Syst. Biol. 62(3):456–466, 2013

DOI:10.1093/sysbio/syt011
Advance Access publication February 15, 2013

Resolving Ambiguity of Species Limits and Concatenation in Multilocus Sequence Data for the Construction of Phylogenetic Supermatrices

DOUGLAS CHESTERS[1,2,3] AND ALFRIED P. VOGLER[1,2,*]

[1]*Department of Entomology, Natural History Museum, London SW7 5BD, UK;* [2]*Division of Biology, Imperial College London, Silwood Park Campus, Ascot, Berkshire SL5 7PY, UK; and* [3]*Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China*

[*]*Correspondence to be sent to: Department of Entomology, Natural History Museum, Cromwell Road, London SW7 5BD, UK. E-mail: apv@nhm.ac.uk.*

Received 25 July 2012; reviews returned 1 November 2012; accepted 8 February 2013
Associate Editor: Karl Kjer

Abstract.—Public DNA databases are becoming too large and too complex for manual methods to generate phylogenetic supermatrices from multiple gene sequences. Delineating the terminals based on taxonomic labels is no longer practical because species identifications are frequently incomplete and gene trees are incongruent with Linnaean binomials, which results in uncertainty about how to combine species units among unlinked loci. We developed a procedure that minimizes the problem of forming multilocus species units in a large phylogenetic data set using algorithms from graph theory. An initial step established sequence clusters for each locus that broadly correspond to the species level. These clusters frequently include sequences labeled with various binomials and specimen identifiers that create multiple alternatives for concatenation. To choose among these possibilities, we minimize taxonomic conflict among the species units globally in the data set using a multipartite heuristic algorithm. The procedure was applied to all available GenBank data for Coleoptera (beetles) including > 10 500 taxon labels and > 23 500 sequences of 4 loci, which were grouped into 11 241 clusters or divergent singletons by the BlastClust software. Within each cluster, unidentified sequences could be assigned to a species name through the association with fully identified sequences, resulting in 510 new identifications (13.9% of total unidentified sequences) of which nearly half were "trans-locus" identifications by clustering of sequences at a secondary locus. The limits of DNA-based clusters were inconsistent with the Linnaean binomials for 1518 clusters (13.5%) that contained more than one binomial or split a single binomial among multiple clusters. By applying a scoring scheme for full and partial name matches in pairs of clusters, a maximum weight set of 7366 global species units was produced. Varying the match weights for partial matches had little effect on the number of units, although if partial matches were disallowed, the number increased greatly. Trees from the resulting supermatrices generally produced tree topologies in good agreement with the higher taxonomy of Coleoptera, with fewer terminals compared with trees generated according to standard filtering of sequences using species labels. The study illustrates a strategy for assembling the tree-of-life from an ever more complex primary database. [BlastClust; data mining; graph theory; incongruence; multipartite matching; species delimitation; supermatrix.]

Metadata from mining of public DNA databases are a rapidly growing resource for molecular phylogenetics (Driskell et al. 2004). Compilations of these data for construction of large phylogenetic trees and multiple gene sets are widely performed (Sanderson et al. 2008; Goloboff et al. 2009; Peters et al. 2011) and software for data manipulation is available (Jones et al. 2011). Whether produced based on existing gene annotations (Jones et al. 2011), or with similarity searches (Hunt and Vogler 2008; Sanderson et al. 2008), the initial products of these compilations are sets of presumed orthologous sequences partitioned by locus. Combined data analysis generally outperforms other methods of phylogenetic inference such as supertrees derived from individual gene fragments (McMahon and Sanderson 2006; de Queiroz and Gatesy 2007; Kupczok et al. 2010; Thomson and Shaffer 2010), and therefore sequences require concatenation across loci to generate a "supermatrix" or "superalignment." There has been great interest in the behavior of supermatrices, for example, regarding the density of taxon sampling, the selection of loci and the number of loci and characters required, the effects of missing data and the optimal way in which to reduce them (Philippe et al. 2004; Delsuc et al. 2005; Gatesy et al. 2007; Simon et al. 2009; Meusemann et al. 2010).

However, little attention has been paid to the key step of defining the terminals for such analysis. Specifically, it is not straightforward how to delimit the species entities that represent the terminals at each locus, and how to combine sequence data from various loci. In other words, although methods have been developed for specifying the columns in a 2-dimensional data matrix (representing orthologous sequences), the rows of the matrix (representing the terminals) have not been addressed. The problem of defining species units is complicated by the recent trend of depositing sequences without full species identification (Ryberg et al. 2008), either labeled with incomplete taxon names or approximate species identifiers ("sp.," "cf.," "aff.," "sp. near," etc.), or with various alphanumerical specimen tags referring to individuals, rather than a taxon. Inconsistent species identifications by various authors exacerbate the problem of name-based concatenation across loci. In addition, the phylogenetic history of various DNA markers may be incongruent, for example, due to hybridization (Funk and Omland 2003; Edwards 2009). This causes different taxonomic affiliations of gene copies depending on the locus under consideration and results in incongruent signal in a concatenated matrix.

This study sets out a procedure for generating a global (multilocus) species delineation matrix by optimizing the concatenation process. This required that multiple exemplars of a species, for example, from population

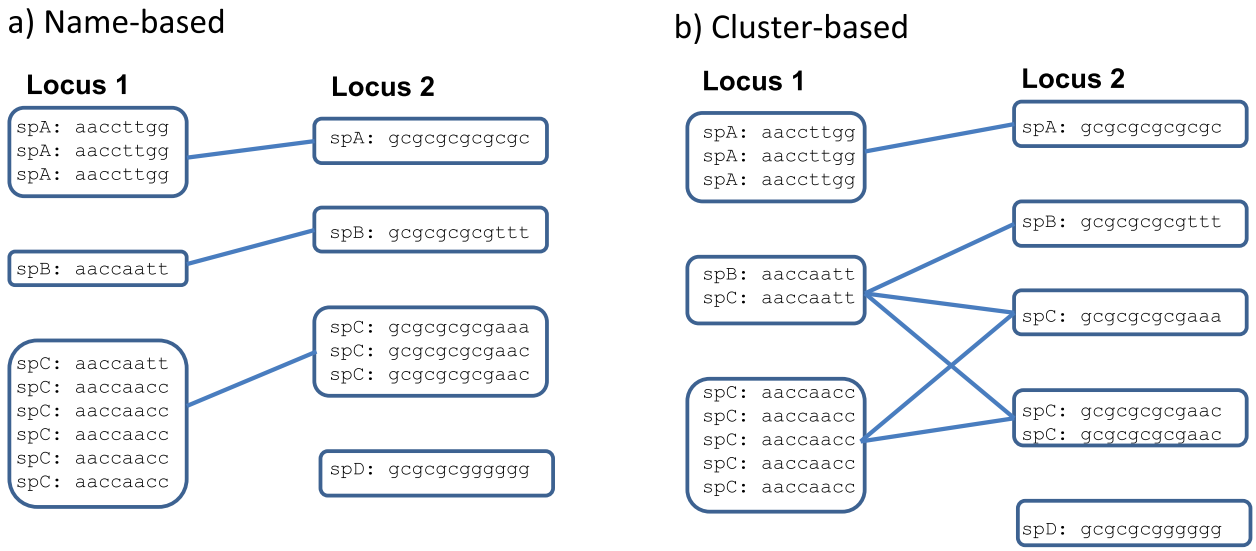

FIGURE 1. Building a supermatrix from species units (boxes). a) Name-based concatenation. A single exemplar sequence is selected to represent each species to be concatenated with a sequence labeled with the same name in a second locus. Missing data are inserted into the matrix where a species name is not represented in a locus (e.g., spD missing for Locus 1). b) Cluster-based concatenation. Clusters for each locus are produced based on sequence similarity, but may result in ambiguity of concatenation if the molecular clusters are incongruent with the species labels. This results in numerous combinations in which loci can be combined based on shared taxonomic labels (blue connecting lines).

