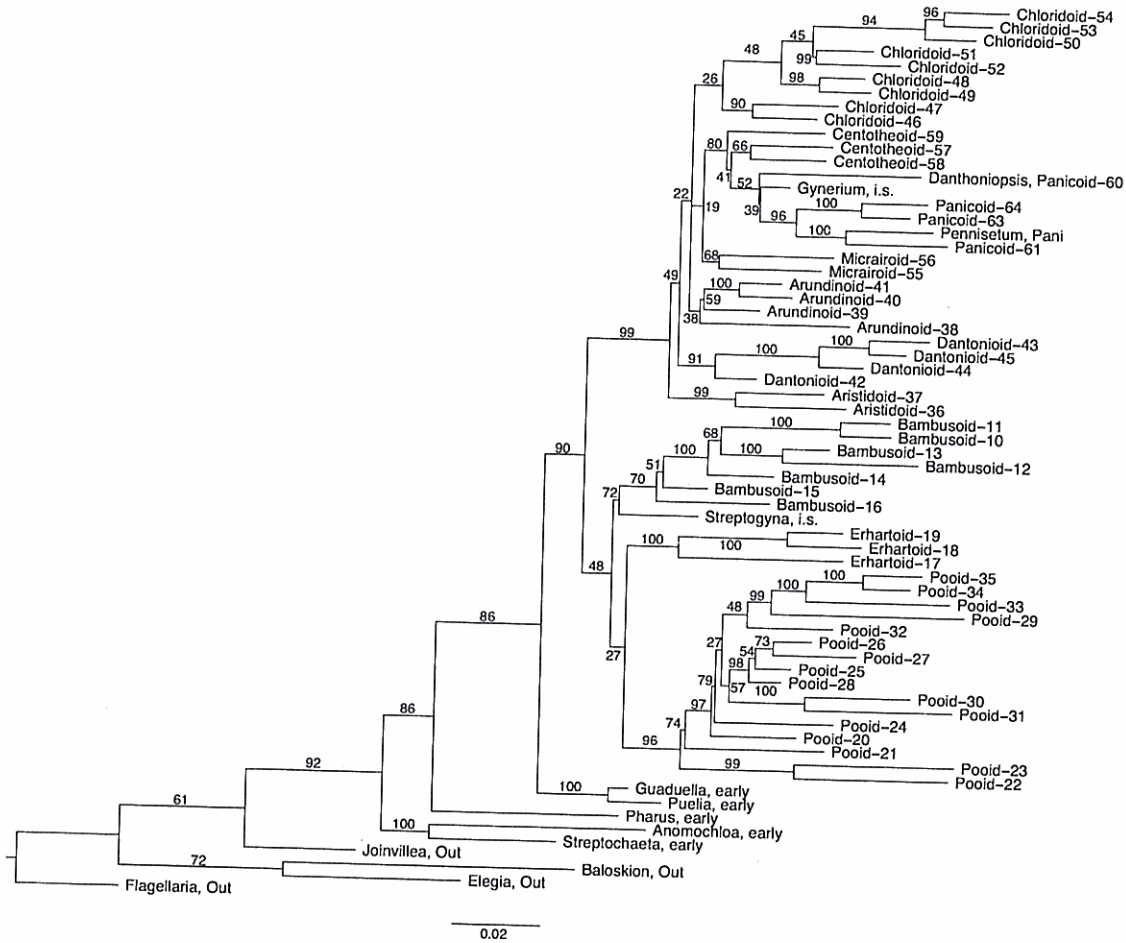

Figure 12: Poaceae phylogeny reconstructed by the medium level approach SuperQP. The numbers at the internal nodes represent the percent occurrence of the corresponding splits in the puzzling step trees, branch lengths were computed automatically from the concatenated alignment. Subfamily names are abbreviated after four letters.

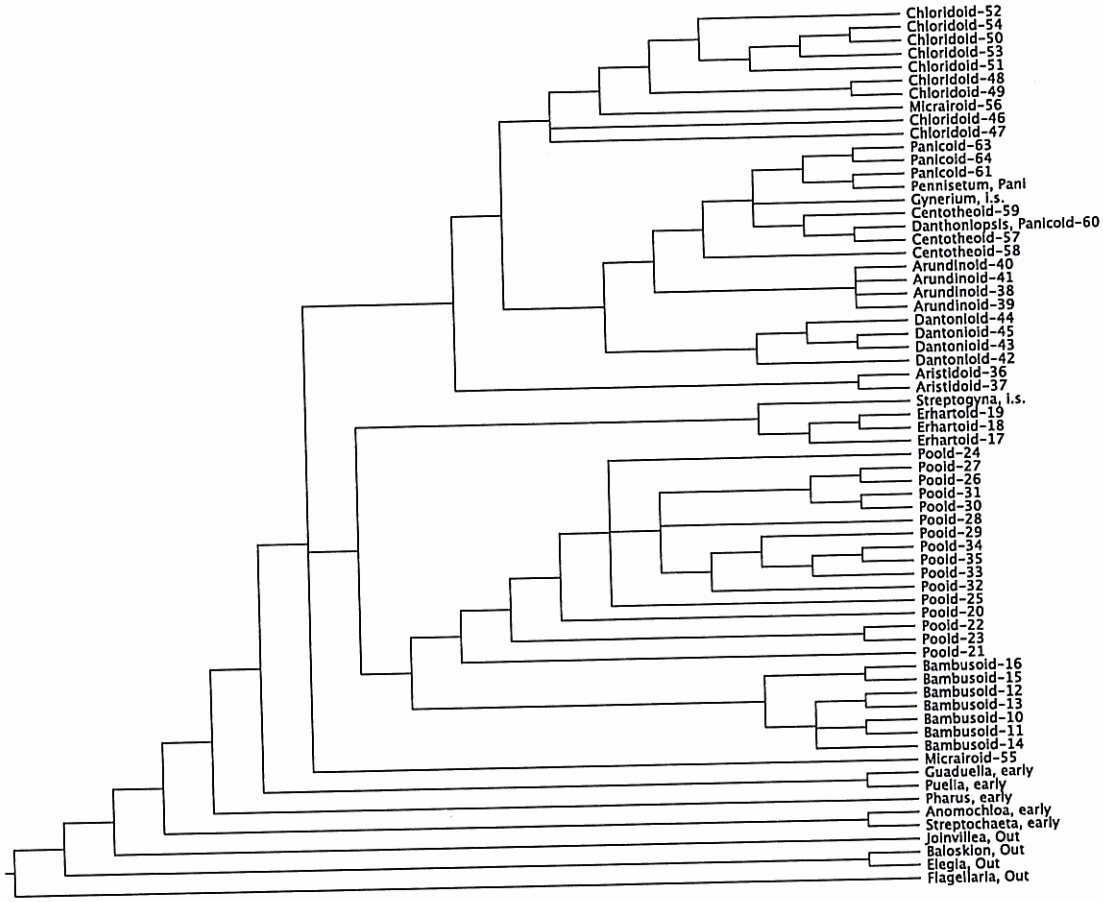

Figure 11: ModMinCut supertree of the Poaceae dataset based on the ML gene trees. Subfamily names are abbreviated after four letters.

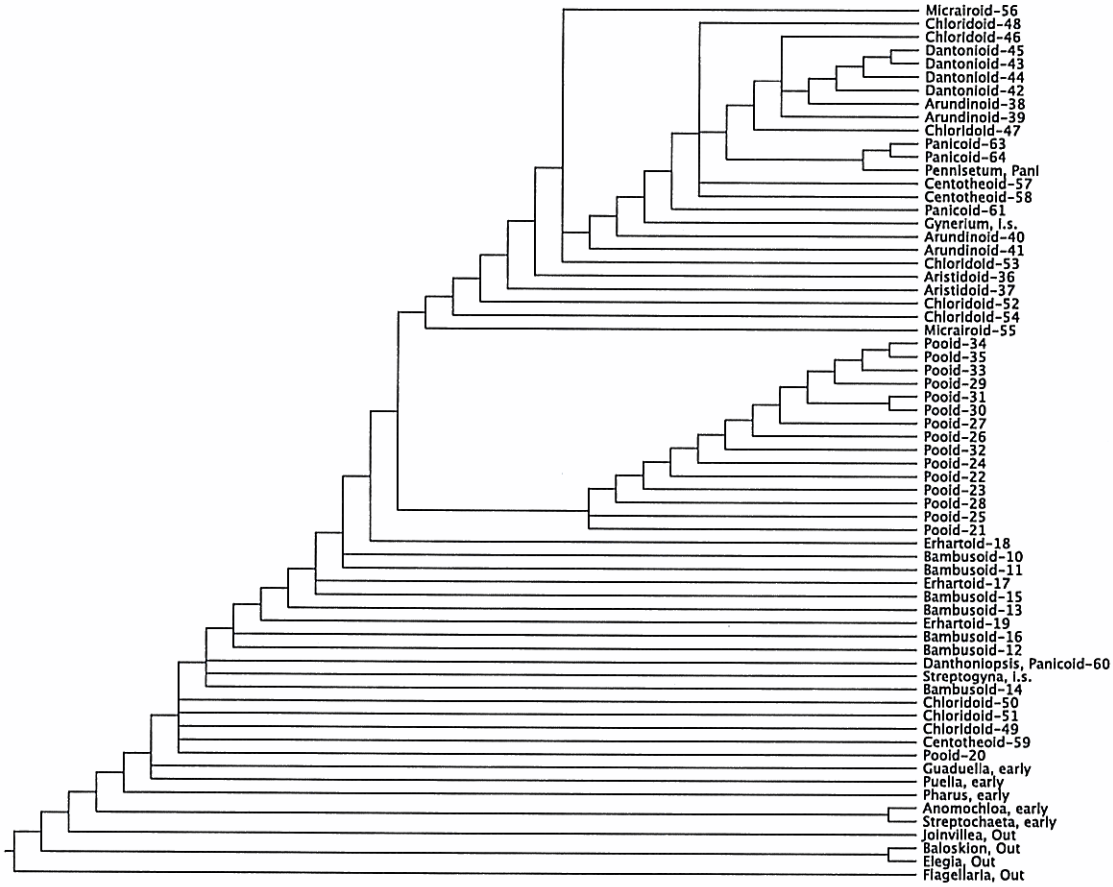

Figure 10: MinCut supertree of the Poaceae dataset based on the ML gene trees. Subfamily names are abbreviated after four letters.

