Points of View

Syst. Biol. 54(4):660–668, 2005
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150590947104

Automated Phylogenetic Taxonomy: An Example in the Homobasidiomycetes (Mushroom-Forming Fungi)

David S. Hibbett,[1] R. Henrik Nilsson,[2] Marc Snyder,[1] Mario Fonseca,[1] Janine Costanzo,[1] and Moran Shonfeld[1]

[1]*Biology Department, Clark University, Worcester, Massachusetts 01610, USA; E-mail: dhibbett@black.clarku.edu (D.S.H.)*
[2]*Current Address: Botanical Institute, Göteborg University, Box 461, 405 30 Göteborg, Sweden; E-mail: henrik.nilsson@botany.gu.se*

Systematists confront a truly formidable task, which is to construct a comprehensive, phylogenetically accurate classification for all of life. Progress toward this goal has accelerated in recent years through analyses of DNA and protein sequences. The dramatic growth of molecular phylogenetics prompts us to ask whether established taxonomic practices are equipped to deal with the rapidly accumulating data. Here, we consider the current status of phylogenetics and classification in the homobasidiomycetes (mushroom-forming fungi), which includes roughly 17,000 described species (Kirk et al., 2001). We find that the available data are not being integrated, and that there is a significant gap between current taxonomy and understanding of phylogenetic relationships. To close this gap, we suggest that there is a need for new approaches that use the tools of bioinformatics to automate the process of phylogenetic analysis and classification, and we describe a prototype software package that we have developed for this purpose. The tools that we have created are applied to homobasidiomycetes, but can be adapted for use in any group of organisms.

Current Status of Homobasidiomycete Systematics

Progress in Phylogenetic Reconstruction

Taxonomy of homobasidiomycetes has been revolutionized through the use of molecular techniques. The most widely sampled locus for studies at high to moderate taxonomic levels (i.e., genera and higher) in homobasidiomycetes is a region of about 1000 base pairs at the 5′ end of the nuclear-encoded large subunit ribosomal DNA (nuc-lsu rDNA), which can be aligned across distantly related taxa, but can often distinguish closely related species (e.g., Moncalvo et al., 2002). As of November 2004, GenBank (http://www.ncbi.nlm.nih.gov/Genbank/index.html) contained about 4204 nuc-lsu rDNA sequences of homobasidiomycetes, representing 2167 distinct names (not all of the sequences are identified to the species level). However, the most comprehensive analyses so far include only 877 sequences (representing 877 species; Moncalvo et al., 2002) or 656 sequences (representing approximately 646 species; Binder et al., 2005), which indicates that the currently available data are not being integrated in comprehensive, simultaneous analyses. This is unfortunate, because comprehensive analyses are needed to determine the composition of clades and develop detailed phylogenetic hypotheses. Comprehensive analyses are also needed to detect mislabeled sequences, which are common in groups like fungi, where identification is difficult (Binder et al., 2005; Bridge et al., 2003; Vilgalys, 2003). Finally, analyses with multiple accessions of morphologically defined species are needed to reveal cryptic lineages.

To enable comprehensive phylogenetic analyses in homobasidiomycetes, it is important that the database of nuc-lsu rDNA sequences be expanded and integrated. However, nuc-lsu rDNA alone is not sufficient to reconstruct homobasidiomycete phylogeny in all of its details, as has been shown in many studies (e.g., Binder and Hibbett, 2002). Fortunately, analyses of multilocus data sets (supermatrices) are becoming more common (e.g., Matheny et al., 2002; Wang et al., 2004), and these are providing improved resolution of major clades of homobasidiomycetes and other groups of fungi (Binder and Hibbett, 2002; Lutzoni et al., 2004).

Numerous single-gene and multilocus studies have contributed to our current understanding of the major groups of homobasidiomycetes. In 2001, Hibbett and Thorn reviewed the results of 26 studies and proposed a "preliminary phylogenetic outline," in which the homobasidiomycetes were divided into eight mutually exclusive major clades that were given informal names (e.g., "euagarics clade"). Subsequently, several other putatively independent clades have been discovered, including the "athelioid clade," "corticioid clade," "trechisporoid clade," and "Gloeophyllum clade" (Hibbett and Binder, 2002; Langer, 2002; Larsson et al., 2004; Binder et al., 2005). There may be other major clades waiting to be discovered. For example, *Jaapia argillacea* and several other species of resupinate basidiomycetes cannot be placed into any of the twelve clades named above based on currently available sequences (Larsson et al., 2004; Binder et al., 2005). Analyses of environmental sequences also suggest that there are major groups of homobasidiomycetes that have not yet been described (Vandenkoornhuyse et al. 2002; Schadt et al., 2003). This is not surprising, given that only about 70,000 species of fungi have been described (Kirk et al., 2001), whereas it is estimated that there are as many as 1.5 million extant species (Hawksworth, 2001).

Progress in Classification

The fungal taxonomic literature is highly dispersed, which makes it difficult to assess the extent to which recent advances in phylogenetic reconstruction have been incorporated into currently accepted classifications. However, there is one major reference that provides a measure of the current status of fungal taxonomy. The *Dictionary of the Fungi* series presents a unified classification for all groups of fungi down to the level of genera, and is now in its ninth edition (Kirk et al., 2001). Between the eighth and ninth editions, the classification of homobasidiomycetes in the *Dictionary* was revised extensively, in response to progress in phylogenetic reconstruction (Hawksworth et al., 1995; Kirk et al., 2001). The eighth edition placed the homobasidiomycetes in 26 orders in the subclass Holobasidiomycetidae, whereas the ninth edition placed most of the homobasidiomycetes in eight orders in the subclass Agaricomycetidae.

The ninth edition of the *Dictionary of the Fungi* represented a major, episodic advance toward a phylogenetic classification of homobasidiomycetes. Nevertheless, based on current views of homobasidiomycete phylogeny (Hibbett and Thorn, 2001; Weiss and Oberwinkler, 2001; Larsson et al., 2004; Binder et al., 2005), the Agaricomycetidae and at least four of the orders within the Agaricomycetidae are probably not monophyletic (Table 1). The most problematic order is the Polyporales, which includes representatives of nine different clades.

It is obviously unfair to expect the classification in the *Dictionary of the Fungi* to reflect the most recent phylogenetic analyses, especially those that appeared after the latest edition of the *Dictionary* was published. Indeed, the discrepancy between the *Dictionary* classification and our current understanding of phylogeny argues for the abandonment of print as a medium for communicating the most up-to-date taxonomy (in fact, on-line classifications already exist for both Basidiomycota, http://www.mycokey.com/AAU/Systematics/SystematicsBasi.html, and Ascomycota, http://www.umu.se/myconet/Myconetxxxx.html). There is evidence that the limitations of print are not the only impediments to translation of phylogenetic trees into taxonomy, however. The *International Code of Botanical Nomenclature* (Greuter et al., 2000) may also be part of the problem (Hibbett and Donoghue, 1998).