analyses, are grouped to provide an estimate of species entities independently of taxon labels or particular identifications (Tautz et al. 2003; Brock et al. 2009; Lee et al. 2012). These sequence clusters established for each locus have to be combined into a supermatrix. This is straightforward if all sequences in the clusters are labeled with the same species name (Fig. 1a), in which case any sequence can be picked to represent a given locus in the supermatrix with equal justification. However, often the sequence variation is not congruent with the taxon assignment, that is, a name may be associated with multiple clusters, or conversely a single taxon label may be distributed over multiple groups. In these cases, combining loci becomes ambiguous (Fig. 1b). As we will show, empirical data sets are greatly affected by complex pattern of ambiguity, causing great difficulties to join sequences in matrix construction.

The proposed procedure concatenates the representatives for each locus in a way that minimizes conflict of the taxon names (including specimen identifiers) attached to each group. Because multiple labels may be present in a single sequence cluster and a single label may be distributed across multiple clusters, there are many different ways in which to link 2 loci (Fig. 2). The challenge of optimally linking these entities is taken as a specific case of the general problem of maximizing matches in graph theory (Cherkassky et al. 1996). Here, the label similarities among species units from multiple loci are represented as a multipartite graph, where nodes in a partite set correspond to the labels of clusters generated in the initial clustering step, and edges link nodes from adjacent partite sets where species names are shared. Decomposing the collection of edges such that nodes have no greater than a single edge to adjacent node sets is by "graph matching" (Fig. 2), in which the algorithm aims to find the greatest possible number of edges or the greatest sum of weights. Thus, the preferred global solution is a combinatorial optimization problem on multipartite graphs. By applying an explicit weighting scheme for name matches and partial matches, the linking may either be very strict, avoiding mismatches of any kind (beyond those already contained within a single-locus cluster), or may be more permissive by allowing mixed clusters with multiple taxon labels. The latter will increase the number of loci that participate in a concatenate and so reduce the number of terminals in the resulting supermatrix, hence increase the gene occupancy, at the cost of greater conflict of names within the concatenates.

We applied this procedure to a large multilocus data set of many thousand taxa for an entire insect order, the Coleoptera (beetles). Starting from DNA-based clusters that were created with the rapid BlastClust procedure for establishing the primary entities, we first assessed which clusters are inconsistent with Linnaean names and therefore constitute difficulties for concatenation. The labels attached to these clusters were used to generate a maximum matching set with the greatest number of edges. This set links various loci in accordance with the taxonomy to the greatest degree possible given the inconsistencies of sequence labels in the clusters. We also compared the outcome of the analysis with a conventional name-based concatenation, to test the properties of the data matrices in regard to the number of terminals, the proportion of missing data, and potential improvement in tree topology.

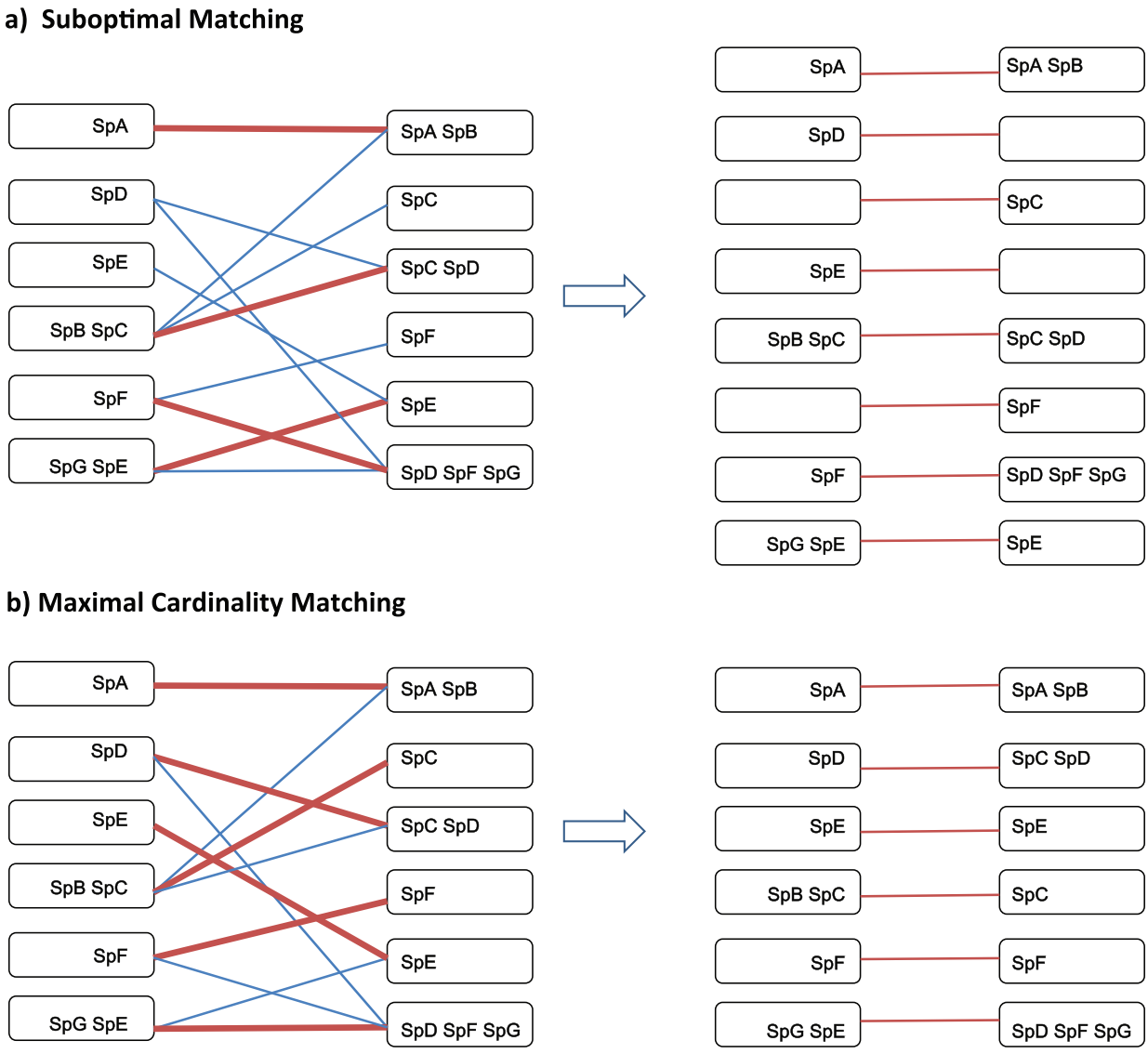

FIGURE 2. Species-level clusters from multiple loci represented as a multipartite graph. Species names associated with hypothetical sequence clusters at 2 loci are shown. Edges link clusters of adjacent partite where taxonomic species are shared (thin and fat lines). a) Graph is decomposed to permit only single links (fat lines), but resulting in a suboptimal matching. No further edges may be matched due to taxonomic dissimilarity of clusters. Matching results in a matrix of 8 global species units, with a cell density of 0.75. b) Maximum matching (fat lines), forming a matrix of 6 global species units and cell density of 1.0. All cases shown here are for the bipartite matching of 2 loci only, which is repeated sequentially for the multilocus (multipartite) data set.