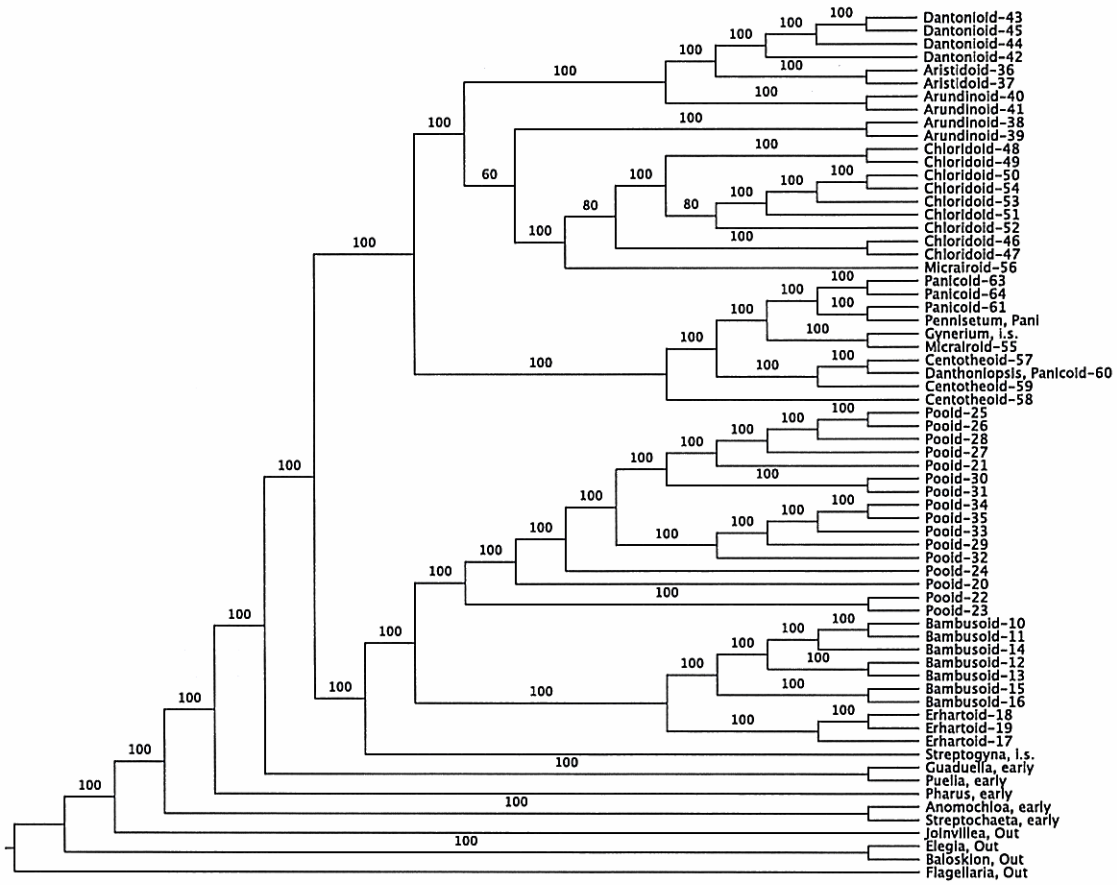

Figure 9: MRF phylogeny of the Poaceae dataset based on ML gene trees and Purvis' coding scheme (MRF-Pu). Five best trees found with same score by the MRF heuristics were summarized by the M_{50} consensus. Subfamily names are abbreviated after four letters.

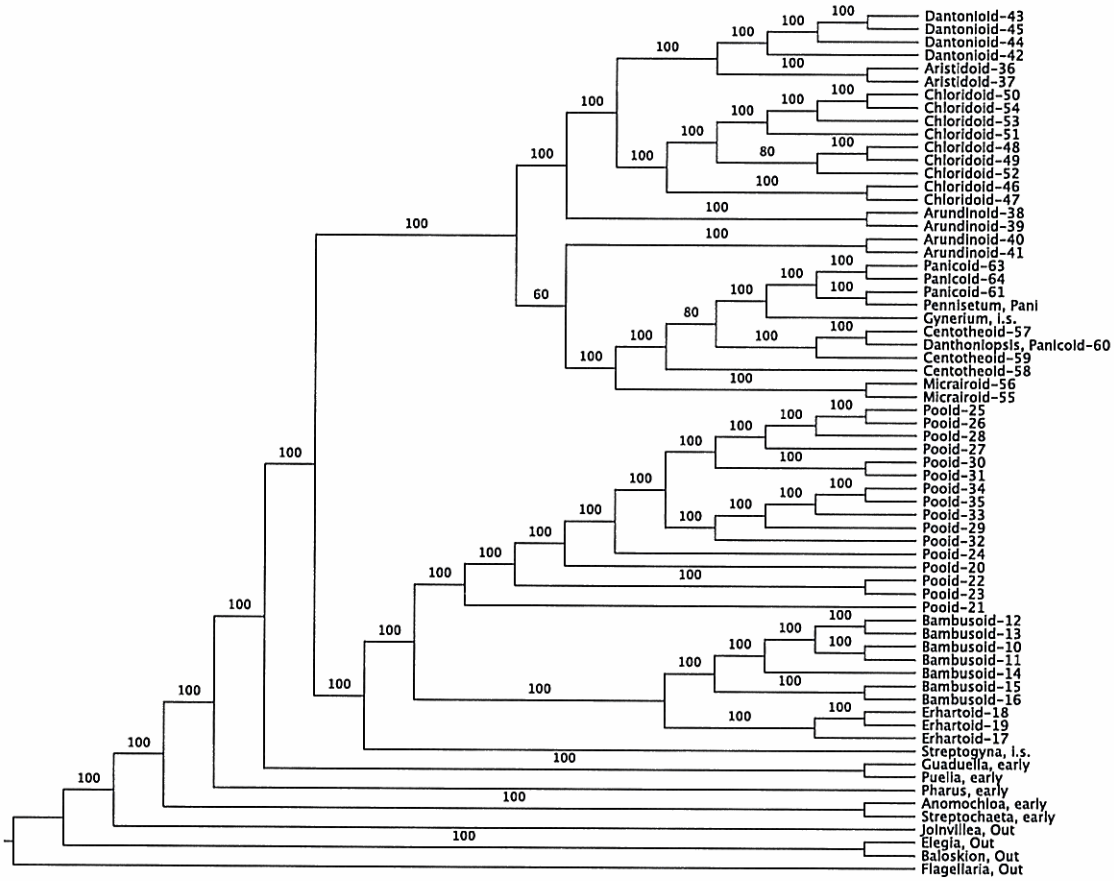

Figure 8: MRF phylogeny from the Poaceae dataset based on ML trees and the Baum/Ragan coding scheme (MRF-BR). Five best trees found with same score by the MRF heuristics were summarized by the M_{50} consensus. Subfamily names are abbreviated after four letters.

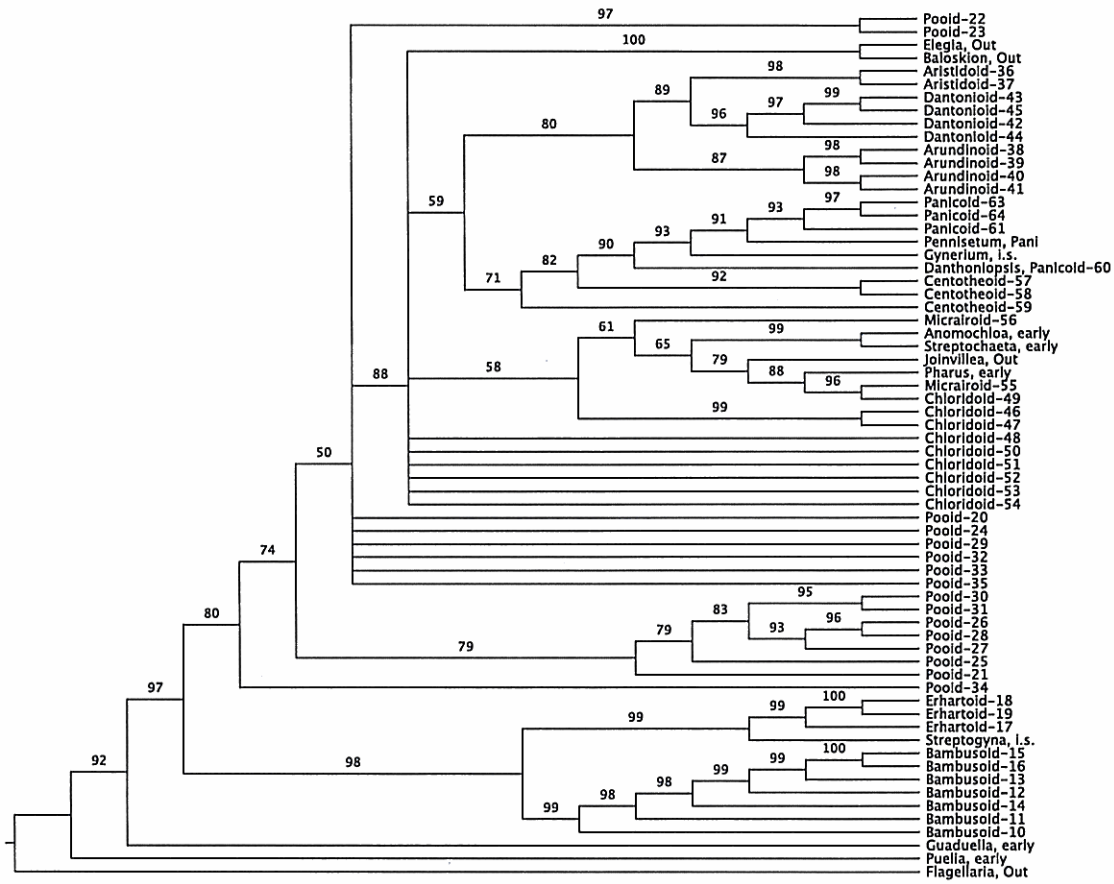

Figure 7: MRP phylogeny of the Poaceae dataset based on ML gene trees and Purvis' coding scheme (MRP-Pu). All equally most parsimonious trees found were summarized by the M_{50} consensus. Subfamily names are abbreviated after four letters.

