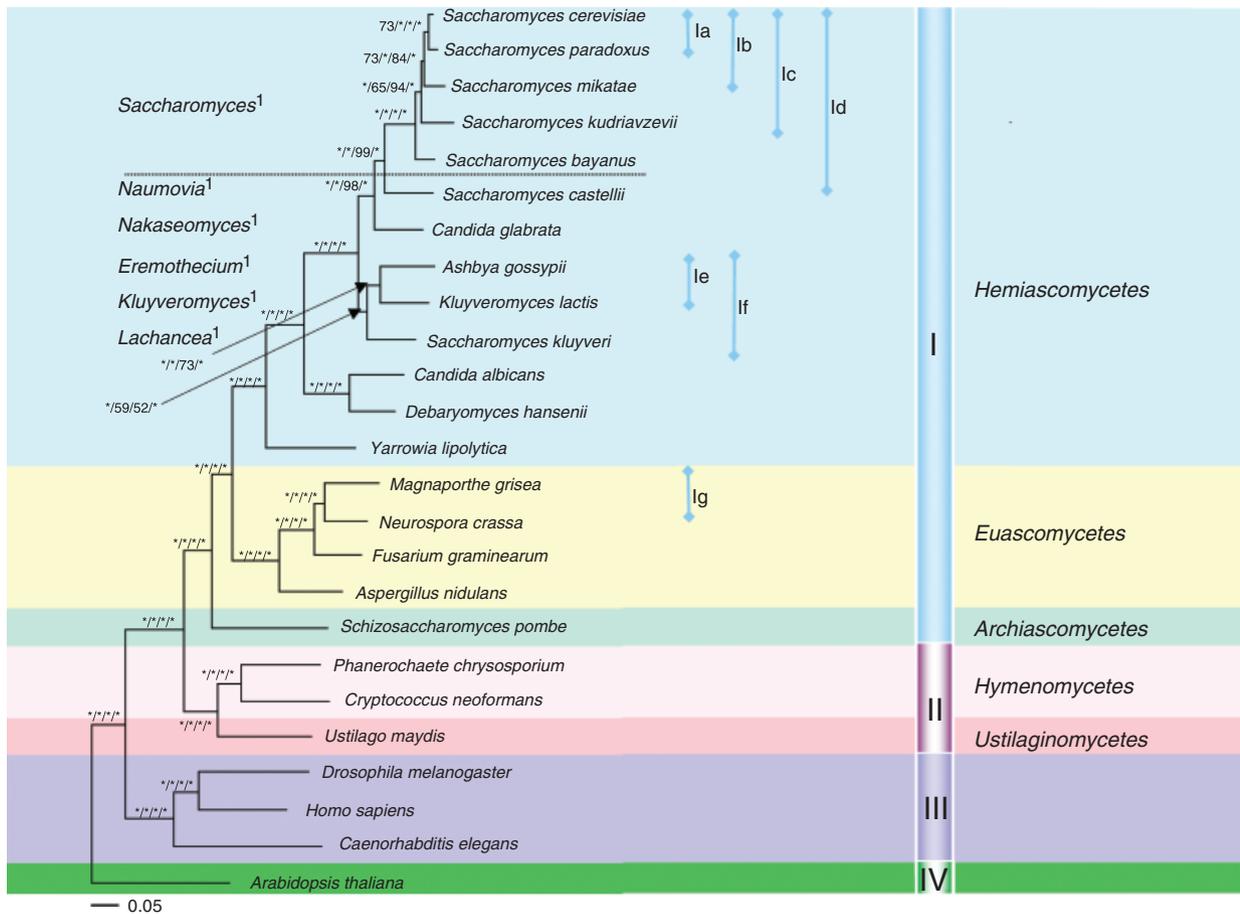
Fig. 2. Phylogenetic tree based on 531 concatenated proteins present in all 25 genomes studied. The numbers from right to left are branch support values (maximum parsimony bootstrap/neighbor-joining bootstrap/maximum likelihood quartet puzzling support values/posterior probabilities from Bayesian inference). Branch lengths were estimated using Bayesian inference. *represents 100%. I, *Ascomycota*; II, *Basidiomycota*; III, animals; IV, plants. Clusters represent the following species: Ia (Sce+Spa), Ib (Sce+Spa+Smi), Ic (Sce+Spa+Smi+Sku), Id (Sce+Spa+Smi+Sku+Sba+Sca), Ie (Ago+Kla), If (Ago+Kla+Skl), Ig (Mgr+Ncr). [1]New genera recently proposed (Kurtzman, 2003).