MATERIALS AND METHODS

Data Mining and Species Clustering at Each Locus

Sequences were obtained from the NCBI database by downloading the invertebrate release (DNA) flatfiles from ftp://ftp.ncbi.nih.gov/genbank/. In addition, we obtained the NCBI taxonomy database from ftp://ftp.ncbi.nih.gov/pub/taxonomy/, from which we built a unique name-based string for each taxon (the "taxon identifier" or "taxon ID") corresponding to the taxonomic hierarchy (Hunt and Vogler 2008). Taxon IDs were generated for every node descended from the "Coleoptera" node (NCBI taxon number 7041), using the Linnaean binomials and 7 hierarchical levels of the Linnaean taxonomic system up to the species level currently applied in this database. The taxon IDs were assigned to DNA sequences from the flatfiles by matching the NCBI taxon numbers. The formatdb NCBI program (ftp://ftp.ncbi.nih.gov/toolbox/FAQ.html) was used to generate a Blast searchable database which was screened for partial gene sequences coding for Cytochrome Oxidase subunit I (COI, 3′-portion), 16S rRNA, 18S rRNA, and 28S rRNA, using multiple queries from a phylogenetically diverse range of Coleoptera for each locus. Sequences showing similarity above the *e*-value cutoff of $1e-5$ were obtained using a BioPerl

(Stajich et al. 2002) wrapper script, in which the Fastacmd tool (ftp://ftp.ncbi.nih.gov/toolbox/FAQ.html) was used for trimming of sequences to the extent of the query and obtaining the reverse complement sequence where necessary.

The first step of the matrix building procedure was the clustering of the DNA sequences into species-level groups. We utilized a rapid Blast-based clustering method implemented in NCBI's BlastClust program to define operational species-level units, as applied previously to microbiological species discovery (Lee et al. 2012). BlastClust performs all-against-all pairwise alignment, followed by single linkage clustering according to user-specified similarity thresholds. Rates of sequence variation are highly locus specific (e.g., Rokas et al. 2002; Danforth et al. 2005), hence the degree of sequence variation within a species is likely to differ across the loci. We therefore assessed a range of single linkage cutoffs for each of the 4 target loci for the highest congruence with the Linnaean species names. For each cutoff, we summed the number of correctly delimited species, that is, cases in which multiple sequences labeled as a particular taxonomic name are grouped into a single cluster that are free of sequences identified as belonging to other species. Unidentified sequences (lacking assignment to Linnaean binomials in GenBank), and sequences labeled with a given taxon name that occurred only once, were excluded from this assessment.

Multipartite Concatenation from Ambiguously Assigned Clusters

The set of operational units, defined as BlastClust clusters for a locus k_i, can be represented as vertices V_i, in an undirected graph G. Relationships between vertices are indicated by edges, where in the current instance, relationships are established based on the sharing of taxonomic names between units. Since units are grouped into classes (loci), with the aim of creating a serial set of matches (a concatenate), the graph is k-partite, where k equals the number of loci. In a k-partite graph, edges exist between vertex classes, but not within a vertex class. Where 2 vertex classes are present, the graph is bipartite. Figure 2 gives a toy example, with 2 partites (loci), each containing 6 vertices (species units), which are composed of one or more species labels (SpA to SpG). Edges represent all taxonomic matches between operational units at different loci. For example, the lower left unit has 2 edges to the adjacent locus, one indicating the shared presence of SpE and one for the shared SpG. Delineation of the graph into global species units requires decomposition so that no vertex has more than a single edge. A suboptimal decomposition is shown in Figure 2a, where 4 units in the first locus have been matched to 4 units in the second locus, resulting in 8 global species units. There are no species matches between partites in the remaining 4 units. This results in missing cells in the resulting global species units (a matrix density of 75%), corresponding to missing character states in the supermatrix. In contrast, all units are linked in the maximal cardinality matching shown in Figure 2b, thus forming 6 global species units with a matrix density of 100%. Finding a maximal matching in general k-partite graphs is NP-hard (Garey and Johnson 1983), but a simple heuristic is to reduce the problem to a series of bipartite matchings (Bandelt et al. 1994), as maximal matching of a bipartite graph is solvable in polynomial time (Papadimitriou and Steiglitz 1982). Additionally, bipartite graphs are well studied, with several algorithmic solutions. Here, we implement a single hub heuristic (Bandelt et al. 1994), as follows: for k loci, L1, L2, L3 … Lk, compute edge weights between L1 and L2, and find maximal matching. The resulting linked and unlinked operational units for L1+L2 become the vertices for a second round of graph matching, whereby edge weights are calculated between L1+L2 and L3, to determine the maximal bipartite matching in this second round. Rounds continue from $n=3\ldots k-1$, with bipartite matching between L1+L2…Ln and Ln+1. Maximal weight matches are calculated according to Galil (1986), as implemented in the Graph::Matching Perl module written by Joris van Rantwijk and freely available at http://jorisvr.nl/maximummatching.html, with the single hub multipartite heuristic implemented in Perl (see Supplementary Material).

Weights were applied to the name matchings based on unique taxon IDs (not individuals) matched between units in 2 loci. As the default weights, we assigned an arbitrary score of +2 if all members in an operational unit (sequence cluster) obtained in one locus are the same taxonomic species as all members in a unit from another locus (full matches). This score is increased by +1 for a match of a particular specimen that links units in different loci via an alphanumerical specimen label attached to the sequence. Specimen labels were automatically identified by pattern matching in the species name field, with strings containing only the letters a–z and a single internal whitespace assumed to be Linnaean, and specimen label otherwise. A score of +1 is assigned to "partial" name matches, if a subset of the sequences in a unit has at least one match to a species name in a unit obtained for a second locus.

The global (multilocus) species units were labeled according to the taxon ID most widely represented across loci. This strategy would avoid that, for example, a single mislabeled sequence would be used to label the entire concatenate. The following algorithm was used for assigning a taxon ID to a specific global species unit: for each locus, obtain all unique taxon IDs, then for each taxon ID, count the number of loci containing it. Where sequences in a specific matching were assigned multiple Linnaean binomials, the one with widest representation across loci was selected, or where binomials were absent, the most frequently non-binomial label was used. In addition to assigning a species label to the concatenate, this procedure was also used to choose the DNA sequence that represents a locus in the concatenate. Where multiple sequences for the preferred binomial were available, the most complete sequence was used

or a sequence was chosen at random if sequences were of equal length. A file with the matched clusters across the 4 loci and a key to the taxon identifiers are available as Supplementary Material (Dryad Digital Repository; doi:10.5061/dryad.qp367).

Building Phylogenetic Trees from Concatenates

The above process of selecting taxon labels and concatenating the labels across loci was followed to build data matrices accordingly. In an initial step, all available sequences from each locus were aligned using BlastAlign (Belshaw and Katzourakis 2005), a program based on the Blast algorithm successfully used to align large numbers of rRNA sequences in Coleoptera (Hunt et al. 2007; Hunt and Vogler 2008). Alignments were produced for each of the 4 loci (COI, 16S, 18S, and 28S) and representatives for each operational unit were selected. (Note that the above step of clustering did not require multiple sequence alignment because the BlastClust algorithm produces clusters according to pairwise similarity.)