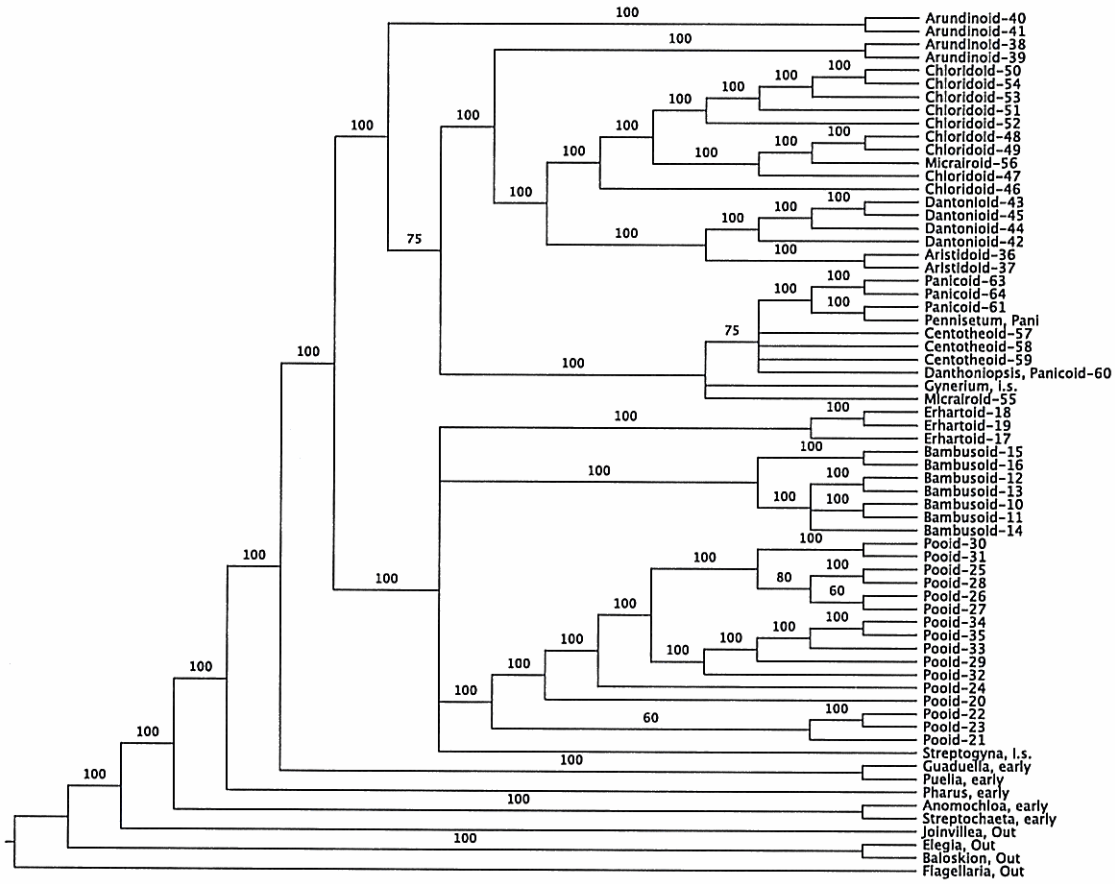

Figure 6: MRP phylogeny from the Poaceae dataset based on ML trees and the Baum/Ragan coding scheme (MRP-BR). All equally most parsimonious trees found were summarized by the M_{50} consensus. Subfamily names are abbreviated after four letters.

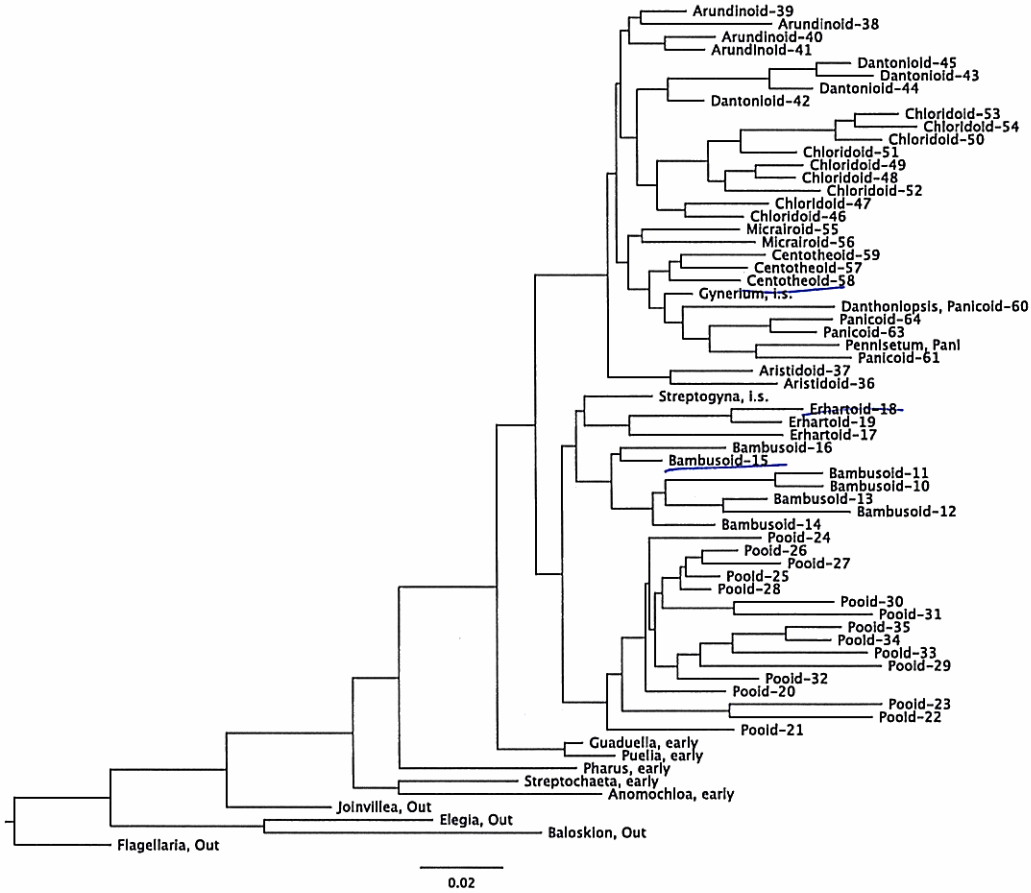

Figure 5: Poaceae ML phylogeny based on the superalignment of the GPWG dataset. Subfamily names are abbreviated after four letters.

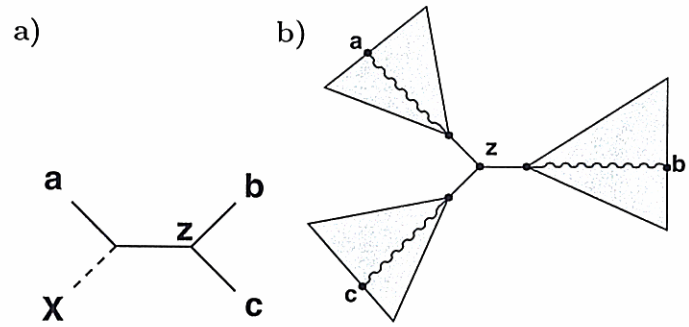

Figure 4: Insertion scores. a) Informative quartet for inserting taxon X, i.e. taxa a, b, and c are already in the tree; b) triplet induced by a, b, and c in their disjoint subtrees and the central node z.

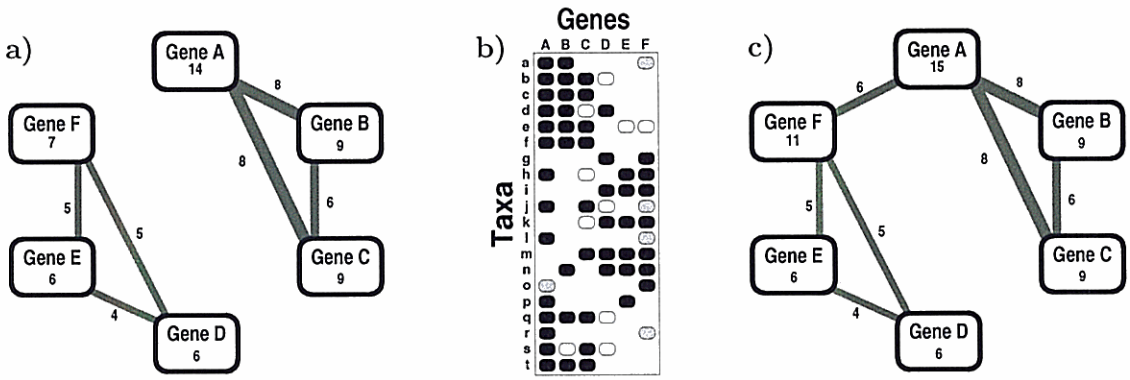

Figure 3: Overlap Graphs: example graphs for datasets of 6 genes with different degrees of completeness. Unconnected overlap graph (a) of the sequences represented by the black dots in the availability matrix (b). Adding the sequences presented by the gray dots (b) produces a single connected component (c). Adding also the white dots would produce an almost completely connected graph missing only overlap among genes B and E.