The constraints on taxonomic progress imposed by the *Code* (Greuter et al., 2000) are illustrated by the controversy over the classification of the genus *Coprinus*, a familiar assemblage of "inky cap" mushrooms. In 1994, and repeatedly thereafter, *Coprinus* was shown to be polyphyletic (Hopple and Vilgalys, 1994; Redhead et al., 2001, and references therein). Redhead et al. (2001) suggested that *Coprinus* s. lat. should be divided into *Coprinus* s. str., *Coprinellus*, *Coprinopsis*, and *Parasola*, but this proposal has become mired in a nomenclatural debate concerning conservation and typification of generic names (Jørgensen et al., 2001; also see discussion at http://www.cbs.knaw.nl/nomenclature/index.htm). Consequently, *Coprinus* s. lat. remains an "accepted" taxon (Kirk et al., 2001) fully a decade after it was shown that it is not a clade.

The example of *Coprinus* demonstrates that the challenges of translating phylogenetic hypotheses into taxonomy can be especially difficult at taxonomic levels at or below the rank of family, because of nomenclatural complexities. Problems such as the one posed by *Coprinus* will only increase in number as more genera are sampled intensively. For example, in the study of Binder et al. (2005), 34 genera were resolved as nonmonophyletic, and in the study of Moncalvo et al. (2002), 44 genera were resolved as nonmonophyletic.

In summary, homobasidiomycete taxonomy is blessed with an abundance of data and analyses, but lacks efficient mechanisms to integrate the emerging data into comprehensive phylogenetic trees, or translate trees into classifications. Over 90% of the homobasidiomycete nuc-lsu rDNA sequences in GenBank have been published since 1998, and there is no reason to expect the rate of data acquisition to decline in the near term. As the sequence database grows, there is a danger that the gap between taxonomy and understanding of phylogenetic relationships will widen. To prevent this from happening, systematists may wish to consider new tools that automate the process of phylogenetic classification.

Toward an Automated Phylogenetic Taxonomy of Homobasidiomycetes

We have developed a Perl script package called *mor* that performs automated phylogenetic taxonomy in homobasidiomycetes. In brief, *mor* is a Web-accessible, interactive program that compiles and screens nuc-lsu rDNA sequences from GenBank, constructs phylogenetic trees, and translates trees into classifications (Fig. 1). Central to *mor* is the use of node-based phylogenetic taxon definitions, which allow the composition of taxa (clades) to be continuously updated as the underlying phylogeny changes. The *mor* package is available on an open-source basis and can be modified to work with any gene and any group of organisms. The following paragraphs provide an overview of the operations performed by *mor*. More detailed information is provided in

TABLE 1. Comparison of higher-level classification of homobasidiomycetes in the *Dictionary of the Fungi* versus current hypotheses of phylogenetic relationships, with exemplar genera. Genera followed by question marks are of uncertain placement.

	Dictionary of the Fungi, 9th ed. (2001)									
Current phylogeny	Agaricales	Boletales	Cantharellales	Ceratobasidiales	Hymenochaetales	Phallales	Polyporales	Russulales	Thelephorales	Tulasnellales
Athelioid clade							*Athelia*	*Stephanospora*?		
Bolete clade		*Boletus*								
Cantharelloid clade	*Multiclavula*		*Cantharellus*	*Ceratobasidium*			*Sistotrema*			*Tulasnella*
Corticioid clade							*Corticium*			
Euagarics clade	*Agaricus*	*Hymenogaster*					*Cylindrobasidium*			
Gloeophyllum clade							*Gloeophyllum*			
Gomphoid-Phalloid clade		*Gastrosporium*				*Phallus*				
Hymenochaetoid clade	*Rickenella*				*Hymenochaete*		*Cotylidia*			
Jaapia		*Jaapia*								
Polyporoid clade						*Beenakia*?	*Polyporus*			
Russuloid clade							*Albatrellus* p.p.	*Russula*		
Thelephoroid clade									*Thelephora*	
Trechisporoid clade							*Trechispora*			