The quality of data sets generated under different edge weighting regimes was assessed on the phylogenetic trees inferred from the concatenated supermatrices. Direct comparisons of trees under likelihood require that the supermatrices obtained under the various weighting regimes have the same number and composition of concatenates (=terminals in the phylogenetic analysis), whereas the heuristic matching of loci under different weights results in matrices of variously formed concatenates. We therefore searched the supermatrices for concatenates which are identical between regimes. This set of equivalent concatenates were retained both in the trees and supermatrices, while all others were pruned (see fig. 1 of Poe (1998) for pruning method). Hence, the tree topology was assessed only for a core set of concatenates recovered in all weighting schemes, whose relationships change due to differences in the variable concatenates (whose relationships were not assessed). Tree searches were performed on each of the resulting supermatrices under ML and the GTRCAT model with RAxML v. 7.2.8 (Stamatakis 2006), and with the NJ method implemented in Paup*4b (Swofford 2002), using ML distances and gamma distributed rates. To test for significance in the likelihood differences across weighting regimes, the branch lengths of the pruned topology were optimized first, and then the likelihood was calculated for each site. Site likelihoods were then bootstrapped using Consel (Shimodaira and Hasegawa 2001) to calculate approximately unbiased (AU) test statistics. The taxonomic retention index (tRI; Hunt and Vogler 2008) was used to assess the fit of a tree to the Linnaean taxonomy. The tRI was obtained by turning the taxon IDs created from the NCBI taxonomy database (see above) for each terminal into a set of binary pseudocharacters for various levels of the taxonomic hierarchy. State changes in each of these characters were scored for the trees using Paup to calculate the RI for each character, and the ensemble tRI for all characters in the matrix was based on taxonomic state changes of 1376 taxonomic groups (898 genera, 161 tribes, 175 subfamilies, 120 families, and 22 superfamilies).

Results

Clustering GenBank Entries and Grouping of Unidentified Sequences

The database contained 23 555 sequences and 10 503 taxon labels, including 7712 Linnaean binomials and 2791 alphanumerical name codes (referred to as "unidentified" in the following). The latter also contained 180 partial identifications ("sp.," "cf.," "nr.," or "aff."). It is not known how many species these unidentified sequences equate to, but an estimate was made for each locus based on the level of intraspecific versus interspecific sequence divergence in a clustering with the BlastClust algorithm. Using all sequences with Linnaean taxon labels present at least twice in the database, similarity cutoffs were varied for the range of 90–100% in steps of 0.25%, and for each value, the proportion of correctly recognized species was determined (Fig. 3). Several hundred of these clusters matched the Linnaean names in the case of the mitochondrial markers and just over one hundred for the nuclear markers (Table 1). The slowly evolving loci (18S and 28S) returned a peak of correctly defined clusters at very high sequence similarity (linkage cutoff 99.75%), that is, intraspecific variation was very low, while the peak for the faster evolving COI and 16S was around 96 and 98.5%, respectively, with a very broad maximum. However, even under these optimal cutoff values, congruence of the clusters with the established taxonomy was low, with at most 52% (COI), 46% (16S), 37% (18S), and 69% (28S) of clusters unambiguously associated with a single Linnaean binomial (Table 1).

We then applied these optimal values to the total database, which returned 4760, 3260, 2092, and 1129 clusters for COI, 16S, 18S, and 28S (total 11 241 clusters; Table 2), including those sequences that were not grouped with any others (singletons). These clusters generated under the optimal cutoff were considered as operational units to represent species-level entities. These entities also contained many unidentified sequences. For example, in the COI locus, 1282 unidentified sequences were members of 1023 units, which included 935 units (including singletons) with all members unidentified, and 88 "mixed" units of unidentified and identified sequences (Table 2). Association with these units resulted in unambiguous species identification where the identified sequences in the cluster belong to a single species. For COI, 119 unidentified sequences were assigned species names in this way. The proportions were similar for other loci, with a total of 295 sequences newly identified in this manner (Table 2). This species assignment was improved by assignment to a name at a secondary locus if sequences for a given unidentified specimen were part of

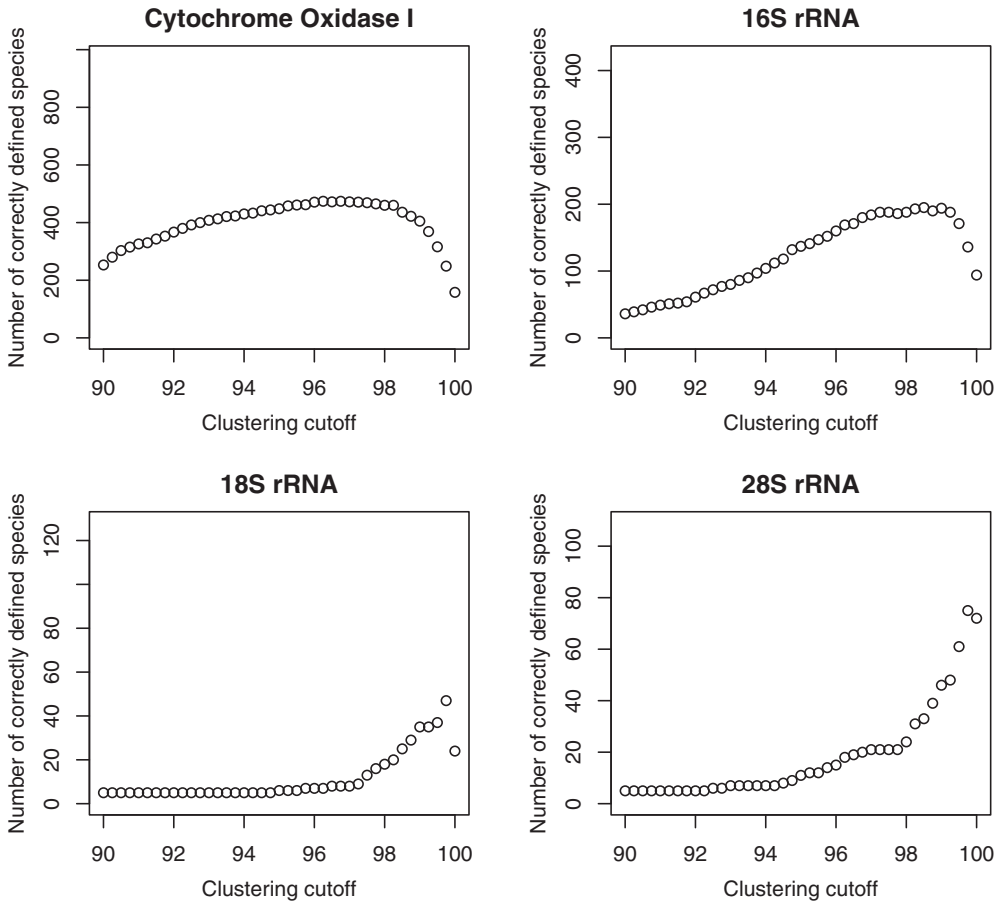

FIGURE 3. Blastclust cutoff values plotted against the number of correctly delimited species, that is, the proportion of clusters that include all sequences associated with a particular Linnaean name. The results are shown for 4 loci and all Coleoptera sequences with multiple entries per species name. The panels are scaled to the maximum possible value of correctly defined species for each gene.