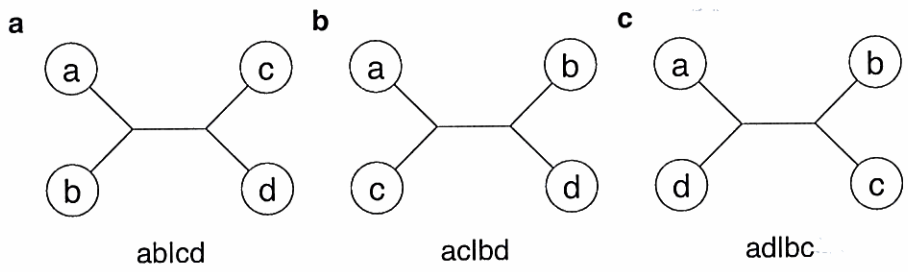

Figure 2: Quartet trees. a)-c) The three informative quartet topologies for quartet $\{a, b, c, d\}$.

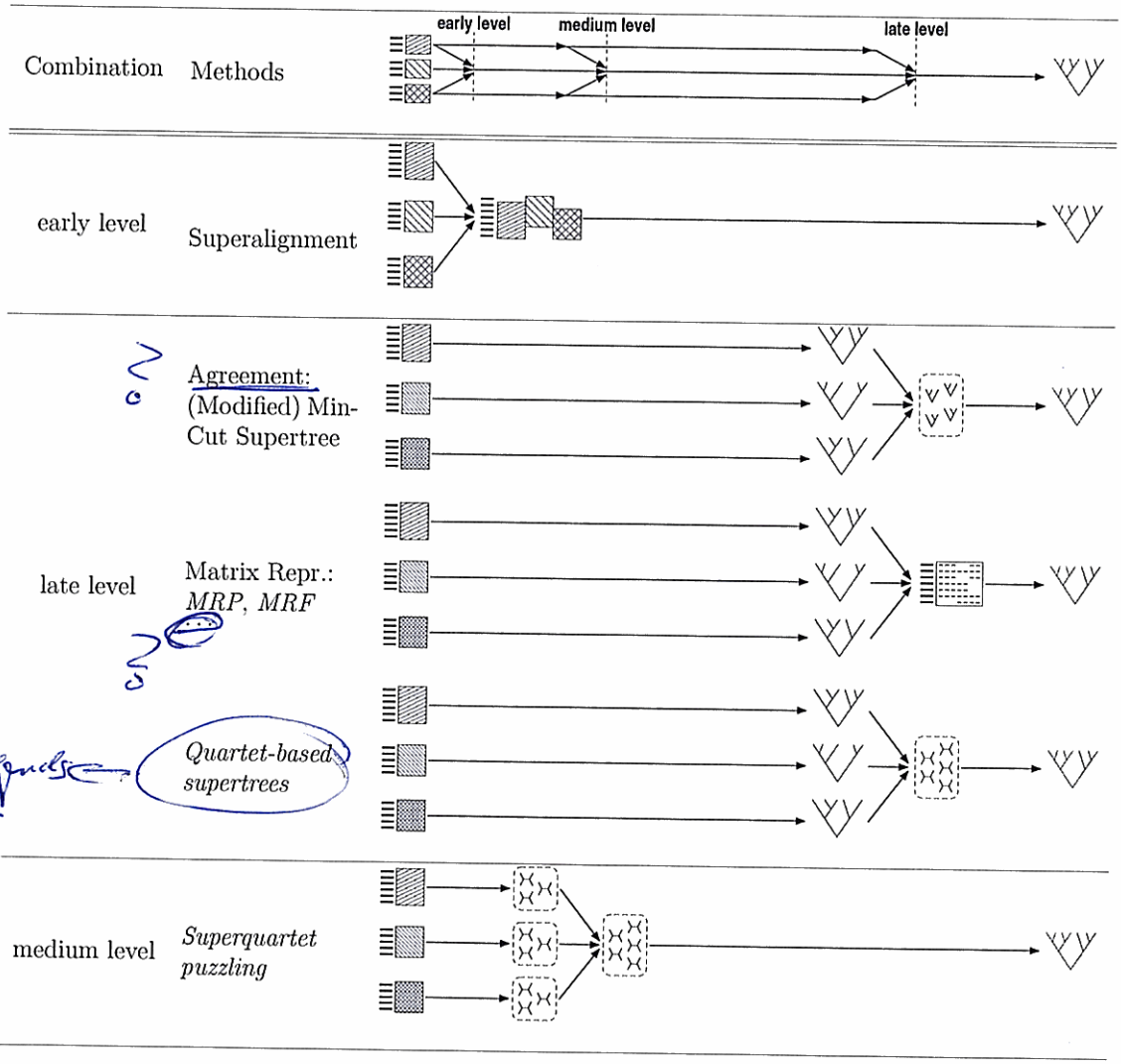

Figure 1: Level of combination with regard to distance from the underlying datasets

Table 2: Reconstructed Poaceae Subfamilies. Results from combined tree reconstruction using superalignment, MRP and MRF with Baum/Ragan (MRP-BR, MRF-BR) and Purvis' (MRP-Pu, MRF-Pu) coding, MinCut and ModMinCut ~~supertrees~~, and SuperQP. '•' denotes monophyletic reconstructed subfamilies. If the Panicoideae are marked 'o', only Panicoideae *sensu stricto* were recovered, i.e. without *Danthoniopsis*.

	superalignment	MRP-BR	MRP-Pu	MRF-BR	MRF-Pu	MinCut	ModMinCut	SuperQP
PACCMAD clade:	•	•		•	•			•
- Panicoideae	•	o	•	o	o		o	•
- Aristidoideae	•	•	•	•	•		•	•
- Centothecoideae	•							
- Chloridoideae	•			•	•			•
- Micrairoideae	•			•				•
- Arundinoideae	•		•				•	•
- Danthonioideae	•	•	•	•	•	•	•	•
BEP clade:	•	•		•	•		•	•
- Bambusoideae	•	•	•	•	•		•	•
- Erhartoideae	•	•	•	•	•		•	•
- Pooideae	•	•		•	•	•	•	•
early grasses	•	•		•	•		•	•
outgroups basal	•	•		•	•	•	•	•

Table 1: Sequences available for analysis per taxon in the Poaceae dataset (Grass Phylogeny Working Group, 2001)

	ndhF	*rbcL*	*rpoC*	*phyB*	*waxy*	ITS
Outgroup:						
1 *Flagellaria*	X	X	-	X	-	-
2 *Elegia*	X	X	-	-	-	-
3 *Baloskion*	X	X	-	-	-	-
4 *Joinvillea*	X	X	X	X	-	X
Early grasses:						
5 *Anomochloa*	X	X	-	X	X	-
6 *Streptochaeta*	X	-	-	X	-	X
7 *Pharus*	X	-	-	X	X	X
8 *Guaduella*	X	X	-	-	-	-
9 *Puelia*	X	X	-	X	-	-
Bambusoideae:						
10 *Eremitis*	X	-	-	X	X	-
11 *Pariana*	X	-	-	X	X	-
12 *Lithachne*	X	X	-	X	-	X
13 *Olyra*	X	-	X	X	-	-
14 *Buergersiochloa*	X	-	-	X	-	-
15 *Pseudosasa*	X	X	X	X	-	-
16 *Chusquea*	X	X	-	X	X	X
Ehrhartoideae:						
17 *Ehrharta*	X	-	X	X	-	X
18 *Oryza*	X	X	X	X	X	X
19 *Leersia*	X	X	-	-	-	X
Pooideae:						
20 *Phaenosperma*	X	-	-	-	-	-
21 *Brachyelytrum*	X	-	-	-	-	X
22 *Lygeum*	X	-	X	X	X	X
23 *Nardus*	X	-	X	X	-	X
24 *Anisopogon*	X	-	X	X	-	X
25 *Ampelodesmos*	X	-	-	-	-	X
26 *Stipa*	X	X	X	-	-	X
27 *Nassella*	X	-	-	X	-	X
28 *Piptatherum*	X	-	-	-	-	X
29 *Brachypodium*	X	-	-	X	-	X
30 *Melica*	X	-	-	X	X	X
31 *Glyceria*	X	-	-	X	X	X
32 *Diarrhena*	X	-	-	X	-	X
33 *Avena*	X	X	X	X	-	X
34 *Bromus*	X	X	X	X	-	X
35 *Triticum*	X	X	-	X	X	X

	ndhF	*rbcL*	*rpoC*	*phyB*	*waxy*	ITS
Aristidoideae:						
36 *Aristida*	X	X	X	X	-	X
37 *Stipagrostis*	X	X	X	-	-	X
Arundinoideae:						
38 *Amphipogon*	X	X	X	-	-	X
39 *Arundo*	X	X	X	-	-	X
40 *Molinia*	X	X	X	X	-	X
41 *Phragmites*	X	X	X	X	-	X
Danthonioideae:						
42 *Merxmuellera m.*	X	X	X	-	X	X
43 *Karoochloa*	X	X	X	-	X	X
44 *Danthonia*	X	X	X	X	-	X
45 *Austrodanthonia*	X	X	X	-	X	X
Chloridoideae:						
46 *Merxmuellera r.*	X	X	X	-	X	X
47 *Centropodia*	X	X	X	-	X	X
48 *Eragrostis*	X	X	X	X	-	X
49 *Uniola*	X	-	-	-	-	-
50 *Zoysia*	X	-	-	-	-	-
51 *Distichlis*	X	-	-	-	-	-
52 *Pappophorum*	X	X	X	-	-	-
53 *Spartina*	X	-	X	-	-	X
54 *Sporobolus*	X	-	-	X	-	X
Micrairoideae:						
55 *Eriachne*	-	X	-	-	-	X
56 *Micraira*	X	-	X	-	-	X
Centothecoideae:						
57 *Thysanolaena*	X	X	X	X	-	X
58 *Chasmanthium*	X	X	X	X	-	X
59 *Zeugites*	X	-	-	-	-	-
Panicoideae:						
60 *Danthoniopsis*	X	-	-	X	X	-
61 *Panicum*	X	-	X	X	-	X
62 *Pennisetum*	X	X	X	X	X	X
63 *Miscanthus*	X	X	X	X	X	X
64 *Zea*	X	X	X	X	X	X
Incertae sedis:						
65 *Gynerium, i.s.*	X	X	X	-	-	X
66 *Streptogyna, i.s.*	X	-	-	X	-	-

Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42:587–596.