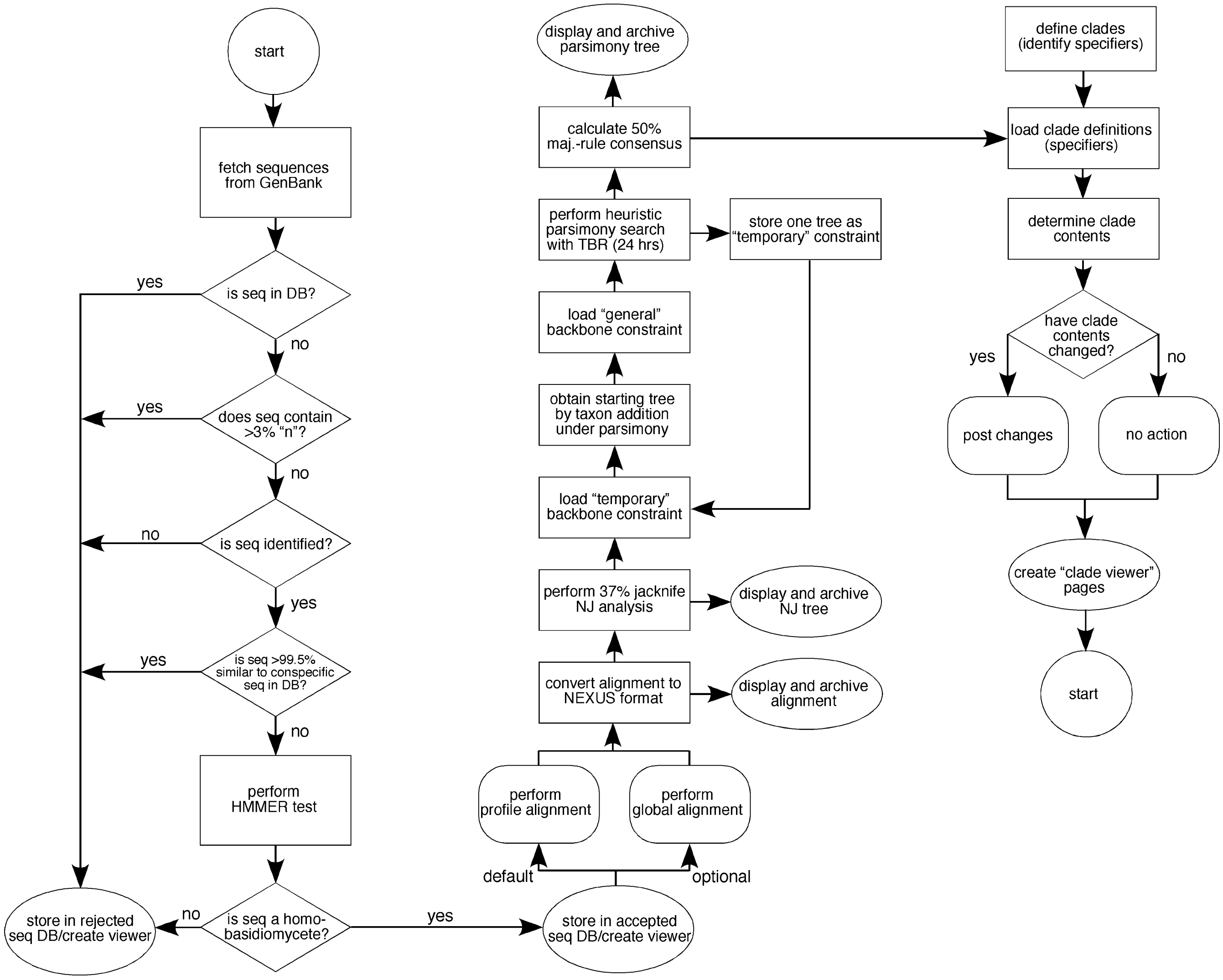

start
fetch sequences from GenBank
is seq in DB?
yes
no
does seq contain >3% "n"?
yes
no
is seq identified?
no
yes
is seq >99.5% similar to conspecific seq in DB?
yes
no
perform HMMER test
is seq a homo-basidiomycete?
no
yes
store in rejected seq DB/create viewer
store in accepted seq DB/create viewer
default
optional
perform profile alignment
perform global alignment
convert alignment to NEXUS format
display and archive alignment
perform 37% jacknife NJ analysis
display and archive NJ tree
load "temporary" backbone constraint
obtain starting tree by taxon addition under parsimony
load "general" backbone constraint
perform heuristic parsimony search with TBR (24 hrs)
store one tree as "temporary" constraint
calculate 50% maj.-rule consensus
display and archive parsimony tree
define clades (identify specifiers)
load clade definitions (specifiers)
determine clade contents
have clade contents changed?
yes
no
post changes
no action
create "clade viewer" pages
start

FIGURE 1. Simplified flow chart representing operations performed by *mor*, including sequence acquisition and screening (left), alignment and phylogenetic analyses (center), and extraction of clade composition information (right).

the program documentation available at the *mor* website (http://mor.clarku.edu).

Automated Phylogenetic Reconstruction

The *mor* software runs in a Linux environment (presently on a 1.6-GHz AMD Athlon PC). Upon being evoked on a weekly basis by the operating system, *mor* uses BioPerl routines (http://bioperl.org) to connect to GenBank and retrieve new sequences using a Boolean search string tailored to match only homobasidiomycete nuc-lsu rDNA sequences between 800 and 1500 base pairs. Downloaded sequences are screened according to several criteria, including percentage of ambiguous sites (sequences with more than 3% of the positions reported as "n" are rejected on the assumption that they are of low quality), and occurrence of putatively conspecific sequences with at least 99.5% identity in the dataset (values of these screening parameters are arbitrary and can be adjusted by modifying the Perl script). The latter criterion is intended to prevent redundant sequences from accumulating in the data set, while still allowing sequences representing cryptic species to be added. Sequences that pass these tests are subjected to a final screening procedure using a profile hidden Markov model constructed with HMMER 2.3.2 (Eddy, 1998; http://hmmer.wustl.edu/) based on a reference alignment of 1275 homobasidiomycete nuc-lsu rDNA sequences. For each candidate sequence, HMMER calculates a bit score, which reflects whether the candidate sequence is a better match to the profile model or a null model of nonhomologous sequences, and an E-value, which reflects the number of false positives expected at that bit score. Threshold values for bit scores and E-values were determined by matching 200 other homobasidiomycete sequences and 200 nonhomobasidiomycete sequences to the reference alignment using the *hmmpfam* component of HMMER. Following the screening procedure, the accepted and rejected sequences are stored in separate Web-accessible MySQL databases (http://www.mysql.com). Reasons for rejection are noted in a rejected sequences viewer.

Accepted sequences are aligned using Clustal W (Thompson et al., 1994) in profile mode (optionally, Clustal W can be set to perform a global alignment). Aligned sequences are analyzed by PAUP* 4.0b10 (Swofford, 2003) using an unconstrained 37% jacknife neighbor-joining (NJ) analysis, which provides a crude measure of clade support, followed by a constrained equally weighted heuristic maximum parsimony (MP) analysis. The jacknife NJ analysis uses 100 replicates and maximum likelihood distances calculated with the GTR model. The MP analysis uses a 202-sequence backbone monophyly constraint (Fig. 2) that was constructed by hand using MacClade 4.0 (Maddison and Maddison, 2001), and which reflects major groupings based on prior studies, including multilocus analyses (Hibbett and Thorn, 2001; Binder and Hibbett, 2002; Larsson et al., 2004; Lutzoni et al., 2004; Binder et al., 2005). The backbone constraint tree does not include representatives of the athelioid clade, trechisporoid clade, or Auriculariales s. lat., because their higher-level placements are poorly resolved.

For the MP analysis, *mor* uses a "twin constraint" search strategy, which eliminates the need to construct a starting tree from scratch through stepwise taxon addition and to perform the early rounds of branch-swapping in each iteration of the program. At the outset, this search protocol requires the presence of a data matrix, one treefile containing a "general constraint" tree (i.e., the 202-sequence backbone constraint tree described above), and another treefile containing a "temporary constraint," which is one of the most parsimonious trees generated in the previous iteration of *mor*. The search proceeds as follows:

1. Update the starting dataset (i.e., download, screen, and add new sequences to the existing alignment).
2. Execute the data set and load the temporary constraint tree as a backbone monophyly constraint; add the new sequences to this tree by stepwise taxon addition, but do not perform branch swapping.
3. Load the general constraint (Fig. 2) as a backbone monophyly constraint (i.e., replace the current constraint); perform a heuristic search using the tree in memory as a starting tree for branch swapping (presently *mor* is set to perform TBR for 24 hours).
4. Write one of the resulting trees to the temporary constraint treefile (i.e., overwrite the old temporary