TABLE 1. Optimal BlastClust cutoffs for the 4 genetic markers and number of species correctly delimited

Locus	COI	16S	18S	28S
Optimal BlastClust cutoff (%)	96.75	98.5	99.75	99.75
Number of species tested	474	420	128	109
Clusters congruent at optimal cutoff	249	195	47	75

a mixed cluster at another locus (Fig. 4). This trans-locus assignment increased the number of newly identified sequences by a further 215, for a total of 13.9% of the 3658 unidentified sequences in the data set (Table 2).

Incongruence of Clustering across Loci

Given the incongruence between sequence clusters and taxonomic species, we tested whether some of these novel groupings were corroborated by independent genetic loci. This test was performed on a subset of clusters that contained sequences labeled with 2 or more different species labels exclusively in that cluster, where these names were also present in other loci. These labels at a secondary locus may be consistent

TABLE 2. Species assignment to unidentified sequences via clustering at optimal clustering cutoff

Locus	COI	16S	18S	28S	Total
Clustering					
Sequences	13 557	5338	2486	2174	23 555
Clusters (including singletons)	4760	3260	2092	1129	11 241
Clusters all members identified	3737	2517	1366	702	8322
Clusters all members unidentified	935	688	674	399	2696
Clusters mixed identified and unidentified	88	55	52	28	223
Clusters with multiple binomials	243	178	95	57	573
Name assignments					
Total binomials	3577	2695	1482	824	8578
Total unidentified sequences	1282	1023	800	553	3658
Non-assignable sequences	1188	987	727	499	3401
Sequences assigned to binomials	119	82	46	48	295
Trans-locus assignments	70	89	16	40	215
Concatenation					
Unambiguous by string matches	4090	2783	1871	979	9723
Names dispersed among clusters	670	477	221	150	1518

Notes: The "clusters" refer to all separate entities produced in a BlastClust analysis at the optimal value (Table 1), including a large number of sequences not grouped with any others (singletons).

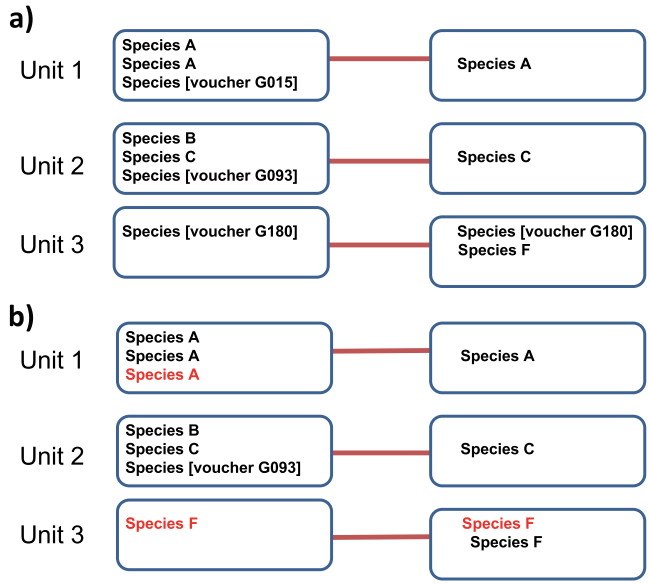

FIGURE 4. Assignment of species names to unidentified sequences. The figure shows a hypothetical example with 3 species units of 2 loci, containing both identified and unidentified members (labeled with a voucher numbers). Sequence labels as initially obtained are given in a), and name assignments (red letters) shown in b). Species names are assigned to the unidentified sequence by membership in a sequence cluster of identified sequences (Unit 1 and Unit 3), or through trans-locus assignment of a given specimen at a secondary locus (Unit 3). Where multiple named species are present in a cluster, a name cannot be unambiguously assigned to the unidentified sequence (Unit 2).

TABLE 3. Congruence in sequence clustering between loci

	Total		Identified		Unidentified		Mixed	
Locus	Cons.	Incons.	Cons.	Incons.	Cons.	Incons.	Cons.	Incons.
COI	271	133	56	22	186	96	29	15
16S	344	190	62	18	262	160	20	12
18S	126	125	11	15	106	93	9	17
28S	146	118	25	16	112	86	9	16

Notes: The table presents the results from tests of congruence in the associations of multiple names and/or alphanumerical codes across loci. Tests were performed on a subset of clusters that contained sequences labeled with 2 different taxon IDs, where these names were not found separated in clusters elsewhere. These "uniquely clustered" sequences are then used to assess each member for their status as being uniquely clustered at other loci (consistent, cons.), or as members of separate (=singleton sequence) or non-unique clusters at other loci (inconsistent, incons.).

with the first, that is, grouping the same labels into a single cluster, or they may be inconsistent, being distributed among several clusters. All members of these mixed clusters were then assessed for "consistent" or "inconsistent" trans-locus clustering. Likewise, we assessed the placement of unidentified sequences (alphanumerical specimen tags) and their linkage with clusters labeled with more than one binomial. The number of trans-locus inconsistencies of clusters affected between half and one-third of all sequences representing taxa with membership in multiple clusters (Table 3). The proportion was particularly high in the slowly varying 18S gene, compared with the faster evolving mitochondrial genes. For example, 125 of the 251 (49.8%) 18S clusters grouped species that were ungrouped at other loci. This was also evident from the (much larger) class of unidentified sequences, where trans-locus clustering was also inconsistent (Table 3).

Search for Optimal Concatenates

The most straightforward category for concatenation across loci were "singleton" sequences that did not group with any others and therefore were unique in representing this binomial. Although a large proportion of entities in a given locus were singletons (e.g. 3655 sequences in COI, against 1105 true clusters composed of more than one sequence), a binomial that is a singleton in one locus was frequently linked with a cluster in another locus. Only 1104 of minimally 7366 terminals (see below) in the final matrix were "singleton concatenates" composed of singletons in each locus available for the taxon. The remainder included many clusters with multiple taxon labels that were dispersed among several clusters in at least one locus. This problem affected 47.5% of all binomials represented by more than one sequence in COI (225 of 474 binomials), 53.6% for 16S, 63.3% for 18S, and 31.2% for 28S. Conversely, a cluster may contain multiple binomials that were dispersed across multiple clusters in at least one other locus. In total, there were 1518 clusters affected by these inconsistencies, corresponding to 13.5% of altogether 11 241 clusters. This leaves 9723 clusters that could be concatenated unequivocally using basic species name string matching (see Table 2 for breakdown by locus).