Zgurski, J. M., H. S. Rai, Q. M. Fai, D. J. Bogler, J. Francisco-Ortega, and S. W. Graham. 2008. How well do we understand the overall backbone of cycad phylogeny? new insights from a large, multigene plastid data set. Mol. Phylogenet. Evol. 47:1232–1237.

Sanderson, M. J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19:301–302.

Sanderson, M. J., C. Ané, O. Eulenstein, D. Fernández-Baca, J. Kim, M. M. McMahon, and R. Piaggio-Talice. 2007. Fragmentation of large data sets in phylogenetic analyses. Pages 199–216 *in* Mathematics of Evolution and Phylogeny (O. Gascuel, ed.). Oxford University Press, Oxford, UK.

Sanderson, M. J. and A. C. Driskell. 2003. The challenge of constructing large phylogenetic trees. Trends Plant Sci. 8:374–379.

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

Semple, C. and M. Steel. 2000. A supertree method for rooted trees. Discr. Appl. Math. 105:147–158.

Strimmer, K., N. Goldman, and A. von Haeseler. 1997. Bayesian probabilities and quartet puzzling. Mol. Biol. Evol. 14:210–213.

Strimmer, K. and A. von Haeseler. 1996. Quartet puzzling: A quartet maximum–likelihood method for reconstructing tree topologies. Mol. Biol. Evol. 13:964–969.

Vinh, L. S. and A. von Haeseler. 2004. IQPNNI: Moving fast through tree space and stopping in time. Mol. Biol. Evol. 21:1565–1571.

Wilkinson, M., J. A. Cotton, and J. L. Thorley. 2004. The information content of trees and their matrix representations. Syst. Biol. 53:989–1001.

Meredith, R. W., M. Westerman, and M. S. Springer. 2008. A timescale and phylogeny for 'Bandicoots' (Peramelemorphia: Marsupialia) based on sequences for five nuclear genes. Mol. Phylogenet. Evol. 47:1–20.

Miyamoto, M. M. and W. M. Fitch. 1995. Testing species phylogenies and phylogenetic methods with congruence. Syst. Biol. 44:64–76.

Novacek, M. J. 2001. Mammalian phylogeny: Genes and supertrees. Curr. Biol. 11:R573–R575.

Page, R. D. M. 2002. Modified mincut supertrees. Pages 537–551 *in* Proceedings of the 2nd Workshop on Algorithms in Bioinformatics (WABI 2002) vol. 2452 of *Lecture Notes in Computer Science* Springer, New York.

Piaggio-Talice, R., G. Burleigh, and O. Eulenstein. 2004. Quartet supertrees. Pages 173–191 *in* Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life (O. R. P. Bininda-Emonds, ed.). Kluwer Academic, Dordrecht.

Purvis, A. 1995. A composite estimate of primate phylogeny. Philos. Trans. R. Soc. Lond. Ser. B 348:405–421.

Ragan, M. A. 1992. Phylogenetic inference based on matrix representation of trees. Mol. Phylogenet. Evol. 1:53–58.

Robinson-Rechavi, M. and D. Graur. 2001. Usage optimization of unevenly sampled data through the combination of quartet trees: An eutherian draft phylogeny based on 640 nuclear and mitochondrial proteins. Isr. J. Zool. 47:259–270.

Sánchez-Ken, J. G., L. G. Clark, E. A. Kellogg, and E. E. Kay. 2007. Reinstatement and emendation of subfamily micrairoideae (poaceae). Syst. Bot. 32:71–80.

Gatesy, J., R. H. Baker, and C. Hayashi. 2004. Inconsistencies in arguments for the supertree approach: Supermatrices versus supertrees of Crocodylia. Syst. Biol. 53:342–355.

Gatesy, J., C. Matthee, R. DeSalle, and C. Hayashi. 2002. Resolution of a supertree/supermatrix paradox. Syst. Biol. 51:652–664.

Gatesy, J. and M. S. Springer. 2004. A critique of matrix representation with parsimony supertrees. Pages 369–388 *in* Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life (O. R. P. Bininda-Emonds, ed.). Kluwer Academic, Dordrecht, The Netherlands.

Goldman, N., J. P. Anderson, and A. G. Rodrigo. 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. 49:652–670.

Gordon, A. D. 1986. Consensus supertrees: The synthesis of rooted trees containing overlapping sets of labelled leaves. J. Classif. 3:335–348.

Grass Phylogeny Working Group. 2001. Phylogeny and subfamilial classification of the grasses (poaceae). Ann. Mo. Bot. Gard. 88:373–457.

Kluge, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). Syst. Zool. 38:7–25.

Kluge, A. G. 1998. Total evidence or taxonomic congruence: Cladistics or consensus classification. Cladistics 14:151–158.

Lewis, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50:913–925.

Discrete Mathematics and Theoretical Computer Science (M. F. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts, eds.) vol. 61. American Mathematical Society, Providence, Rhode Island.

Chen, D., O. Eulenstein, and D. Fernández-Baca. 2004. Rainbow: a toolbox for phylogenetic supertree construction and analysis. Bioinformatics 20:2872–2873.

Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein. 2001. Introduction to Algorithms. 2 ed. The MIT Press, Cambridge, Massachusetts.

de Queiroz, A., M. J. Donoghue, and J. Kim. 1995. Separate versus combined analysis of phylogenetic evidence. Annu. Rev. Ecol. Syst. 26:657–681.

Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6:361–375.

Dutilh, B. E., V. van Noort, R. T. J. M. van der Heijden, T. Boekhout, B. Snel, and M. A. Huynen. 2007. Assessment of phylogenomic and orthology approaches for phylogenetic inference. Bioinformatics 23:815–824.

Eernisse, D. and A. G. Kluge. 1993. Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. Mol. Biol. Evol. 10:1170–1195.

Eulenstein, O., D. Chen, J. G. Burleigh, D. Fernández-Baca, and M. J. Sanderson. 2004. Performance of flip supertree construction with a heuristic algorithm. Syst. Biol. 53:299–308.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Department of Genetics, University of Washington Seattle distributed by the author.

Bininda-Emonds, O. R. P. 2005. Supertree construction in the genomic age. Pages 745–757 *in* Molecular Evolution: Producing the Biochemical Data, Part B (E. A. Zimmer and E. Roalson, eds.) vol. 395 of *Methods Enzymol.* Elsevier, North Holland.

Bininda-Emonds, O. R. P., M. Cardillo, K. E. Jones, R. D. E. MacPhee, R. M. D. Beck, R. Grenyer, S. A. Price, R. A. Vos, J. L. Gittleman, and A. Purvis. 2007. The delayed rise of present-day mammals. Nature 446:507–512.

Bininda-Emonds, O. R. P., J. L. Gittleman, and M. A. Steel. 2002. The (super)tree of life: Procedures, problems, and prospects. Annu. Rev. Ecol. Syst. 33:265–289.

Bininda-Emonds, O. R. P., K. E. Jones, S. A. Price, M. Cardillo, R. Grenyer, and A. Purvis. 2004. Garbage in, garbage out: data issues in supertree construction. Pages 267–280 *in* Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life (O. R. P. Bininda-Emonds, ed.). Kluwer Academic, Dordrecht, The Netherlands.

Bouchenak-Khelladi, Y., N. Salamin, V. Savolainen, F. Forest, M. van der Bank, M. W. Chase, and T. R. Hodkinson. 2008. Large multi-gene phylogenetic trees of the grasses (Poaceae): Progress towards complete tribal and generic level sampling. Mol. Phylogenet. Evol. 47:488–505.

Bryant, D. and M. Steel. 1995. Extension operations on sets of leaf-labeled trees. Adv. Appl. Math. 16:425–453.

Chen, D., L. Diao, O. Eulenstein, D. Fernández-Baca, and M. J. Sanderson. 2003. Flipping: A supertree construction method. Pages 135–160 *in* DIMACS Series in

SPP-1174) is gratefully acknowledged.

References

Aho, A. V., Y. Sagiv, T. G. Szymanski, and J. D. Ullman. 1981. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. SIAM J. Comput. 10:405–421.

Angiosperm Phylogeny Group. 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. Bot. J. Linn. Soc. 141:399–436.

Baum, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41:3–10.

Bininda-Emonds, O. R. P. 2003a. MRP supertree construction in the consensus setting. Pages 231–242 *in* DIMACS Series in Discrete Mathematics and Theoretical Computer Science (M. F. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts, eds.) vol. 61. American Mathematical Society, Providence, Rhode Island.

Bininda-Emonds, O. R. P. 2003b. Supertrees are a necessary not-so-evil: A comment on gatesy et al. Syst. Biol. 52:724–729.

Bininda-Emonds, O. R. P., ed. 2004a. Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life. Kluwer Academic, Dordrecht.

Bininda-Emonds, O. R. P. 2004b. Trees versus characters and the supertree / supermatrix "paradox". Syst. Biol. 53:360–361.

Conclusion

We have presented a quartet-based medium-level method to combine multiple sequence data with missing data for phylogeny reconstruction. The results obtained from the GPWG Poaceae dataset show that superalignment and SuperQP performs best among the methods for phylogeny reconstruction from overlapping datasets with missing data tested here.