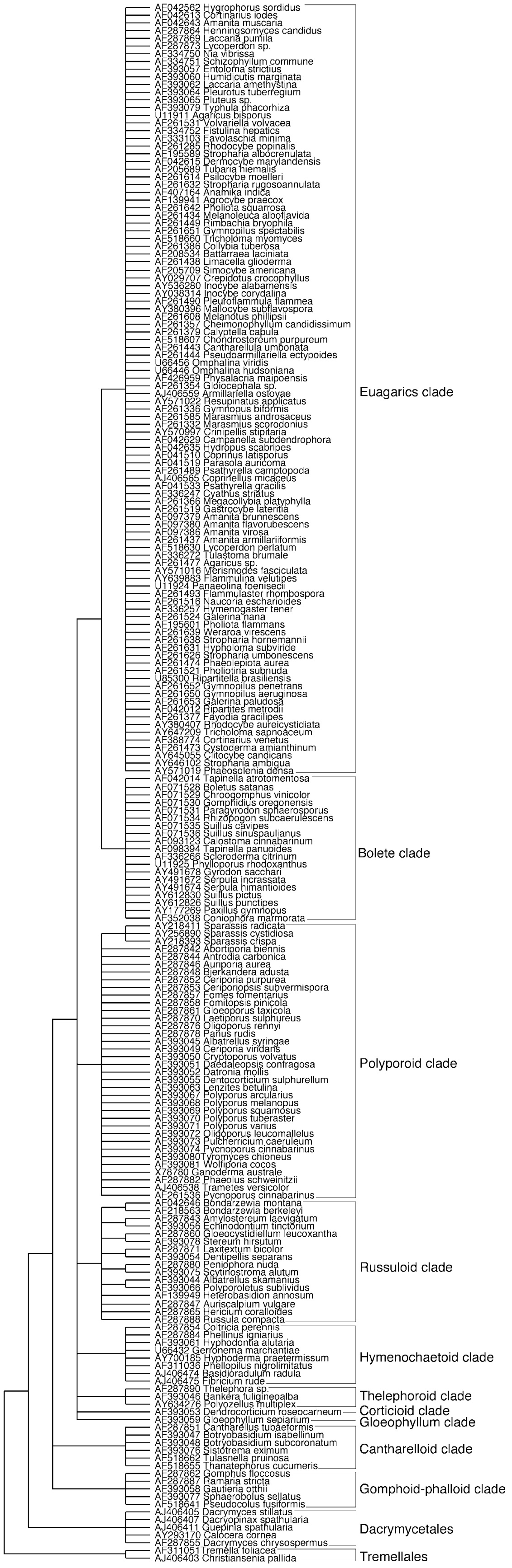

AF042562 Hygrophorus sordidus
AF042613 Cortinarius iodes
AF042643 Amanita muscaria
AF287864 Henningsomyces candidus
AF287869 Laccaria pumila
AF287873 Lycoperdon sp.
AF334750 Nia vibrissa
AF334751 Schizophyllum commune
AF393057 Entoloma strictius
AF393060 Humidicutis marginata
AF393062 Laccaria amethystina
AF393064 Pleurotus tuberregium
AF393065 Pluteus sp.
AF393079 Typhula phacorhiza
U11911 Agaricus bisporus
AF261531 Volvariella volvacea
AF334752 Fistulina hepatics
AF333103 Favolaschia minima
AF261285 Rhodocybe popinalis
AF195589 Stropharia albocrenulata
AF042615 Dermocybe marylandensis
AF205689 Tubaria hiemalis
AF261614 Psilocybe moelleri
AF261632 Stropharia rugosoannulata
AF407164 Anamika indica
AF139941 Agrocybe praecox
AF261642 Pholiota squarrosa
AF261434 Melanoleuca alboflavida
AF261449 Rimbachia bryophila
AF261651 Gymnopilus spectabilis
AF518660 Tricholoma myomyces
AF261386 Collybia tuberosa
AF208534 Battarraea laciniata
AF261438 Limacella glioderma
AF205709 Simocybe americana
AY029707 Crepidotus crocophyllus
AY536280 Inocybe alabamensis
AY038314 Inocybe corydalina
AF261490 Pleuroflammula flammea
AY380396 Mallocybe subflavospora
AF261608 Melanotus phillipsii
AF261357 Cheimonophyllum candidissimum
AF261379 Calyptella capula
AF518607 Chondrostereum purpureum
AF261443 Cantharellula umbonata
AF261444 Pseudoarmillariella ectypoides
U66456 Omphalina viridis
U66446 Omphalina hudsoniana
AF426959 Physalacria maipoensis
AF261354 Gloiocephala sp.
AJ406559 Armillariella ostoyae
AY571022 Resupinatus applicatus
AF261336 Gymnopus biformis
AF261585 Marasmius androsaceus
AF261332 Marasmius scorodonius
AY570997 Crinipellis stipitaria
AF042629 Campanella subdendrophora
AF042635 Hydropus scabripes
AF041510 Coprinus latisporus
AF041519 Parasola auricoma
AF261489 Psathyrella camptopoda
AJ406565 Coprinellus micaceus
AF041533 Psathyrella gracilis
AF336247 Cyathus striatus
AF261366 Megacollybia platyphylla
AF261519 Gastrocybe lateritia
AF097379 Amanita brunnescens
AF097380 Amanita flavorubescens
AF097386 Amanita virosa
AF261437 Amanita armillariiformis
AF518630 Lycoperdon perlatum
AF336272 Tulastoma brumale
AF261477 Agaricus sp.
AY571016 Merismodes fasciculata
AY639883 Flammulina velutipes
U11924 Panaeolina foenisecii
AF261493 Flammulaster rhombospora
AF261516 Naucoria escharioides
AF336257 Hymenogaster tener
AF261524 Galerina nana
AF195601 Pholiota flammans
AF261639 Weraroa virescens
AF261638 Stropharia hornemannii
AF261631 Hypholoma subviride
AF261626 Stropharia umbonescens
AF261474 Phaeolepiota aurea
AF261521 Pholiotina subnuda
U85300 Ripartitella brasiliensis
AF261652 Gymnopilus penetrans
AF261650 Gymnopilus aeruginosa
AF261653 Galerina paludosa
AF042012 Ripartites metrodii
AF261377 Fayodia gracilipes
AY380407 Rhodocybe aureicystidiata
AY647209 Tricholoma sapnoaceum
AF388774 Cortinarius venetus
AF261473 Cystoderma amianthinum
AY645055 Clitocybe candicans
AY646102 Stropharia ambigua
AY571019 Phaeosolenia densa
Euagarics clade
AF042014 Tapinella atrotomentosa
AF071528 Boletus satanas
AF071529 Chroogomphus vinicolor
AF071530 Gomphidius oregonensis
AF071531 Paragyrodon sphaerosporus
AF071534 Rhizopogon subcaerulescens
AF071535 Suillus cavipes
AF071536 Suillus sinuspaulianus
AF093123 Calostoma cinnabarinum
AF098394 Tapinella panuoides
AF336266 Scleroderma citrinum
U11925 Phylloporus rhodoxanthus
AY491678 Gyrodon sacchari
AY491672 Serpula incrassata
AY491674 Serpula himantioides
AY612830 Suillus pictus
AY612826 Suillus punctipes
AY177269 Paxillus gymnopus
AF352038 Coniophora marmorata
Bolete clade
AY218411 Sparassis radicata
AY256890 Sparassis cystidiosa
AY218393 Sparassis crispa
AF287842 Abortiporia biennis
AF287844 Antrodia carbonica
AF287846 Auriporia aurea
AF287848 Bjerkandera adusta
AF287852 Ceriporia purpurea
AF287853 Ceriporiopsis subvermispora
AF287857 Fomes fomentarius
AF287858 Fomitopsis pinicola
AF287861 Gloeoporus taxicola
AF287870 Laetiporus sulphureus
AF287876 Oligoporus rennyi
AF287878 Panus rudis
AF393045 Albatrellus syringae
AF393049 Ceriporia viridans
AF393050 Cryptoporus volvatus
AF393051 Daedaleopsis confragosa
AF393052 Datronia mollis
AF393055 Dentocorticium sulphurellum
AF393063 Lenzites betulina
AF393067 Polyporus arcularius
AF393068 Polyporus melanopus
AF393069 Polyporus squamosus
AF393070 Polyporus tuberaster
AF393071 Polyporus varius
AF393072 Oligoporus leucomallelus
AF393073 Pulcherricium caeruleum
AF393074 Pycnoporus cinnabarinus
AF393080Tyromyces chioneus
AF393081 Wolfiporia cocos
X78780 Ganoderma australe
AF287882 Phaeolus schweinitzii
AJ406538 Trametes versicolor
AF261536 Pycnoporus cinnabarinus
Polyporoid clade
AF042646 Bondarzewia montana
AF218563 Bondarzewia berkeleyi
AF287843 Amylostereum laevigatum
AF393056 Echinodontium tinctorium
AF287860 Gloeocystidiellum leucoxantha
AF393078 Stereum hirsutum
AF287871 Laxitextum bicolor
AF393054 Dentipellis separans
AF287880 Peniophora nuda
AF393075 Scytinostroma alutum
AF393044 Albatrellus skamanius
AF393066 Polyporoletus sublividus
AF139949 Heterobasidion annosum
AF287847 Auriscalpium vulgare
AF287865 Hericium coralloides
AF287888 Russula compacta
Russuloid clade
AF287854 Coltricia perennis
AF287884 Phellinus igniarius
AF393061 Hyphodontia alutaria
U66432 Gerronema marchantiae
AY700185 Hyphoderma praetermissum
AF311036 Phellopilus nigrolimitatus
AJ406474 Basidioradulum radula
AJ406475 Fibricium rude
Hymenochaetoid clade
AF287890 Thelephora sp.
AF393046 Bankera fuligineoalba
AY634276 Polyozellus multiplex
Thelephoroid clade
AF393053 Dendrocorticium roseocarneum
Corticioid clade
AF393059 Gloeophyllum sepiarium
Gloeophyllum clade
AF287851 Cantharellus tubaeformis
AF393047 Botryobasidium isabellinum
AF393048 Botryobasidium subcoronatum
AF393076 Sistotrema eximum
AF518662 Tulasnella pruinosa
AF518655 Thanatephorus cucumeris
Cantharelloid clade
AF287862 Gomphus floccosus
AF287887 Ramaria stricta
AF393058 Gautieria otthii
AF393077 Sphaerobolus sellatus
AF518641 Pseudocolus fusiformis
Gomphoid-phalloid clade
AJ406405 Dacrymyces stillatus
AJ406407 Dacryopinax spathularia
AJ406411 Guepinia spathularia
AY293170 Calocera cornea
AF287855 Dacrymyces chrysospermus
Dacrymycetales
AF311051Tremella foliacea
AJ406403 Christiansenia pallida
Tremellales