Varying Edge Weighting Regimes according to Matches in Linnaean Taxonomy

Applying the sequential bipartite matching algorithm to the 11 241 operational units from the 4 loci, maximal matchings were generated under a series of edge weighting regimes. For the 2:1:1 weighting regime (for full, partial, and specimen match; see "Material and Methods" section), the procedure returned 7366 terminals, and this value was essentially unchanged under different weights for partial matches (Table 4). Likewise, the number of chimerical concatenates was similar across weighting regimes, as was the proportion of missing data (62%). However, if the "partial" match score was set to 0 (weighting scheme 2:0:0), that is, the algorithm does not consider any solution that implements an edge between units with multiple names, the number of terminals increased, and consequently the resulting data matrix showed an increase in the proportion of missing data (64.8%; Table 4, final column). This increase in the number of terminals affected in particular the mixed-name units, which increased to 565 over 520 units obtained with the 2:1:1 concatenation scheme. When partial matching is permitted, these

TABLE 4. Comparing the impact of different edge weighting regimes on the resulting matrices and trees

Match	2	2	2	2	2
Partial match	2	1	2	1	0
Specimen match	0	0	1	1	0
Supermatrices					
Number of Concatenates	7367	7367	7366	7366	7955
Missing cells (proportion)	0.620	0.620	0.620	0.620	0.648
Chimerical concatenates	520	520	518	518	565
Tree statistics, ML					
Ensemble tRI	0.904	0.908	0.904	0.907	0.907
−logL	1 216 096	1 216 324	1 216 531	1 216 155	1 215 679
AU statistic	0.096	0.023	0.008	0.102	0.957
Tree statistics, NJ					
Ensemble tRI	0.812	0.811	0.813	0.813	0.811
−logL	1 321 219	1 321 182	1 320 998	1 320 753	1 322 788
AU statistic	0.112	0.084	0.242	0.896	<0.001

Notes: The weight for 2 species clusters paired between partites was varied according to taxonomic match, as given at the top. The composition of the resulting supermatrices is given in the upper rows, whereas the lower panel provides basic tree statistics obtained from the ML and NJ trees. Indicators of tree quality are the tRI, and the AU statistic for comparing likelihoods.

mixed-name clusters are mostly concatenated with mixed-name clusters in other loci, resulting in overall fewer terminals.

The concatenates were used for phylogenetic tree searches by retrieving the corresponding sequence information, retaining a single DNA sequence for each cluster for each locus, and concatenating these across loci. For each tree, we recorded the fit to the taxonomy (using the tRI, calculated from 1376 taxonomic groups of 5 ranks) and the likelihood of trees, after pruning all concatenates that are not universal to all weighting schemes to ensure comparability (see "Materials and Methods" section). The correspondence between lnL and tRI was low across weighing regimes, for the ML trees. For example, the tree attaining the highest tRI (2:1:0) was rejected ($AU < 0.05$) according to bootstrap analysis of site likelihoods. The quality of NJ trees was much lower than those obtained under ML, as indicated by the 2 measures, although a higher correspondence between tRI and lnL is observed, with the weighting regime of 2:1:1 ranking as the most likely, and gaining the (equal) highest tRI, and the regime 2:0:0 rejected according to lnL and (equal) lowest tRI (Table 4).

Comparing Supermatrices from Sequence-based and Name-based Data Sets

We also generated a concatenated supermatrix of the 4 loci based on Linnaean taxon labels (species names and unidentified specimen labels). The original sequences were filtered to leave a single sequence per species label, using the most complete sequence or, where sequences were of the same length, randomly choosing one individual. This "name-based" supermatrix included

TABLE 5. Number of matched species units with data present/absent for the 4 target loci, separate for the name-based and cluster-based supermatrices

Number of Loci	COI	16S	18S	28S	Cluster based	Name based	% Reduction
4 Loci	✓	✓	✓	✓	278	404	31.2
3 Loci	✓	✓	✓		233	241	3.3
	✓	✓		✓	327	363	9.9
	✓		✓	✓	31	40	22.5
		✓	✓	✓	44	57	22.8
2 Loci	✓	✓			1085	1360	20.2
		✓	✓		132	112	−17.8
			✓	✓	158	177	10.7
	✓		✓		222	197	−12.7
		✓		✓	64	89	28.1
	✓			✓	82	91	9.9
1 Locus	✓				2477	2174	−13.9
		✓			1097	1092	−0.5
			✓		993	1053	5.7
				✓	144	155	7.1

Note: The last column gives the percentage by which the number of cluster-based concatenates is lower than name-based concatenates (negative values if the number of cluster-based concatenates is greater).

7605 terminals with greater intra-genus representation compared with the cluster-based matrix, including a few hyperdiverse genera with >200 differently named terminals possibly derived from population studies. The proportion of missing data at 59.7% was lower than the cluster-based analysis (62.0%). The sequence-based concatenation resulted in a general reduction in concatenates, and in particular in the multilocus concatenates (Table 5). Contrary to this trend, the number of COI clusters was greatly increased in the sequence-based clusters, suggesting that this fastest diverging gene frequently split Linnaean entities, unlike the other loci. This increase was not observed in the multilocus concatenates of the COI gene, which suggests that the concatenation step reduces the number of entities, that is, clusters that split Linnaean species are concatenated based on alternative names in mixed clusters present at other loci.

For comparisons of the resulting tree topologies, we pruned all terminals that were not identical in both trees, leaving a core set of 5485 terminals. As expected for a matrix of greater size, the search time differed significantly for the taxon-based matrix compared with the sequence-based matrix. Based on the 1000 bootstrap replicates, a mean search time of 16.4 ± 1.1 h was required for the name filtered, against 12.8 ± 1.0 h for the cluster-based data matrix ($P = 0.012$, $W = 758\,579$, unpaired Wilcoxon rank sum test). After extensive searches, the tRI was 0.890 and 0.889 for the sequence-based and name-based data sets, respectively. To obtain an indication of significance for these values, the ensemble tRI was repeated for each of 1000 bootstrap trees for both data sets, resulting in a tRI of 0.8634 ± 0.0002 for the sequence-based versus 0.8600 ± 0.0002 for the name-based analysis, which was a highly significant difference ($P \leqslant 0.001$, $W = 262741$, unpaired Wilcoxon rank sum test). The tree from sequence-based

concatenation showed a small but significantly increased bootstrap support (0.604 ± 0.014, vs. 0.602 ± 0.014; $P = 0.019$, $V = 762\,162$, paired Wilcoxon signed rank test), whereby searches were performed only on a set of shared sequences to find equivalent nodes. Finally, the likelihoods of the 2 topologies were assessed after a round of thorough likelihood optimization. While not significant, the tree from the cluster-based matrix ranked as showing higher likelihood (AU = 0.859, PP = 1.0, SH = 0.864).

Discussion

It is widely recognized that gene annotations are insufficient and data mining therefore requires careful partitioning of orthologous sequences (Sanderson et al. 2008; Smith et al. 2009; Peters et al. 2011), as the first step in building a supermatrix. However, the literature has not addressed the allied problem of defining the terminals. Instead, virtually all recent studies define the species axis of the matrix by using the Linnaean names, which seems no longer adequate given the increasing proportion of un-named sequences and inconsistent identifications. The magnitude of the problem is apparent from the Coleoptera database that contains nearly 1000 "unidentified" species-level clusters (including singleton sequences) without assignment of Linnaean names and >1200 unidentified sequences for the COI gene alone. Altogether some 15.5% out of 23 555 sequences are "unidentified" (Table 2). More importantly, the names in databases do not coincide with species-level clusters obtained from the sequences themselves. Blast clustering produced species-level entities containing multiple names in some 23% of clusters with >1 sequence, even after careful selection of the best performing cutoff values. In as many as 50% of the binomials represented by >1 sequence, a given taxon ID was split among different clusters, that is, the annotations contradict the sequence-based entities, and to make matters worse, these splits frequently are "inconsistent" between loci (Table 3). Choosing representatives for each cluster and deciding what annotation to use in a supermatrix therefore become very complicated. We here developed a procedure for consolidating these single locus species clusters into a supermatrix. The approach maximizes the number of matches of taxon IDs among loci globally over the entire database, using established procedures from graph theory. The key feature of the method is that it matches each set of names (each cluster) with no more than one other set, until no more matching sets are available, to produce the maximum matching set. This process of bipartite matching is iterated for multiple loci to provide a heuristic solution for the NP-hard problem of multipartite matching.