The early (superalignment) and medium level methods (SuperQP) have shown the best results for the Poaceae data. MRF-BR produced the best results among the supertree methods and the modifications to the MinCut algorithm suggested by Page (2002, ModMinCut supertree) have shown to substantially improve the MinCut result for the grass dataset. Yet, none of the methods to analyze incomplete phylogenetic data presented here has proven to be the final choice. Facing the problem that it is unlikely to have complete molecular and morphological datasets in the near future, methods combining overlapping incomplete datasets remain a valuable tool to reconstruct evolution from multiple datasets (cf. also Bininda-Emonds, 2003a). Always several of them should be used to compare the results to identify possible pitfalls. Tree reconstruction using superquartet puzzling (SuperQP) provides a new tool for such work on real data allowing to incorporate gene-specific constraints.

Acknowledgments

The authors thank Elizabeth A. Kellogg for providing access to the extended Poaceae data of the GPWG. Financial support of the WWTF (Wiener Wissenschafts-, Forschungs- und Technologiefonds) and DFG (German Research Foundation,

but superalignment were not able to reveal Centothecoideae monophyletic and some could only reconstruct the Panicoideae *sensu stricto* (without *Danthoniopsis*). All methods (except MinCut), however, support the monophyly of Centothecoideae with the Panicoideae and *Gynerium i.s.*. Only MRF-BR does also join *Eriachne* into the subtree.

The methods which were able to reconstruct best the grass subfamilies, PACCMAD, and BEP clade were superalignment, SuperQP, MRF-BR, and finally ModMinCut approaches. Early combination was able resolve all subfamilies followed by the medium-level method SuperQP before the late level supertree approaches.

Further extensions of the overlap graph.— Besides checking whether multi-gene dataset is connected by overlapping information and, hence, amenable for combination, the overlap graph can help to detect deficiencies of the available data matrix. By identifying minimum cuts (sets of edges with minimum total weights separating the graph into subgraphs) one might determine genes which would substantially improve the dataset, if added by directly sequencing in certain taxa,

We expect that the overlap graph-based insertion order might also improve other methods based on stepwise insertion due to the increased degree of information available for positioning taxa.

Furthermore, the order in which the taxon-subsets are chosen from the front in the overlap-guided puzzling step, might be improved by weighting the choice preferring taxon-sets with larger overlap or more taxa.

et al., 2000, for review). Such tests assess whether the computed likelihoods of reconstructed trees are significantly worse or better than others. Yet such tests require a model of evolution as well as a dataset that enables a comparison of topologies. Due to the large amount of missing data in the GPWG dataset (37.85% gaps and missing characters in the superalignment and a range of gaps and ambiguous characters from 5.42% in *Zea mays* up to 76.41% for *Eriachne*) and the different evolutionary constraints of genes this framework is not applicable for such data (cf. also Novacek, 2001). Hence, we had to rely on biological expert knowledge (like Grass Phylogeny Working Group, 2001; Sánchez-Ken et al., 2007; Bouchenak-Khelladi et al., 2008) to validate the results.

Examining the reconstructed trees shows that most methods can more or less resolve the taxonomic structure of the grasses (see Tab. 2). If subfamilies could not be reconstructed, their taxa are usually placed in multifurcations which do not contradict the taxonomic classification. Only MinCut supertree and MRP-Pu present completely displaced taxa and groups. While the problems of MinCut might be attributed to its resemblance to Adams consensus (Mike Steel, pers. comm.), MRP-Pu might be hampered by discarding crucial information in Purvis' coding (cf. Wilkinson et al., 2004). The latter caused the problem that MRP-Pu was unable to group the early and outgroup taxa together, thus, substantially perturbing the resulting tree topology. Interestingly, MRF seems not hampered by the Purvis encoding.

Furthermore, the phylogenetic status of the Micrairoideae remains unclear. Five of the methods position *Eriachne* and *Micraira* separate in the trees. The suggested positions vary among the five methods. The remaining three methods, superalignment, SuperQP, and MRF-BR, group the Micrairoideae together. Interestingly, these methods comprise one from every level of combination (early, medium, and late). All methods

using meta-data like matrix representation (MR) derived from the input trees, or they use the trees directly. The sequence data is usually not used in the combination process. Although this is an advantage when input trees obtained from literature are combined, it discards valuable information when the underlying sequence data is available. It has been stated by Bininda-Emonds (2003a, p. 275) that *"the inherent loss of information from using the source trees is a necessary trade-off to be able to combine all possible sources of phylogenetic information."* The medium level SuperQP, however, proposed in this article combines the datasets at the quartet level guided by the likelihood of underlying data and, thus, tries to reduce the loss of information in the combination process. Furthermore, the method offers the possibility to use appropriate models of evolution for the different sources of data. Yang (1996) has shown that combined analysis of datasets is possible even if the parameters as well as the branch lengths differ among the datasets. In general, it should even be possible to combine likelihoods of different types of data like DNA, amino acid, and restriction data. A prerequisite for combining such different data sources is the availability of applicable ML models to compute the quartet topologies. Recently, also new likelihood-based models for morphological data have been suggested (Lewis, 2001). The necessity to adequately normalize the substitution models for different types of data remains to be elucidated.

Evaluation of Reconstructed Trees.— We used the dataset of Poaceae to test our method in a real-world scenario. Fortunately, the grasses and their subfamilies are well-studied. Hence, we compare the trees constructed using different methods based on biological background knowledge by Grass Phylogeny Working Group (2001); Sánchez-Ken et al. (2007); Bouchenak-Khelladi et al. (2008).

Frequently the quality of trees reconstructed from real data is compared by testing the ML values of topologies in a maximum likelihood framework (see Goldman

grasses, and outgroups (Fig. 11, Tab. 2). If we ignored the position of the Micrairoid taxa, ModMinCut would, in addition, reveal the Chloridoideae and the PACCMAD clade itself.

SuperQP tree.— SuperQP was able to resolve all but one subfamily, the two major clades, and the early grasses (Fig. 12, Tab. 2). Even the Micrairoideae group could be recovered from the GPWG dataset. Only the Centothecoideae could not be established as by none of the supertree methods presented here.

Discussion

Problems of Dataset-Combining Methods.— There is an ongoing debate on the question whether early-level or late-level combination should be preferred (e.g. de Queiroz et al., 1995, and references therein). A main argument against superalignment methods is the problem how to choose an appropriate model of evolution for concatenated datasets. It is well-known that different areas of the genome ~~develop~~ evolve under different evolutionary constraints. It is hard if not impossible to choose one appropriate model for different sequences coding for proteins, structural or functional RNAs, and 'function-less' parts of the genome. Missing data also causes problems because some methods discard alignment columns containing gaps. In this case only few or even no alignment columns remain for the analysis. Nevertheless, the superalignment approach has been reported to frequently produce very good results (e.g., Dutilh et al., 2007).

Since consensus and supertree methods reconstruct gene trees separately, they can easily incorporate different evolutionary models matching the different constraints of genes. Supertree methods have been developed to handle the problem of missing data. For the reconstruction of an overall tree, consensus and supertree methods are

Micrairoid taxon *Micraira* would be regarded *incertae sedis* and, thus, its position ignored in the evaluation, also the Chloridoideae were reconstructed.

MRP supertree using Purvis coding.— The M_{50} consensus of 1 million reconstructed most parsimonious MRP-Pu trees shows some strange features. Outgroups and early grasses are scattered throughout the tree. Furthermore, the center part of the tree is more or less unresolved, possibly exhibiting a rooting problem. Nevertheless MRP-Pu was able to recover two BEP and four PACCMAD subfamilies (Fig. 7, Tab. 2).

MRF supertree using Baum/Ragan coding.— MRF-BR reconstructed both major clades, all BEP, and four PACCMAD subfamilies plus the Panicoideae *s.str.* group, early grasses and outgroups (Fig. 8, Tab. 2). It ~~even~~ also supported the Micrairoideae.

MRF supertree using Purvis coding.— MRF-Pu reconstructed almost all subfamilies and clades present in the MRF-BR tree, except that the Micrairoideae sequences are located disjoint in the tree (Fig. 9, Tab. 2).

MinCut supertree.— The MinCut supertree (Fig. 10, Tab. 2) located the outgroup sequences basal in the tree followed by the early grass lineages. Besides, only the Pooideae and Danthonioideae were recovered. The remaining tree presented a by and large a mixture of caterpillar-like and star-like areas mixing most of the subfamilies. This problem has been assigned to the fact that the MinCut approach resembles the Adams consensus (Mike Steel, pers. comm.) placing unclear taxa basal to the respective subtrees.

ModMinCut supertree.— The modifications suggested by Page (2002) in the ModMinCut supertree algorithm lead to a biologically much more reasonable tree (Fig. 11) compared to MinCut. All BEP subfamilies and the BEP clade itself were recovered and three PACCMAD subfamilies plus the Panicoideae *s.str.* group, the early

species *Eriachne*, *Isachne*, and *Micraira* by Sánchez-Ken et al. (2007) extending the former PACCAD to a PACCMAD clade. Bouchenak-Khelladi et al. (2008), however, could only support Micrairoideae excluding *Isachne*. The Micrairoid species had still been considered *incertae sedis* (*i.s.*, 'of uncertain placement') by the Grass Phylogeny Working Group (2001). As the GPWG data does not contain any *Isachne* sequences, we assume the Micrairoideae subfamily as correct.

Reconstructed Poaceae Trees

Table 2 summarizes the recovered grass subfamilies using the superalignment approach (Fig. 5), MRP with Baum/Ragan (Fig. 6) or Purvis coding (Fig. 7), MRF with Baum/Ragan (Fig. 8) or Purvis coding (Fig. 9), MinCut (Fig. 10), ModMinCut (Fig. 11), and SuperQP (Fig. 12).

We ignore the placements of species *incertae sedis* in the evaluation. If only Panicoideae *sensu stricto* (*s.str.*, that means, excluding *Danthoniopsis*) was recovered, we have marked Panicoideae with an open circle (○). The Centothecoid subfamily could not be resolved by any method except by using superalignment.