(Fig. 2) with most internal branches supported with 100%, suggesting robustness of our dataset. Support values smaller than 100% were only observed for some branches of the *Saccharomyces* species complex. In BI analysis, the tree topologies became constant after stationarity of the Markov chains was reached. Contrary to previous fungal phylogenetic analyses based on one or a few genes (e.g. Kurtzman & Robnett, 2003; Tehler *et al.*, 2003; Lutzoni *et al.*, 2004), our analysis resulted in fully resolved phylogenetic trees with highly supported internal branches. The *Eumycota* had the *Ascomycota* and *Basidiomycota* as sister groups (Fig. 2I and II), and there was a well-supported tripartition of the *Ascomycota* into basal *Archiascomycetes*, and *Hemiascomycetes* and *Euascomycetes* as sister groups. This is consistent with a recent phylogenetic study using two-gene, three-gene and four-gene alignments (Lutzoni *et al.*, 2004). Our results indicate that *Sch. pombe* forms a basal lineage within the *Ascomycota*, which is in line with previous suggestions

(Eriksson *et al.*, 1993), but is in contradiction with data presented by Prillinger *et al.* (2002) and Diezmann *et al.* (2004). The construction of phylogenetic trees based on whole genome data resolved the position of the *Archiascomycetes* as represented by *Sch. pombe*.

Previous analyses using multigene sequences (e.g. 18S, 5.8S, ITS and 26S rRNA, EF-1α, mitochondrial small-subunit rDNA and COX II) separated *Saccharomyces sensu stricto* and *Saccharomyces sensu lato* species (Kurtzman & Robnett, 2003). In our analysis, the phylogenetic positions of the *sensu stricto* species *Saccharomyces cerevisiae*, *Saccharomyces paradoxus*, *Saccharomyces mikatae*, *Saccharomyces kudriavzevii* and *Saccharomyces bayanus* corroborated previous results of Rokas *et al.* (2003), who used 106 concatenated genes. In contrast, our tree topology was somewhat different from that found by Edwards-Ingram *et al.* (2004) by comparative genomic hybridization (CGH), the multigene analysis presented by Kurtzman & Robnett (2003), and the
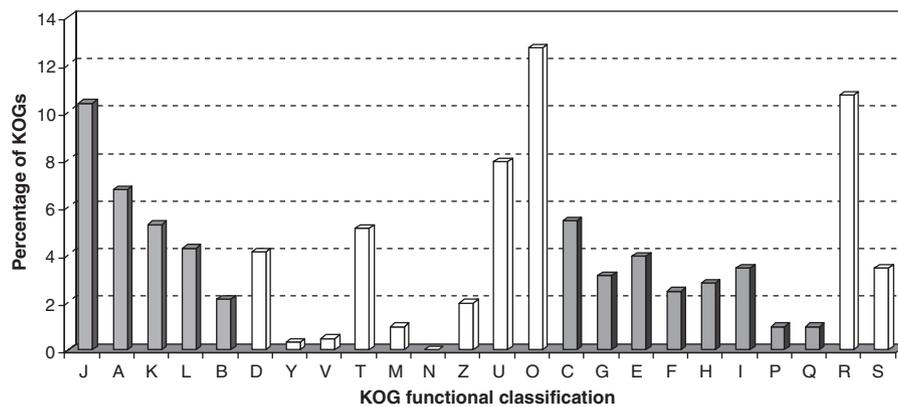
Fig. 3. Percentage of proteins in each KOG functional category present in all 25 genomes. The categories are grouped into four main categories: (1) information storage and processing ([J] translation, ribosomal structure and biogenesis, [A] RNA processing and modification, [K] transcription, [L] replication, recombination and repair, [B] chromatin structure and dynamics); (2) cellular processing and signaling ([D] cell cycle control, cell division, chromosome partitioning, [Y] nuclear structure, [V] defense mechanisms, [T] signal transduction mechanisms, [M] cell wall/membrane/envelope biogenesis, [N] cell motility, [Z] cytoskeleton, [U] intracellular trafficking, secretion, and vesicular transport, [O] posttranslational modification, protein turnover, chaperones); (3) metabolism ([C] energy production and conversion, [G] carbohydrate transport and metabolism, [E] amino acid transport and metabolism, [F] nucleotide transport and metabolism, [H] coenzyme transport and metabolism, [I] lipid transport and metabolism, [P] inorganic ion transport and metabolism, [Q] secondary metabolite biosynthesis, transport and catabolism); (4) poorly characterized ([R] general function prediction only, [S] function unknown) (Tatusov *et al*., 2003).