FIGURE 2. Two hundred two-sequence backbone monophyly constraint tree used by *mor*. Major clades are bracketed and labeled. Several of the clades delimited in *mor* (Fig. 3) are not represented in this constraint tree, including the trechisporoid clade, athelioid clade, Auriculariales, and Sebacinales. These groups were excluded from the backbone constraint because their higher-level placements are uncertain.

constraint treefile); this becomes the starting tree for the next run of the program.
5. Calculate a 50% majority-rule consensus tree.
6. Go to step 1.

The majority-rule consensus MP tree and the jackknife NJ tree are converted to tree graphics in ASCII characters and are made available on the *mor* site through the Apache Web server (http://httpd.apache.org). Terminals in the trees are labeled with complete genus and species names and GenBank accession numbers, and are provided with links that allow viewers to toggle between the parsimony and NJ trees, view the GenBank accession directly, make and view comments, or search for images on the internet. Each week, the current alignment and MP and NJ trees are archived on the *mor* website (examples of these files have been deposited at the *Systematic Biology* website http://www.systbiol.org).

Automated Phylogenetic Taxonomy

The routines described above produce comprehensive, continuously updated phylogenetic trees, but they do not generate taxonomy. For this purpose, *mor* includes a "clade parser," which delimits the contents of individual clades, and thus translates trees into classifications. The clade parser takes as its input a tree in Newick format (http://evolution.genetics.washington.edu/phylip/newicktree.html), representing the 50% majority-rule consensus tree from the MP analysis, and sets of up to 20 sequences identified by GenBank accession numbers that act as "specifiers" in node-based taxon definitions (de Queiroz and Gauthier, 1992). A clade is thus defined as the most recent common ancestor of its specifiers, and all of its descendants (i.e., the least inclusive monophyletic group that contains all of the specifiers).