Superalignment tree.— The ML tree based on the superalignment shows the best result (Fig. 5) with respect to the Poaceae classification. The tree resolves both major clades and all ten subfamilies, including even the *Chloridoideae* (Tab. 2). Also the outgroups and early grasses were recovered.

MRP supertree using Baum/Ragan coding.— MRP-BR produced a good result resolving the outgroup, the early grasses, the major clades, the three BEP subfamilies and two of the PACCMAD subfamilies (Fig. 6, Tab. 2). Panicoideae and Centothecoideae form a subtree but only the Panicoideae *s.str.* could be resolved. If the

MinCut (Semple and Steel, 2000) and the ModMinCut supertree algorithm (Page, 2002). The reconstruction was performed by using Rod Page's SUPERTREE software included in the Rainbow package (Chen et al., 2004).

Medium Level Combination with SuperQP.— Superquartet puzzling was performed as described above. For the six genes ML quartet trees were computed with TREE-PUZZLE 6.0. The resulting log-likelihoods were combined to superquartets. Finally, the superquartet topologies and the overlap graph were fed to the TREE-PUZZLE 6.0 program to reconstruct the overall tree. 100,000 puzzling steps were performed to construct the M_{glob} consensus tree.

Poaceae Subfamily Structure

We compare the output of the methods with respect to the classifications suggested by the Grass Phylogeny Working Group (2001) also including recent improvements discussed by Sánchez-Ken et al. (2007) and Bouchenak-Khelladi et al. (2008). The grass dataset contains four outgroup taxa (*Flagellaria*, *Baloskion*, *Elegia*, *Joinvillea*) and the three early branching lineages Anomochloideae (*Anomochloa* and *Streptochaeta*), Pharoideae (*Pharus*), and Puelioideae (*Guaduella* and *Puelia*).

The main group of grasses consists of two clades where the clade names reflect the initials of the contained subfamilies: the PACCMAD clade comprising the subfamilies Panicoideae, Aristidoideae, Centothecoideae, Chloridoideae, Micrairoideae, Arundinoideae, and Danthonioideae and the BEP clade containing the subfamilies Bambusoideae, Erhartoideae, and Pooideae. The taxa in the Trees (Figs. 5-12) are denoted by their subfamily name and an index.

The grass subfamily Micrairoideae was suggested recently to comprise the

carboxylase/oxygenase large subunit (*rbcL*), RNA polymerase II β" subunit (*rpoC2*), and phytochrome B (*phyB*). The latest GPWG dataset comprising 66 taxa. The available sequences are listed in Table 1.

Evaluated Methods

Early Level Combination.— Superalignment analysis was performed by concatenating the gene alignments into one large superalignment. From the superalignment a tree was constructed with IQPNNI (Vinh and von Haeseler, 2004).

Late Level Combination.— For the late level methods, ML trees were reconstructed using IQPNNI (Vinh and von Haeseler, 2004) for each gene. The resulting gene trees were used as input for various late level methods.

MRP and MRF supertrees.— These methods use a binary matrix representation of the input trees which are constructed with the r8s program (Sanderson, 2003) using Baum/Ragan encoding (1992) as well as Purvis' scheme (1995). From the two matrix representations MRP trees (MRP-BR, MRP-Pu) and MRF trees (MRF-BR, MRF-Pu) were constructed. Most parsimonious trees were constructed using PAUP* (MRP) performing 100 repetition with TBR setting maxtrees to 10^4 (10^6 for Purvis coding due to the huge number of equally most parsimonious trees found). MRF trees were computed using the MRF heuristics implemented in HeuristicMFT2 (Eulenstein et al., 2004).

The equally best trees were summarized with CONSENSE from the PHYLIP package (Felsenstein, 1993) using 50% majority rule consensus (M_{50}) to get a better resolution than with strict consensus.

MinCut supertrees.— These were constructed from the input trees with the

of taxon-sets and taxa.

Global Relative Majority Consensus.—

To summarize the set of intermediate trees we construct a consensus tree. To that end we order all observed splits in descending order of occurence. Into the consensus tree we insert all splits with are compatible to all splits with higher or equal rank (no matter whether they were incorporated into the consensus tree or not).

In other words, for each split incorporated into the consensus tree there exists no incompatible split occuring more frequently in the set of intermediate trees. Every split, thus, has the the relative majority globally over all its incompatible splits. Hence, we will refer to it as the global relative majority consensus (M_{glob}).

Example: The Phylogeny of the Grasses

To evaluate SuperQP, we applied it to a multi-gene dataset of the well-studied group of the grasses (Poaceae). We use the reconstruction of their subfamilies and main clades (Grass Phylogeny Working Group, 2001; Sánchez-Ken et al., 2007; Bouchenak-Khelladi et al., 2008) as benchmark to compare the results of SuperQP with those of superalignment and supertree approaches, respectively.

The Dataset

The molecular sequences of Poaceae collected by the GPWG (Grass Phylogeny Working Group, 2001) comprises three nuclear loci, NADH dehydrogenase subunit F (*ndhF*), the internal transcribed spacer (ITS) of ribosomal DNA, and granule bound starch synthase I (GBSSI or *waxy*), as well as three chloroplast genes, ribulose 1,5-bisphosphate

overlap graph $G_{ovl} = (\mathcal{V}, \mathcal{E})$ with node set

$$\mathcal{V} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots \mathcal{S}_k\} \tag{6}$$

(7)

and edge set

$$\mathcal{E} = \{e_{(\mathcal{S}_i,\mathcal{S}_j)} | |\{\mathcal{S}_i \cap \mathcal{S}_j\}| \geq 3\} \quad 1 \leq i < j \leq k. \tag{8}$$

Furthermore we assign edge weights $w_{(\mathcal{S}_i,\mathcal{S}_j)} = |\mathcal{S}_i \cap \mathcal{S}_j|$. The overlap graph provides insight into several properties of the data.

We can construct an overall tree if and only if G_{ovl} is connected (Fig. 3c). If G_{ovl} is connected, then the overlap-guided puzzling step will lead to an overall tree.

Overlap-guided puzzling step.— Starting from a randomly picked taxon-set $\mathcal{S}' \in \{\mathcal{S}_1 \ldots \mathcal{S}_k\}$, we define a front set $\mathcal{F}$ containing all unused taxon-sets from the overlap graph connected by an edge to any taxon-set already used in building the tree (only $\mathcal{S}'$ at the beginning). The taxa in $\mathcal{S}'$ are inserted using the voting scheme gaining a start tree $T(\mathcal{S}')$. Then we randomly draw one taxon-set $\mathcal{S}''$ randomly from $\mathcal{F}$. We remove $\mathcal{S}''$ from $\mathcal{F}$ and add to $\mathcal{F}$ all (unused) taxon-sets connected to $\mathcal{S}''$, ~~but not yet in $\mathcal{F}$.~~ The unused taxa of $\mathcal{S}''$ are then added to the tree. Then we pick a new $\mathcal{S}''$ from $\mathcal{F}$ and repeat the procedure until all $X \in \mathcal{S}$ are placed in the tree. This taxon set selection procedure is similar to Prim's minimum spanning tree algorithm (Cormen et al., 2001).

This selection of taxa-sets guarantees that it will be always possible to construct an overall tree although not all quartets are represented ~~by trees.~~

As in QP many intermediate trees are constructed using different random orders

$$I(e) = \frac{\text{Bon}_e - \text{Pen}_e}{\text{Bon}_e + \text{Pen}_e + \text{Miss}_e} \tag{5}$$

The relative score reflects the number of quartets favoring e (Bon_e) minus those contradicting e (Pen_e) normalized by the the number of all quartets leading through e including unresolved and missing quartets. Using the bonus and penalty we solve the problem, that discarding all missing quartets can lead to edges or whole subtrees with low penalty scores, but not because they are proper edges to insert X but merely due to missing data. By using the relative score, we avoid possible influences caused by different numbers of quartets leading through an edge depending on its position within the tree.

Taxon X is inserted into the edge with maximal relative score. If the maximal relative score is attained by more than one edge, one of them is chosen randomly.

Again this procedure is repeated until we construct an intermediate tree containing all taxa. To avoid any bias the construction of intermediate trees is repeated many times and a majority rule consensus is computed (see below).

The overlap condition.— Contrary to the QP algorithm we cannot simply pick a random taxon X not already in the tree, because the set of superquartets is not necessarily equal to $\mathcal{Q}$. Thus, it may happen that X cannot be placed on the partially reconstructed tree because no superquartet trees with taxon-set $\{a, b, c, X\}$ is available. To avoid such degenerate cases we introduce the so-called *overlap condition.*

For two taxon-sets $\mathcal{S}_i$ and $\mathcal{S}_j$ their pairwise overlap $|\mathcal{S}_i \cap \mathcal{S}_j|$ must be at least 3. This is related to the concepts of tree-graphs (Sanderson et al., 1998) and groves (Sanderson et al., 2007).

To assess whether $\mathcal{S}_1, \ldots, \mathcal{S}_k$ are suitable to build ~~one~~ a tree, we construct the

Draws regarding equally supported quartet topologies (cf. Strimmer et al., 1997) or insertion edges are broken randomly.

QP requires that that quartet trees are available for all quartets in $\mathcal{Q}$. Due to missing data this is usually not the case when combining datasets of genes. Two problems arise: (a) how to deal with missing quartet trees in the voting scheme to insert new taxa into the tree and (b) how to choose the insertion order of taxa to guarantee available quartet tree information to guide the insertion process.

The SuperQP voting scheme.— Using only the number of contradicting quartet trees for an edge (as in QP) is not sufficient in the context of missing data.

To cope with missing quartet trees we use the following voting scheme. A quartet tree $ab|cX$ (a, b, c already in the tree; Fig. 4a) does provide two types of information: First, not to insert X into the path $b \leftrightarrow b$ and second, that X should be placed on the path $z \leftrightarrow c$ (Fig. 4a) because this placement preserves the quartet tree $ab|cX$.