analysis based on 25S rRNA gene sequences (Dujon *et al*., 2004). The positions of *S. mikatae* and *S. kudriavzevii* were different from those in the CGH-based tree, whereas those of *S. cerevisiae*, *S. paradoxus* and *S. mikatae* differed if compared with the four-gene analysis (Kurtzman & Robnett, 2003). Most likely, the four-gene dataset is too limited to fully resolve the phylogenetic position of these species, which is also suggested by the relatively low bootstrap values observed in that study (i.e. 69% for *S. mikatae* and *S. paradoxus*, and 68% for *S. cerevisiae*, *S. paradoxus* and *S. mikatae*). This is even more true for the 25S rRNA gene analysis, which gave only 36% bootstrap for *S. paradoxus* and *S. mikatae*, and 39% for *S. paradoxus*, *S. mikatae* and *S. kudriavzevii* (Dujon *et al*., 2004). Combining sequence data of 531 proteins using the four phylogenetic methods enabled us to clarify the phylogenetic positions of *A. gossypii*, *K. lactis* and *S. kluyveri*, which was not possible using 106 concatenated genes (Hittinger *et al*., 2004).

The placement of the hemiascomycetous yeasts as a sister group to the filamentous Euascomycetes, the presence of the dimorphic yeast species *Y. lipolytica*, *C. albicans* and *A. gossypii* towards the base of the yeast cluster, and the basal position of the fission yeast *Sch. pombe* within the *Ascomycota* (Fig. 2) suggests that the hemiascomycetous yeasts have secondarily evolved from a filamentous or dimorphic ancestor.

Number of concatenated proteins needed to resolve branches of the fungal phylogenetic tree

To estimate the number of proteins needed to fully resolve the phylogenetic relationships among the species studied here, we established the similarity of the phylogenetic in-

formation in each KOG. To discover this, we first determined the correlations between each KOG protein distance matrix and the distance matrix generated by the 531 concatenated KOGs. Subsequently we concatenated varying numbers of proteins according to their correlation values (see Materials and methods). The correlation values ranged from –0.018 to 0.98 (Table S1). Approximately 70% of the KOG families have high (> 0.70) correlation values (Fig. 4). The number of concatenated proteins needed to resolve most of the internal nodes with maximum support (i.e. 100% bootstrap) is about 270 (Fig. 5). These proteins have correlation values ranging from 0.83 to 0.98 (Table S1) when compared to the distance matrix inferred from the 531 concatenated proteins. However, the cluster of *A. gossypii*, *K. lactis* and *S. kudriavzevii* was not well resolved using those 270 concatenated proteins (Figs 2 and 5 cluster If). This latter cluster needs concatenation of all 531 commonly occurring proteins (Fig. 2).

We have shown that whole-genome comparisons are the ultimate informative data for estimating accurately the phylogenetic relationships among fungal species. Although there are many fungal sequencing genome projects ongoing, mainly by the Broad Institute – Fungal Genome Initiative (FGI; http://www.broad.mit.edu/annotatio/fgi), most concern species belonging to the two phyla *Ascomycota* and *Basidiomycota*. Future studies to understand the fungal TOL should also include representative species of the remaining phyla: *Chytridiomycota*, *Glomeromycota* and *Zygomycota*. In addition, we demonstrated that concatenation of shared orthologous proteins is a robust method to infer fungal phylogenetic relationships. It seems that the noise caused by proteins with different evolutionary histories, in this study