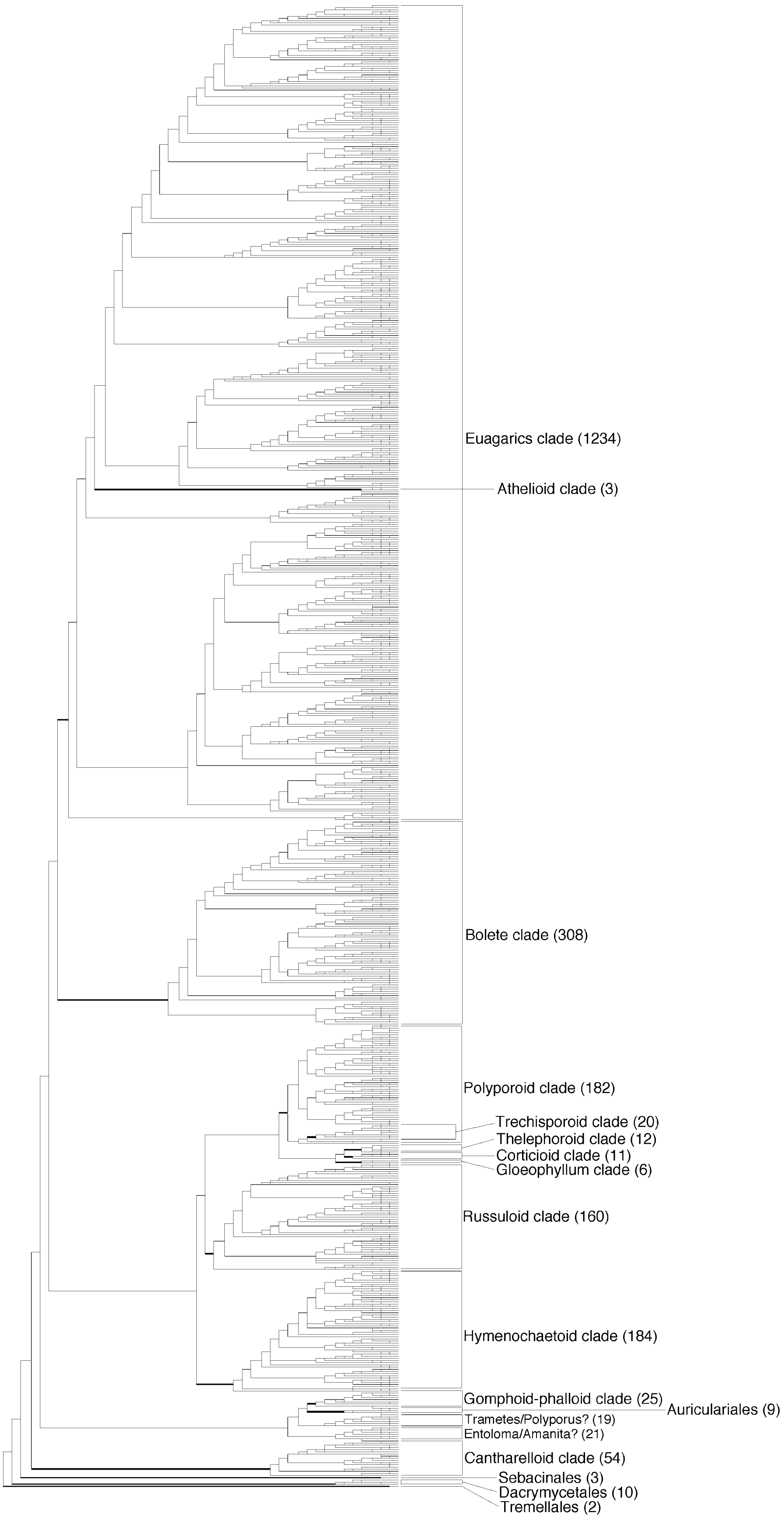
To determine the contents of each clade, the clade parser locates the first and last specifiers listed in the Newick string. Beginning from the first specifier, the clade parser traverses the tree in a down-pass to the root, marking each node that is traversed. Next, the clade parser traverses the tree in a down-pass toward the root starting from the last specifier in the Newick string. When a node that was visited on the first down-pass is reached, that node is marked as the most recent common ancestor of the clade. An up-pass is then performed, leading to all of the terminals in the clade, which are compiled in a list. A "clade viewer" web page is then produced, which presents a tree graphic (in the same form as the full MP and NJ trees), a list of all included sequences (with links to literature and images), and a link to download a NEXUS format (Maddison et al., 1997) file with the aligned sequences and tree in Newick format for the clade. The tree graphic on the clade viewer pages is produced using the CONSENSE component of the PHYLIP package (Felsenstein, 1989). The information on the clade viewer pages is automatically updated each week as the underlying phylogeny changes. A summary of recent changes in the compositions of all clades is posted, and the records of such changes are archived and accessible via the *mor* web site. In addition, an "incertae sedis" page is created, which lists all of the sequences that are not placed in any defined clade. Users may add new clades to *mor* by supplying a list of specifiers via web-accessible forms, along with a name for the clade and a set of author's comments.

As of this writing, *mor* has downloaded 2416 sequences, which is 57% of the homobasidiomycete nuc-lsu rDNA sequences in GenBank. The remaining 1788 putative homobasidiomycete nuc-lsu rDNA sequences were not downloaded because they do not meet the minimum length criterion (800 base pairs). The screening functions of *mor* rejected 177 of the downloaded sequences, most often because they are more than 99.5% similar to conspecific sequences already in the dataset (85 sequences) or include too many ambiguous sites (79 sequences). Two thousand two hundred thirty-nine sequences have been aligned and analyzed, which represent 1774 species in 537 genera (about 10% of the described species of homobasidiomycetes), and parsimony trees of 45,036 steps (CI = 0.09721; RI = 0.80799) have been produced. Seventeen clades have been defined, based primarily on major clades recognized by Hibbett and Thorn (2001), Larsson et al. (2004), and Binder et al. (2005) (Fig. 3). The clades that have been delimited in *mor* have been given informal names, consistent with their use in the publications of Hibbett and Thorn (2001), Larsson et al. (2004), and Binder et al. (2005). Eventually, it should be possible to define clades in *mor* in a way that will correspond to formal taxon definitions under the PhyloCode (http://www.ohiou.edu/phylocode/).

Some aspects of the topology produced by *mor* appear to be erroneous, which is not surprising give the size of the dataset and the limitations of the analysis. For example, several exemplars of the genera *Amanita* and *Entoloma*, which are gilled mushrooms, and *Trametes* and *Pycnoporus*, which are poroid forms, are placed as close relatives of the "jelly fungi" Auriculariales s. lat., which is highly unlikely (Fig. 3). Nevertheless, in general, the composition of the clades agrees well with their limits as estimated in the phylogenetic studies cited above. The largest clade of homobasidiomycetes delimited in *mor* is the euagarics clade (Agaricales), which is represented by 1219 sequences, whereas the smallest is the athelioid clade, which is represented by three sequences.

Conclusions and Future Directions

The *mor* system demonstrates that the core elements of phylogenetic taxonomy can be automated. This system also demonstrates an advantage of rank-free taxonomy relative to traditional taxonomy, which is that taxa in a rank-free system are amenable to algorithmic interpretation. Of course, the groups delimited by *mor* could be classified using Linnaean categories. However, that would negate another advantage of rank-free classification, which is that the names of taxa are stable in the face of rearrangements in tree topologies. This last feature is critical for automated taxonomic systems, because trees may change from week to week,


Euagarics clade (1234)
Athelioid clade (3)
Bolete clade (308)
Polyporoid clade (182)
Trechisporoid clade (20)
Thelephoroid clade (12)
Corticioid clade (11)
Gloeophyllum clade (6)
Russuloid clade (160)
Hymenochaetoid clade (184)
Gomphoid-phalloid clade (25)
Auriculariales (9)
Trametes/Polyporus? (19)
Entoloma/Amanita? (21)
Cantharelloid clade (54)
Sebacinales (3)
Dacrymycetales (10)
Tremellales (2)

resulting in arrangements that violate the hierarchy of Linnaean ranks. Conceivably, algorithms could be devised that would detect conflicts between the hierarchy of Linnaean ranks and the nested relationships among clades, but automated solutions to such conflicts are probably not a realistic possibility (as shown by the *Coprinus* example). If taxonomy is to be fully automated, it will be necessary to adopt a system of rank-free classification.