The procedure is flexible in that it can incorporate a weighting function to reflect the greater certainty about the link of some vertices than others. This is particularly relevant if clusters can be linked across loci via sequences obtained from a single individual; in these cases, the link is unambiguous, and this gives confidence that these clusters should be concatenated beyond the taxon ID. We implemented this by adding extra weight to the matching function, which was set arbitrarily to add 50% greater weight above that for a "specimen match" (=identical taxon ID in clusters in 2 loci). Conversely, the certainty of a match may be reduced if certain taxon IDs differ in one or both of the clusters (=a "partial match"), which can be given lower weight than the match of all taxon IDs (=a "full match"). It may also be desirable to avoid partial matches altogether, which we implemented with the "partial match = 0" setting. In the Coleoptera data, only this latter approach had a substantial effect on the supermatrix, showing the greater number of terminals and more missing data expected if partial matches are disallowed. Other weighting schemes are conceivable. For example, the current approach does not take into account the number of sequences in each locus to be used in the weighting, which may lead to an over-proportional impact of a small number of aberrant sequences. This could be addressed by different weighting schemes that take into account the number of sequences supporting any one taxon ID. Likewise, additional information associated to particular sequences could be considered, such as geographic provenance of sequences, to preferentially match up clusters based on sequences from the same collecting locality or habitat that may be indicative of species membership. The "specimen match" could also be implemented by using the specimen voucher information if available on GenBank, instead of using alphanumerical name codes to insure that unique specimens are used, rather than unidentified entities. It also remains to be seen to what degree other parts of GenBank are affected by incongruence in species-level clusters. Currently, the Coleoptera database is well curated and consists of a majority of sequences assigned to Linnaean binomials, for example, 3577 binomials in the database for 4760 clusters in COI (75.1%) and a similar proportion of binomials to clusters in the other loci (Table 2). However, the proportion of unidentified sequences is likely to increase steeply in the future, for example, through the rapidly growing DNA barcodes (COI sequences) labeled by a "barcode index number" (Ratnasingham and Hebert 2007), an arbitrary alphanumerical code referring to DNA-based clusters similar to the clusters created here, and the adoption of high-throughput sequencing technologies, which places greater demand on taxonomic identifications via clustering.

Perhaps the greatest effect on the composition of the supermatrix is exerted by the clustering step preceding the concatenation. We used BlastClust as a pragmatic approach based on similarity cutoffs in a 2-step procedure whereby we first established the cutoff level that is most appropriate to the specific database by comparing clusters and Linnaean names to mirror traditional species circumscriptions, and then applied the preferred value to the wider database including

"unidentified" or partially identified sequences. The approach is simplistic, by using a similarity criterion and universal cutoff values for species delimitation, and assuming correct taxonomic identifications and taxon concepts in GenBank submissions. Database entries are notorious for a low fit of sequence clusters and Linnaean names (Meier et al. 2006), unlike studies based on dedicated sequencing efforts, which estimated this discrepancy to affect at most 5% of all clusters (Hajibabaei et al. 2006). This suggests inconsistent naming as well as a focus on problematic groups in the primary studies from which the database is built. The discrepancies also differ among loci. For example, in the slowly evolving 18S and 28S loci, we found that multiple binomials were commonly collapsed into a single entity, indicating over-clustering that was not observed in fast-changing mitochondrial loci (Table 2). The specifics of the clustering procedure may have contributed to the high incongruence with the Linnaean binomials, while other clustering procedures have recently been found to perform better (Lee et al. 2012), and tree-based procedures (Pons et al. 2006) may be preferable on theoretical grounds.

Once loci are partitioned and species clusters are delineated, the matrix building step has to draw on the units as a set of fixed entities that are to be linked in the most appropriate way. The linking parameters affect the relative proportion of missing data and internally conflicting terminals, which then affect tree inferences and nodal support. However, the preferred settings do not necessarily produce optimal conditions for tree inference. We compared tree topologies obtained under different parameters, by assessing only those entities that were common to the matrices from various weighting schemes, and pruning all other terminals. For the ML searches, the best likelihood values were with the 2:0:0 scheme, that is, not permitting partial matches (AU = 0.95), while in the NJ searches, the best values were for the 2:1:1 scheme (AU = 0.23). Yet, the tRI values did not differ greatly among weighting schemes. The weighting schemes should affect the matrices in complex ways, resulting in different proportions of missing data, and differences in composition of concatenates and the labels attached to them. As we only retain the widely uniform concatenates, that is, a set of "core" sequences not directly affected by differences in the weighting function, we measure indirect effects of these parameters on the trees, which may mask substantially greater differences between matrices. For example, such differences are evident from the greatly expanded number of concatenates under the 2:0:0 weighting scheme which includes nearly 600 additional terminals (Table 4). Although the outcomes of these tree searches are somewhat inconclusive, the proposed procedure will improve the matrix by the more efficient use of the rapidly growing sequences from incompletely identified individuals, objective selection of exemplars for inclusion in the supermatrix, and formal resolution of inconsistencies in public sequence repositories. In addition, the clustering step that precedes the concatenation also results in a substantially greater proportion of data that bear on the phylogeny of a focal group. Compared with conventional name-based analyses, the use of sequence clusters reduces the number of terminals with little phylogenetic information content from population studies. In the case of the Coleoptera database, this resulted in ~300 fewer terminals, a ~20% faster search time and a slightly, but significantly improved topology as measured by the tRI, lnL, and bootstrap support.

It remains to be studied what these mixed concatenates mean biologically. In our analysis, there are 2 layers contributing to mixed species units, at the level of the primary clusters and when matching clusters across loci. Mixed sequence clusters concern tip-level incongruence within genera, although a few clusters also included among-genus, among-subfamily and even higher taxa inconsistencies. The latter probably result from errors in gene annotations or identification, while at the tip-level true gene tree incongruence may be responsible for the clustering of multiple names (in addition to inappropriate clustering methods and parameter settings).

The second layer responsible for mixed terminals results from the locus matching step itself, which may aggravate the problems from mixed primary clusters. However, in the Coleoptera data set, this potentially negative effect is offset by the overall reduced proportion of mixed clusters. For example, under the 2:1:1 scheme that permits partial matches the number of chimerical concatenates was reduced by ~10% compared with the 2:0:0 scheme without partial matches (Table 4), as mixed-name clusters are mostly concatenated with mixed-name clusters in other loci. Likewise, the splitting of COI sequence clusters compared with name-based groups, which is evident in single-locus terminals, is counteracted by the concatenation with other loci (Table 5). Hence, linking partially matching clusters reduces, rather than increases the number of mixed terminals, by combining clusters that would otherwise remain unmatched. Establishing the best matching of mixed clusters may produce biologically meaningful entities from multiple loci whose components may deserve further investigation. The advantage of the procedure for building supermatrices will increase with the rapid growth of sequence-based taxonomy and the expansion of loci available for phylogenetic analysis.

Supplementary Material

A Perl script implementing the multipartite matching is available under the GNU General Public License at http://sourceforge.net/projects/multilocusmotu/files/. Supplementary material, including data files and/or online-only appendices, can be found in the Dryad data repository at http://datadryad.org doi:10.5061/dryad.qp367.

FUNDING

This work was supported by NERC CASE studentship NER/S/A/2006/14013 (to D.C.).