Fig. 4. Percentage of KOGs per interval of correlation values determined by comparing each protein's distance matrix with the distance matrix derived from the concatenated alignment of 531 proteins.
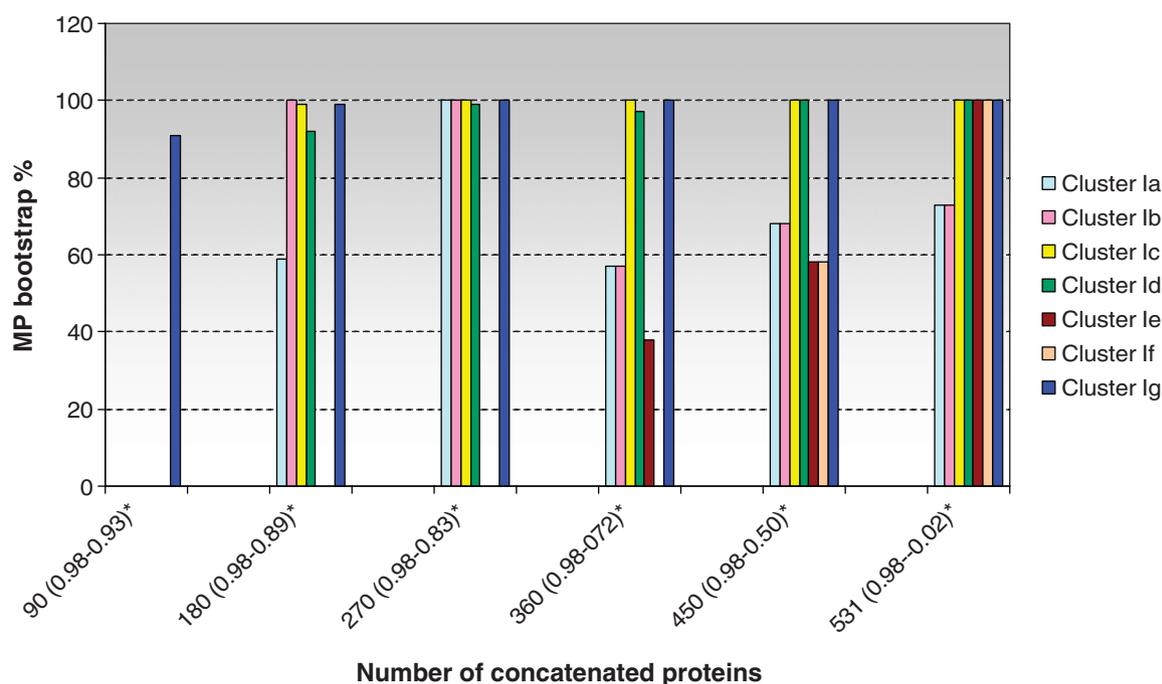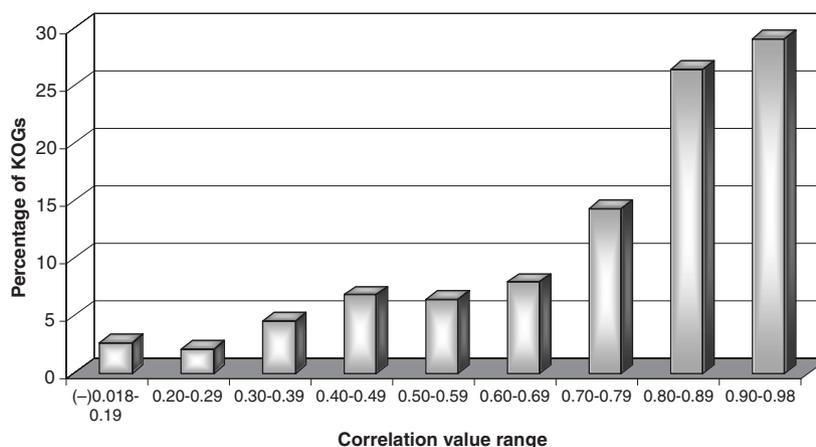

Fig. 5. Percentage of bootstrap values per node obtained after maximum parsimony analysis using 90, 180, 270, 360, 450 and 531 concatenated proteins shared by 25 eukaryotic genomes. Clusters Ia, Ib, Ic, Id, Ie, If and Ig are illustrated in Fig. 2. *The cophenetic correlation values for the groups of concatenated proteins. The absence of bootstrap values for some clusters means that they show a different topology in the respective analysis.

probably those proteins with low correlation values for the distance matrix of the concatenated dataset, does not affect the fungal phylogenetic tree based on 531 proteins and represented by 25 genomes. Finally, we have shown that the number of concatenated proteins needed for a robust phylogenetic hypothesis is likely to depend on the phylogenetic distance and the number of the species to be analyzed.

Acknowledgements

References

Baldauf SL (2003) The deep roots of Eukaryotes. *Science* **300**: 1703–1706.

Berbee ML & Taylor JW (2001) Fungal molecular evolution: gene trees and geologic time. *The Mycota: a Comprehensive Treatise on Fungi as Experimental Systems for Basic and Applied Research. Vol. VII* (McLaughlin DJ, McLaughlin EG & Lemke PA, eds), pp 229–245. Springer, Berlin.

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

Daubin V, Gouy M & Perrière G (2002) A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res* **12**: 1080–1090.