One could argue that automatically updated taxonomy would actually impede communication among systematists, because the compositions of taxa would be unstable (although their phylogenetic definitions would be conserved). This problem could be circumvented by archiving trees, as is done in *mor*. Systematists would then have to refer to taxonomic concepts in the context of a specific phylogenetic tree, identified by its date of deposition. This would not be so different from the present situation, in which the compositions of taxa often cannot be known unless a specific taxonomic work is cited (for example, Russulales sensu Hawksworth et al. [1995] is not equivalent to Russulales sensu Kirk et al. [2001]).

Many enhancements are possible in the alignment and phylogenetic reconstruction procedures used in *mor*. For example, the parsimony analysis in *mor* is susceptible to being trapped by local optima (because it amounts to one long single-threaded heuristic search), and should be replaced by methods that have been designed for large datasets, such as the parsimony ratchet (Goloboff, 1999; Nixon, 1999). Even with such improvements, *mor* is unlikely to provide state-of-the-art analyses, because of the restrictions imposed by such a large data set, as well as its reliance on a single locus (nuc-lsu rDNA). Nevertheless, *mor* can incorporate insights into phylogeny resulting from more sophisticated analyses, including multilocus studies, which can be used to guide refinements of the general constraint.

Another weakness of *mor* is that it does not incorporate the large database of sequences from the internal transcribed spacer (ITS) regions of nuclear rDNA. ITS is more variable than nuc-lsu rDNA and it has become a preferred region for studies at low taxonomic levels and for identification purposes in molecular ecology (Horton and Bruns, 2001). As of November 2004, there were at least 8209 homobasidiomycete ITS sequences in GenBank. Addition of these data to *mor* will greatly improve resolution of the terminal clades of homobasidiomycetes. However, ITS is too divergent to be aligned across all groups of homobasidiomycetes, so it will be necessary to create clade-specific ITS data sets and trees, and combine them with the nuc-lsu rDNA tree using supertree methods (Sanderson et al., 1998; Bininda-Emonds, 2004). A major challenge of such analyses will be to develop automated protocols for determining one-to-one correspondence (ontology) of ITS and nuc-lsu rDNA sequences.

Even with the incorporation of ITS data, the classification generated by *mor* would be restricted to organisms that are represented by sequences in GenBank. This limitation could also be overcome, however, because taxonomic hierarchies have an inherent tree structure, which could be combined with sequence-based trees using supertree methods. Again, establishment of correspondences among sequences and named species will be challenging (especially in cases where the type specimen has not been sequenced). Still, in principle, it should be possible to develop automated methods that produce continuously updated classifications that embody the total knowledge of biodiversity and phylogeny, based on molecular data and traditional taxonomy.

FIGURE 3. Overview of the 2239-sequence MP tree produced by *mor* on October 29, 2004. The clades that have been delimited in *mor* are labeled and the branches that support them are drawn with heavy lines. Numbers in parentheses are the numbers of sequences in each group. The trechisporoid clade is nested within polyporoid clade, and the athelioid clade is nested within the euagarics clade. The positions of the *"Entoloma/Amanita"* group and the *"Trametes/Polyporus"* group (which are not defined clades) are probably artifacts, as discussed in the text. A fully expanded view of the tree is available at http://mor.clarku.edu.

ACKNOWLEDGMENTS

We are grateful to the National Science Foundation, which supported this work through the "Assembling the Fungal Tree of Life" project (award DEB-0228657), including Research Experiences for Undergraduates supplements that supported M. Snyder and J. Costanzo. We also thank Manfred Binder, Olaf Bininda-Emonds, P. Brandon Matheny, Scott Redhead, Nicolas Salamin, Vincent Savolainen, and an anonymous reviewer for their helpful comments, and the Department of Mathematics and Computer Science of Clark University for access to computer facilities.

REFERENCES

Binder, M., and D. S. Hibbett. 2002. Higher-level phylogenetic relationships of homobasidiomycetes (mushroom-forming fungi) inferred from four rDNA regions. Mol. Phylogenet. Evol. 22:76–90.

Binder, M., D. S. Hibbett, K.-H. Larsson, E. Larsson, E. Langer, and G. Langer. 2005. The phylogenetic distribution of resupinate forms across the major clades of mushroom-forming fungi (Homobasidiomycetes). Syst. Biodivers. 3:113–157.

Bininda-Emonds, O. R. P. 2004. Phylogenetic supertrees: Combining information to reveal the Tree of Life. Kluwer Academic, Dordrecht, The Netherlands.

Bridge, P. D., P. J. Roberts, B. M. Spooner, and G. Panchal. 2003. On the unreliability of published DNA sequences. New Phytol. 160:43–48.

de Queiroz, K., and J. Gauthier. 1992. Phylogenetic taxonomy. Annu. Rev. Ecol. Syst. 23:449–480.

Eddy, S. R. 1998. Profile hidden Markov models. Bioinformatics 14:755–763.

Felsenstein, J. 1989. PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics 5:164–166.

Goloboff, P. A. 1999. Analyzing large datasets in reasonable times: Solutions for composite optima. Cladistics 15:415–428.

Greuter, W., J. Mcneill, F. R. Barrie, H.-M. Burdet, V. Demoulin, T. S. Filgueiras, D. H. Nicolson, P. C. Silva, J. E. Skog, P. Trehane, N. J. Turl, and, D. L. Hawksworth. 1999. International Code of Botanical Nomenclature (St Louis Code). Regnum Vegetabile 138. Koeltz Scientific Books, Königstein, Germany.

Hawksworth, D. L. 2001. The magnitude of fungal diversity: The 1.5 million species estimate revisited. Mycol. Res. 105:1422–1432.

Hawksworth, D. L., P. M. Kirk, B. C. Sutton, and D. N. Pegler. 1995. Ainsworth and Bisby's Dictionary of the Fungi, 8th ed. CAB International, Wallingford, UK.

Hibbett, D. S., and Binder, M. 2002. Evolution of complex fruiting body morphologies in homobasidiomycetes. Proc. R. Soc. Lond. B 269:1963–1969.

Hibbett, D. S., and M. J. Donoghue. 1998. Integrating phylogenetic analysis and classification in fungi. Mycologia 90:347–356.