ACKNOWLEDGMENTS

The authors would like to thank Patrick Brock for useful discussions on Blast-based taxonomic clustering and Peter Foster for advice on maximum-likelihood analyses. We thank Karl Kjer, Rod Page, and an anonymous referee for numerous helpful comments. Computation was performed using the resources of Imperial College Bioinformatic Support Service and the Scientific Computing Grid (SCGrid) of the Chinese Academy of Sciences.

REFERENCES

Bandelt H.J., Crama Y., Spieksma F.C.R. 1994. Approximation algorithms for multidimensional assignment problems with decomposable costs. Discrete Appl. Math. 49:25–50.

Belshaw R., Katzourakis A. 2005. *BlastAlign*: a program that uses *blast* to align problematic nucleotide sequences. Bioinformatics 21: 122–123.

Brock P.M., Doring H., Bidartondo M.I. 2009. How to know unknown fungi: the role of a herbarium. New Phytol. 181:719–724.

Cherkassky B.V., Goldberg A.V., Radzik T. 1996. Shortest paths algorithms: theory and experimental evaluation. Math. Program. 73:129–174.

Danforth B.N., Lin C.P., Fang J. 2005. How do insect nuclear ribosomal genes compare to protein-coding genes in phylogenetic utility and nucleotide substitution patterns? Syst. Entomol. 30:549–562.

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

de Queiroz A., Gatesy J. 2007. The supermatrix approach to systematics. Trends Ecol. Evol. 22:34–41.

Driskell A.C., Ane C., Burleigh J.G., McMahon M.M., O'Meara B.C., Sanderson M.J. 2004. Prospects for building the tree of life from large sequence databases. Science 306:1172–1174.

Edwards S.V. 2009. Is a new and general theory of Molecular Systematics emerging? Evolution 63:1–19.

Funk D.J., Omland K.E. 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annu. Rev. Ecol. Syst. 34:397–423.

Galil Z. 1986. Efficient algorithms for finding maximum matching in graphs. ACM Comput. Surv. 18:23–38.

Garey M.R., Johnson D.S. 1983. Crossing number is NP-complete. SIAM J. Algebr. Discrete Methods 4:312–316.

Gatesy J., DeSalle R., Wahlberg N. 2007. How many genes should a systematist sample? Conflicting insights from a phylogenomic matrix characterized by replicated incongruence. Syst. Biol. 56: 355–363.

Goloboff P.A., Catalano S.A., Mirande J.M., Szumik C.A., Arias J.S., Kallersjo M., Farris J.S. 2009. Phylogenetic analysis of 73 060 taxa corroborates major eukaryotic groups. Cladistics 25:211–230.

Hajibabaei M., Janzen D.H., Burns J.M., Hallwachs W., Hebert P.D.N. 2006. DNA barcodes distinguish species of tropical Lepidoptera. Proc. Natl Acad. Sci. U S A 103:968–971.

Hunt T., Bergsten J., Levkanicova Z., Papadopoulou A., John O.S., Wild R., Hammond P.M., Ahrens D., Balke M., Caterino M.S., Gomez-Zurita J., Ribera I., Barraclough T.G., Bocakova M., Bocak L., Vogler A.P. 2007. A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. Science 318:1913–1916.

Hunt T., Vogler A.P. 2008. A protocol for large-scale rRNA sequence analysis: towards a detailed phylogeny of Coleoptera. Mol. Phylogenet. Evol. 47:289–301.

Jones M.O., Koutsovoulos G.D., Blaxter M.L. 2011. iPhy: an integrated phylogenetic workbench for supermatrix analyses. BMC Bioinformatics 12:Art. No. 30.

Kupczok A., Schmidt H.A., Haeseler A.v. 2010. Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. Algorithms Mol. Biol. 5:Art. No. 37.

Lee J.H., Yi H., Jeon Y.S., Won S., Chun J. 2012. TBC: a clustering algorithm based on prokaryotic taxonomy. J. Microbiol. 50: 181–185.

McMahon M.M., Sanderson M.J. 2006. Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes. Syst. Biol. 55:818–836.

Meier R., Shiyang K., Vaidya G., Ng P.K.L. 2006. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Syst. Biol. 55:715–728.

Meusemann K., von Reumont B.M., Simon S., Roeding F., Strauss S., Kuck P., Ebersberger I., Walzl M., Pass G., Breuers S., Achter V., von Haeseler A., Burmester T., Hadrys H., Wagele J.W., Misof B. 2010. A phylogenomic approach to resolve the arthropod tree of life. Mol. Biol. Evol. 27:2451–2464.

Papadimitriou C.H., Steiglitz K. 1982. Combinatorial optimization: algorithms and complexity. Englewood Cliffs (NJ): Prentice-Hall.

Peters R.S., Meyer B., Krogmann L., Borner J., Meusemann K., Schutte K., Niehuis O., Misof B. 2011. The taming of an impossible child: a standardized all-in approach to the phylogeny of Hymenoptera using public database sequences. BMC Biol. 9:55.

Philippe H., Snell E.A., Bapteste E., Lopez P., Holland P.W.H., Casane D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol. Biol. Evol. 21:1740–1752.

Poe S. 1998. Sensitivity of phylogeny estimation to taxonomic sampling. Syst. Biol. 47:18–31.

Pons J., Barraclough T.G., Gomez-Zurita J., Cardoso A., Duran D.P., Hazell S., Kamoun S., Sumlin W.D., Vogler A.P. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst. Biol. 55:595–609.

Ratnasingham S., Hebert P.D.N. 2007. BOLD: The Barcode of Life Data System (www.barcodinglife.org). Mol. Ecol. Notes 7:355–364.

Rokas A., Nylander J.A.A., Ronquist F., Stone G.N. 2002. A maximum-likelihood analysis of eight phylogenetic markers in gallwasps (Hymenoptera : Cynipidae): implications for insect phylogenetic studies. Mol. Phylogenet. Evol. 22:206–219.

Ryberg M., Nilsson R.H., Kristiansson E., Topel M., Jacobsson S., Larsson E. 2008. Mining metadata from unidentified ITS sequences in GenBank: a case study in *Inocybe* (Basidiomycota). BMC Evol. Biol. 8:Art. No. 50.

Sanderson M.J., Boss D., Chen D., Cranston K.A., Wehe A. 2008. The PhyLoTA browser: processing GenBank for molecular phylogenetics research. Syst. Biol. 57:335–346.

Shimodaira H., Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17: 1246–1247.

Simon S., Strauss S., von Haeseler A., Hadrys H. 2009. A phylogenomic approach to resolve the basal pterygote divergence. Mol. Biol. Evol. 26:2719–2730.

Smith S.A., Beaulieu J.M., Donoghue M.J. 2009. Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. BMC Evol. Biol. 9:Art. No. 37.

Stajich J.E., Block D., Boulez K., Brenner S.E., Chervitz S.A., Dagdigian C., Fuellen G., Gilbert J.G.R., Korf I., Lapp H., Lehvaslaiho H., Matsalla C., Mungall C.J., Osborne B.I., Pocock M.R., Schattner P., Senger M., Stein L.D., Stupka E., Wilkinson M.D., Birney E. 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 12:1611–1618.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Swofford D.L. 2002. PAUP*: phylogenetic analysis using parsimony. Version 4.0b. Sunderland (MA): Sinauer Associates.

Tautz D., Arctander P., Minelli A., Thomas R.H., Vogler A.P. 2003. A plea for DNA taxonomy. Trends Ecol. Evol. 18:70–74.

Thomson R.C., Shaffer H.B. 2010. Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. Syst. Biol. 59:42–58.