Diezmann S, Cox CJ, Schonian G, Vilgalys RJ & Mitchell TG (2004) Phylogeny and evolution of medical species of *Candida* and related taxa: a multigenic analysis. *J Clin Microbiol* **42**: 5624–5635.

Dujon B, Sherman D, Fischer G, *et al.* (2004) Genome evolution in yeasts. *Nature* **430**: 35–44.

Edwards-Ingram LC, Gent ME, Hoyle DC, Hayes A, Stateva LI & Oliver SG (2004) Comparative genomic hybridization provides new insights into the molecular taxonomy of the *Saccharomyces sensu stricto* complex. *Genome Res* **14**: 1043–1051.

Eriksson OE, Svedskog A & Landvik S (1993) Molecular evidence for the evolutionary hiatus between *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. *Syst Ascomyc* **11**: 119–162.

Felsenstein J (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol* **266**: 418–427.

Galagan JE, Henn MR, Ma L, Cuomo CA & Birren B (2005) Genomics of the fungal kingdom: insights into eukaryotic biology. *Genome Res* **15**: 1620–1631.

Hawksworth DL (2001) The magnitude of fungal diversity: the 1.5 million species estimate revisited. *Mycol Res* **105**: 1422–1432.

Hittinger CT, Rokas A & Carroll SB (2004) Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. *Proc Natl Acad Sci USA* **101**: 14144–14149.

Huelsenbeck JP & Ronquist FR (2001) MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754–755.

Kurtzman CP (2003) Phylogenetic circumscription of *Saccharomyces*, *Kluyveromyces* and other members of the *Saccharomycetaceae*, and the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygotorulaspora*. *FEMS Yeast Res* **4**: 233–245.

Kurtzman CP & Robnett CJ (2003) Phylogenetic relationships among yeasts of the *Saccharomyces* complex determined from multigene sequence analyses. *FEMS Yeast Res* **3**: 417–432.

Lutzoni F, Kauff F, Cox CJ, *et al.* (2004) Assembling the fungal tree of life: progress, classification and evolution of subcellular traits. *Am J Bot* **91**: 1446–1480.

McLaughlin DJ, McLaughlin EG & Lemke PA (2001) *The Mycota: Systematics and Evolution. Vol VII A*. Springer, Berlin.

Prillinger H, Lopandic K, Schweigkofler W, Deak R, Aarts HJM, Bauer R, Sterflinger K, Kraus GF & Maraz A (2002) Phylogeny and systematics of the fungi with special reference to the *Ascomycota* and *Basidiomycota*. *Chem Immunol* **81**: 207–295.

Rambaut A & Drummond A (2003) Tracer. MCMC Trace Analysis Tool. 1.0. University of Oxford, UK.

Rokas A, Williams BL, King N & Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**: 798–804.

Rubin GM, Yandell MD, Wortman JR, *et al.* (2000) Comparative genomics of the Eukaryotes. *Science* **287**: 2204–2215.

Sanderson MJ, Driskell AC, Ree RH, Eulenstein O & Langley S (2003) Obtaining maximal concatenated phylogenetic data sets from large sequence databases. *Mol Biol Evol* **20**: 1036–1042.

Schmidt HA, Strimmer K, Vingron M & von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**: 502–504.

Schüßler A, Schwarzott D & Walker C (2001) A new fungal phylum, the *Glomeromycota*: phylogeny and evolution. *Mycol Res* **105**: 1413–1421.

Snel B, Bork P & Huynen MA (1999) Genome phylogeny based on gene content. *Nat Genet* **21**: 108–110.

Snel B, Lehmann G, Bork P & Huynen MA (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res* **28**: 3442–3444.

Tatusov RL, Fedorova ND, Jackson JD, *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**: 41–54.

Tehler A, Little DP & Farris JS (2003) The full-length phylogenetic tree from 1551 ribosomal sequences of chitinous fungi, *Fungi*. *Mycol Res* **107**: 901–916.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F & Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**: 4876–4882.

Whelan S & Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. *Mol Biol Evol* **18**: 691–699.

Wolf YI, Rogozin IB, Grishin NV, Tatusov RL & Koonin EV (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol Biol* **1**: 8–28.

Wolf YI, Rogozin IB & Koonin EV (2004) Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. *Genome Res* **14**: 29–36.

Supplementary material

The following supplementary material is available online:

Table S1. Core of 531 common KOGs shared among 21 fungal, three animal and one plant genomes.

This material is available as part of the online article from http://www.blackwell-synergy.com