Hibbett, D. S., and Thorn, R.G. 2001. Basidiomycota: Homobasidiomycetes. Pages 121–168 *in* The Mycota VII, systematics and evolution, Part B (D. J. McLaughlin, E. G. McLaughlin, and P. A. Lemke, eds.) Springer Verlag, Berlin.

Hopple, J. S., Jr., and R. Vilgalys. 1994. Phylogenetic relationships among coprinoid taxa and allies based on data from restriction site mapping of nuclear rDNA. Mycologia 86:96–107.

Horton, T. R., and T. D. Bruns. 2001. The molecular revolution in ectomycorrhizal ecology: Peeking into the black-box. Mol. Ecol. 10:1855–1871.

Jørgensen, P. M., S. Ryman, W. Gams, and J. A. Stalpers. 2001. (1486) Proposal to conserve the name *Coprinus* Pers. (*Basidiomycota*) with a conserved type. Taxon 50:909–910.

Kirk, P. M., P. F. Cannon, J. C. David, and J. A. Stalpers. 2001. Ainsworth and Bisby's Dictionary of the Fungi, 9th ed. CAB International, Wallingford, UK.

Langer, E. 2002. Phylogeny of non-gilled and gilled basidiomycetes: DNA sequence inference, ultrastructure and comparative morphology. Habilitationsschrift, Universität Tübingen, Tübingen, Germany.

Larsson, K.-H., E. Larsson, and U. Kõljalg. 2004. High phylogenetic diversity among corticioid homobasidomycetes. Mycol. Res. 108:983–1002.

Lutzoni, F., F. Kauff, C. J. Cox, D. McLaughlin, G. Celio, B. Dentinger, M. Padamsee, D. Hibbett, T. Y. James, E. Baloch, M. Grube, V. Reeb, V. Hofstetter, C. Schoch, A. E. Arnold, J. Miadlikowska, J. Spatafora, D. Johnson, S. Hambleton, M. Crockett, and R. Shoemaker, G.-H. Sung, R. Lücking, T. Lumbsch, K. O'Donnell, M. Binder, P. Diederich, D. Ertz, C. Gueidan, K. Hansen, R. C. Harris, K. Hosaka, Y.-W. Lim, B. Matheny, H. Nishida, D. Pfister, J. Rogers, A. Rossman, I. Schmitt, H. Sipman, J. Stone, J. Sugiyama, R. Yahr, R. Vilgalys. 2004. Where are we in assembling the fungal tree of life, classifying the fungi, and understanding the evolution of their subcellular traits? Am. J. Bot. 91:1446–1480.

Maddison, D. R., and W. P. Maddison. 2001. MacClade 4: Analysis of phylogeny and character evolution. Version 4.02. Sinauer Associates, Sunderland, Massachusetts.

Maddison, D. R., D. L. Swofford, and W. P. Maddison. 1997. NEXUS: An extensible file format for systematic information. Syst. Biol. 46:590–621.

Matheny, P. B., Y. J. Liu, J. F. Ammirati, and B. D. Hall. 2002. Using RPB1 sequences to improve phylogenetic inference among mushrooms (*Inocybe*, Agaricales). Am. J. Bot. 89:688–698.

Moncalvo, J.-M., R. Vilgalys, S. A. Redhead, J. E. Johnson, T. Y. James, M. C. Aime, V. Hofstetter, S. J. W. Verduin, E. Larsson, T. J. Baroni., R. G. Thorn, S. Jacobsson, H. Clémençon, and O. K. Miller, Jr. 2002. One hundred and seventeen clades of euagarics. Mol. Phylogenet. Evol. 23:357–400.

Nixon, K. C. 1999. The Parsimony Ratchet, a new method for rapid parsimony analysis. Cladistics 15:407–414

Redhead, S. A., R. Vilgalys, J.-M. Moncalvo, J. Johnson, and J. S. Hopple, Jr. 2001. *Coprinus* Pers. and the disposition of *Coprinus* species *sensu lato*. Taxon 50:203–241.

Sanderson, M., A. Purvis, and C. Henze. 1998. Phylogenetic supertrees: Assembling the trees of life. Trends Ecol. Evol. 13:105–109.

Schadt, C. W., A. P. Martin, D. A. Lipson, and S. K. Schmidt. 2003. Seasonal dynamics of previously unknown fungal lineages in tundra soils. Science 301:1359–1361.

Swofford, D. L. 2003. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acid Res. 11:4673–4680.

Vandenkoornhuyse, P., S. L. Baldauf, C. Leyval, J. Straczek, and J. P. W. Young. 2002. Extensive fungal diversity in plant roots. Science 295:2051–2051.

Vilgalys, R. 2003. Taxonomic misidentification in public DNA databases. New Phytol. 160:4–5.

Wang, Z., M. Binder, Y.-C. Dai, and D. S. Hibbett. 2004. Phylogenetic relationships of *Sparassis* inferred from nuclear and mitochondrial ribosomal DNA and a protein-coding gene (rpb2). Mycologia 96:1013–1027.

Weiss, M., and F. Oberwinkler. 2001. Phylogenetic relationships in Auriculariales and related groups–hypotheses derived from nuclear ribosomal DNA sequences. Mycol. Res. 105:403–415.

First submitted 14 September 2004; reviews returned 2 November 2004; final acceptance 19 November 2004
Associate Editor: Vincent Savolainen

Syst. Biol. 54(4):668–672, 2005
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150590947113

Getting to the Roots of Matrix Representation

Olaf R. P. Bininda-Emonds,[1] Robin M. D. Beck,[2,3] and Andy Purvis[2]

[1]*Lehrstuhl für Tierzucht, Technical University of Munich, Hochfeldweg 1, 85354 Freising-Weihenstephan, Germany; E-mail: Olaf.Bininda@tierzucht.tum.de (O.R.P.B.-E.)*
[2]*Department of Biological Sciences, Imperial College London, Silwood Park Campus, Ascot SL5 7PY, United Kingdom; E-mail: robin.beck@student.unsw.edu.au (R.M.D.B.); a.purvis@imperial.ac.uk (A.P.)*
[3]*The Natural History Museum, Cromwell Road, London SW7 5BD, United Kingdom*

Many supertree methods rely on the matrix representation (MR) of relationships in a set of source trees. The most common coding method is based on additive binary coding (Farris et al., 1970): for each informative node in a source tree (i.e., ones that correspond to a parsimony-informative character), taxa that are descended from that node are scored as 1; those that are not, but are present on the tree are scored as 0; and those that are absent on that source tree, but present on other trees in the set are scored as ?. MR supertree construction has usually been performed using source trees that are rooted. This rooting can be accomplished either by re-rooting all trees using a single taxon common to all

Olaf R. P. Bininda-Emonds and Robin M. D. Beck contributed equally to this work.