Now we assign a penalty score of 1 for the edges on the paths $z \leftrightarrow a$, $z \leftrightarrow b$, and a bonus sore of 1 to the edges on $z \leftrightarrow c$. Note, that z is uniquely defined by the triplet induced by $\{a, b, c\}$. As in QP the procedure is repeated for all quartets $\{a, b, c, X\}$ when a superquartet tree is available. We compute the sum of penalty scores Pen_e and bonus scores Bon_e for each edge e. If we cannot decide which out of two quartet topologies is the best (Strimmer et al., 1997, cf.), we add penalty and bonus scores of $\frac{1}{2}$ for each of the two topologies, accordingly. If all three quartet topologies are plausible, we treat the corresponding quartet as missing quartet. For all quartets $\{a, b, c, X\}$ for which no bonus and penalty score were assigned since there was no phylogenetic information available, we introduce missing score Miss_e. For every such quartet, at each edge e on the subtree $\{a, b, c\}$ a missing score of 1 is added to Miss_e.

Then we compute for each edge e the relative score

denotes the summed log-likelihood over all genes that are sequenced for taxa a, b, c, and d. Similarly we define

$$\ell_{ac|bd} = \sum_{g=1}^{k} L_{ac|bd}^{(g)} \quad \text{and} \quad \ell_{ad|bc} = \sum_{g=1}^{k} L_{ab|bc}^{(g)} \tag{4}$$

If a quartet is not represented by an alignment we called it *missing quartet* and ℓ_τ or $\widehat{\ell_\tau}$ are not computed.

In the following we will use the log-likelihoods ℓ_τ (Eqs. 3-4) to determine those quartet topologies supported by the data. Moreover, we employ discrete posterior weights allowing also for partly resolved quartets (i.e. two trees are equally supported) and unresolved quartets (if all three trees are equally likely) (see Strimmer et al., 1997). These quartet topologies are called superquartet trees.

From (incomplete) Superquartets to Trees

Quartet puzzling (QP).— Superquartet puzzling is related to the quartet puzzling algorithm (Strimmer and von Haeseler, 1996) in that it uses topological information of superquartet trees to insert taxon after taxon to construct an overall tree. Hence, we will quickly recap the insertion step of the QP algorithm. To insert taxon X into a tree, QP inspects all quartets of the form $\{a, b, c, X\} \in \mathcal{Q}$ where taxa a, b, and c are already in the tree. Assume, that the topology $ab|cX$ is supported, then inserting X into the path connecting a and b ($a \leftrightarrow b$) would conflict with the quartet topology. Hence, a penalty of 1 is added to each edge on $a \leftrightarrow b$. This is repeated for all quartets $\{a, b, c, X\}$ and finally X is inserted into the edge with lowest penalty, i.e. the edge contradicted by the least quartets. This voting scheme is repeated until all taxa are added to the tree.

METHOD

Combining Quartets to Superquartets: Notation and Method

Let $\mathcal{S} = \{s_1, \ldots, s_n\}$ denote the total taxon-set. $\mathcal{S}_g \subseteq \mathcal{S}$ denotes the taxon-(sub)subset represented by sequences available for gene g. For each of the k genes a multiple sequence alignment of orthologous sequences with taxon-sets $\mathcal{S}_1, \ldots, \mathcal{S}_k$ is available. $\mathcal{Q}$ denotes the collection of all possible subsets of size four (quartets) derived from $\mathcal{S}$, while $\mathcal{Q}_g$ denotes quartets derived from taxon-set $\mathcal{S}_g$. To avoid degenerate cases, we will always require that $|\mathcal{S}_g| \geq 4$ and $\bigcup_{g=1}^{k} \mathcal{S}_g = \mathcal{S}$. For each quartet $\{a, b, c, d\}$ three resolved trees, namely $ab|cd$, $ac|bd$, $ad|bc$, are possible. a, b, c, and d represent taxa from $\mathcal{S}$ and '|' represents the internal edge of the four-taxon tree (Fig. 2a-c).

$L_{\max}(ab|cd)$ denotes the maximum log-likelihood value for the tree $ab|cd$ computed from the four sequences from taxa a, b, c, and d. With

$$L^{(g)}_{ab|cd} \equiv L_{\max}(ab|cd) \cdot \delta^{(g)}_{\{a,b,c,d\}} \tag{1}$$

we denote the maximum log-likelihood of $ab|cd$ for quartet $\{a, b, c, d\} \in \mathcal{Q}$ using gene g. The Kronecker δ in defined as

$$\delta^{g}_{\{a,b,c,d\}} = \begin{cases} 1 & \text{if } \{a, b, c, d\} \in \mathcal{Q}_g \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Thus, $L^{(g)}_{ab|cd} = 0$ if no sequence of gene g is available for ~~at least one taxon~~ a, b, c, ~~or~~ d. Finally,

$$\ell_{ab|cd} = \sum_{g=1}^{k} L^{(g)}_{ab|cd} \tag{3}$$

et al., 1981; Bryant and Steel, 1995; Semple and Steel, 2000; Page, 2002). For an overview we refer to Bininda-Emonds (2004a).

There has been a long debate whether to use supertree/consensus methods (e.g., de Queiroz et al., 1995; Miyamoto and Fitch, 1995; Bininda-Emonds et al., 2002; Bininda-Emonds, 2003b, 2004b; Bininda-Emonds et al., 2004) or superalignment approaches (e.g., Gatesy et al., 2002, 2004; Eernisse and Kluge, 1993; Kluge, 1989, 1998). Superalignment approaches have been criticized because they do not take into account different evolutionary models that may act on the different genes. On the other hand supertree approaches are criticized for the loss of information, since the underlying data usually is discarded prior to combining the input trees. Furthermore, supertree methods have been criticized for possible unwanted data duplication and weighting (Gatesy et al., 2002; Gatesy and Springer, 2004) especially if tree topologies have been collected from literature.

Here we suggest an approach to build a multi-gene species tree that combines the data at a medium level between the early-level (supermatrix) and late-level (supertree) methods. The so-called superquartet puzzling (SuperQP) approach aims to overcome disadvantages discussed for early and late-level combination approaches.

As an illustrative example, SuperQP is applied to the Poaceae data (Grass Phylogeny Working Group, 2001) and the resulting tree is compared to the trees derived from supertree and superalignment approaches.

Efficient sequencing techniques and genome projects bestow exponential growth upon sequence data in public primary databases. Notwithstanding this growth rate, the mutual coverage with respect to phylogenetically interesting taxa and the completeness of gene sequences available for each taxon is far from being satisfactory (Sanderson and Driskell, 2003; Bininda-Emonds, 2005). If one wants to analyze complete data, i.e., data comprising all genes for each examined taxon, this would result in small datasets with respect to number of taxa and/or sequences. To take advantage of the available data one needs methods which incorporate also patchy data, that means sets of genes with missing sequences for various taxa. Different motivations exist to combine multiple genes, proteins, etc. Some researchers try to increase the number of taxa to construct large scale trees (e.g., Bininda-Emonds et al., 2007; Angiosperm Phylogeny Group, 2003; Delsuc et al., 2005), while others aim at extending the data basis to draw better conclusions about the phylogenetic or systematic relationships of their family of interest (e.g., Meredith et al., 2008; Zgurski et al., 2008).

At present mainly two strategies that combine sets of genes to infer phylogenies are used. The first strategy combines the data very early (Fig. 1) by concatenating all gene alignments into one large superalignment. This superalignment is then used to reconstruct a tree. These approaches are also known as supermatrix (sensu Sanderson et al., 1998) or 'total evidence' (sensu Kluge, 1989) approaches. The second strategy combines the data late (Fig. 1). First trees are constructed for each gene separately. Then the resulting gene trees are combined into a supertree (sensu Gordon, 1986). Quite a number of supertree methods have been suggested, based for example on decompositions of the usually rooted trees into a binary matrix representation (Baum, 1992; Ragan, 1992; Purvis, 1995; Chen et al., 2003), quartets (Robinson-Rechavi and Graur, 2001; Piaggio-Talice et al., 2004), or rooted triples and common nestings (Aho

Abstract

We present an algorithm to reconstruct a species-phylogeny from a collection of sequence alignments (genes), where not necessarily each genes is complete with respect to the collection of species one is interested in. Researchers use incomplete multi-gene datasets extend the data underlying their studies to increase evidence and/or the enlarge the set of taxa to get more exhaustive insights into, e.g., the tree of life.

The algorithm proceeds as follows; for each gene the likelihood of the three trees for each possible quartet are computed. Subsequently the likelihoods of the genes are combined to obtain overall likelihoods for each quartet from the species-set. Finally, the quartet species trees are used to compute the global species tree taking missing data into account (superquartet puzzling, SuperQP).

Contrary to superalignment and supertree approaches we combine the data at a medium level (the quartets), thus, we can specifically include the evolutionary parameters for each gene. This medium level combination tries to reduce the loss of primary information provided by the genes.

We apply the SuperQP approach to reconstruct the taxonomic classification of the grasses and compare the results to the findings of other superalignment and supertree approaches.

[Multi-gene datasets; missing data; medium-level combination; phylogenetic tree; superquartet puzzling; Poaceae]

Superquartet Puzzling: Combined Phylogenetic Analysis Between Supertrees and Superalignment Approaches

Heiko A. Schmidt and Arndt von Haeseler

Address:

Center for Integrative Bioinformatics Vienna,

Max F. Perutz Laboratories,

University of Vienna,

Medical University of Vienna,

University of Veterinary Medicine Vienna,

Dr. Bohr-Gasse 9, A-1030 Vienna, Austria

Corresponding Author for proofs:

Heiko A. Schmidt

Center for Integrative Bioinformatics Vienna (CIBIV),

Max F. Perutz Laboratories,

Dr. Bohr-Gasse 9, A-1030 Vienna, Austria

Tel.: ++43 +664 / 6027724021

Fax.: ++43 +1 / 79044 - 4551

Email: heiko.schmidt@univie.ac.at