OPINION

Direct optimization overly optimizes data

Kazunori Yoshizawa

Systematic Entomology, Hokkaido University, Sapporo, Japan

Correspondence: Systematic Entomology, Graduate School of Agriculture, Hokkaido University, Sapporo, 060-8589, Japan. E-mail: psocid@res.agr.hokudai.ac.jp

Introduction

Direct optimization is a criterion that recognizes sequence alignment and tree search as a single epistemological problem, which is performed simultaneously. When multiple datasets are analyzed under the direct optimization criterion, all data partitions are combined and optimized simultaneously along with the same tree topology (Wheeler & Hayashi, 1998; Wheeler, 2003). Each data partition of the combined data is independent in the sense that the homologies are not shifted between partitions, but every data partition has the potential to influence the optimizations of other partitions (Simmons, 2004). The criterion has been adopted mainly for analyses of DNA sequences but has been adopted also for behavioral (Robillard et al., 2006) and morphological (Agolin & D'Haese, 2009) characters for which homology assessment between taxa is ambiguous (dynamic homology). *Systematic Entomology* has published many important contributions to insect systematics partly or fully based on the direct optimization criterion, with little or no discussion of the procedures and methodological assumptions (Edgecombe et al, 2002; Hebsgaard et al, 2004; Robertson et al., 2004; Whiting & Whiting, 2004; Damgaard et al., 2005; Jarvis et al., 2005; Schuh et al., 2009; Kehlmaier & Assmann, 2010).

One of the most important contributions to systematic entomology based on the direct optimization criterion is by Terry & Whiting (2005a), who provide new insights into very poorly understood phylogenetic relationships among orthopteroid insect orders. Among the many gene sequences presented in the paper, the 18S sequence of *Zorotypus hubbardi* (order Zoraptera: AY521890; voucher BYU_ACZO001) was used subsequently by Xie et al. (2009), who analyzed its secondary structure. This showed that the 18S of Zoraptera had similar secondary structure to that of other insects. In contrast, previous studies had shown that 18S of zorapterans contains unusual evolutionary motifs including modifications of secondary structure (Kjer, 2004; Yoshizawa & Johnson, 2005). The sequence was used here as a BLAST search (http://blast.ddbj.nig.ac.jp/: Sept. 24, 2009), but the search returned no other 18S from Zoraptera, not even two other 18S known to be from the same species, among the top 10,000 hits. Pairwise similarity between this peculiar sequence and the other 18S sequences from *Z. hubbardi* (DQ013288, AY521892) was 73-76% [In the original publication AY521892 was labeled as *Z. snyderi*.] In contrast, BLAST search using 28S from the same sample (BYU_ACZO001: AY521823) returned 28S of *Z. hubbardi* (AY521825) and *Z. snyderi* (AF354702) as the highest matches (99%). Furthermore, the peculiar zorapteran 18S sequence showed its highest match (97%) to 18S of *Tagalina* sp. (AY521838), an earwig (Dermaptera). The hypervariable region (E23_3 of Xie et al., 2009) of AY521890 shared a couple of unambiguously alignable regions (≥ 15 bp each) with dermapterans, but not with other zorapterans (see Supporting Information SI). This evidence shows that the peculiar 18S sequence of *Z. hubbardi* may be a contaminant. Use of contaminant sequence is a serious problem, but this contaminant sequence seems to illuminate a peculiar behavior of the direct optimization method (Wheeler, 2003) and the software package implementing the method, POY (Wheeler et al., 1996-2003; Varón et al., in press).

Terry & Whiting (2005a) used the combined 18S rRNA, 28S rRNA and Histone 3 gene loci as well as morphology. POY analysis of the combined data extracted signals from 18S supporting the monophyly of Zoraptera (partitioned Bremer support value: PBS 3). In contrast, in the recovered tree derived from a Clustal alignment, PBS from 18S for Zoraptera has an extremely negative value (-25). In comparison, there was significant support from 28S (PBS 49), Histone 3 (PBS 17) and morphology (PBS 33). Clustal alignments can be problematic also with data containing many insertions/deletions (Golubchik et al., 2007), but considering the contaminant zorapteran 18S sequence, the resulting PBS values from the Clustal alignment are more reasonable than those estimated from the POY analysis.

Among the data partitions, morphology, Histone 3 and some conserved regions of the18S and 28S sequences were pre-aligned manually and not allowed to change homology statements during the direct optimization (Terry & Whiting, 2005a). If the 28S and Histone 3 of zorapterans were extracted, amplified, sequenced, and edited correctly, the pre-aligned regions of these genes probably contained strong signals for the monophyly of Zoraptera, as indicated by PBS. Furthermore, morphological data were coded as constant throughout each order such that the morphological data contained very strong signal for the monophyly of each order. When all data are combined and optimized simultaneously under the direct optimization criterion, the signals from each partitions can influence the homology statements of other partition (Simmons, 2004). Therefore, it is likely that the contaminant 18S was forced to homologize with the other included zorapteran 18S sequences. As a result, pseudo-signal support for a monophyletic Zoraptera would be recovered by POY from the highly variable regions of the contaminant 18S. Much of 18S is invariant among insect orders, whereas some hypervariable regions are essentially randomized (Whiting, 2002; Xie et al., 2009). The signal from the 18S loci supportive of zorapteran monophyly suggests that, using POY, positive signal for the tree may be extracted even though non-homologous sequence has been included (Simmons, 2004).

This supposition seems to be supported from other tests. Using the deep orthopteroid phylogeny data, Terry & Whiting (2005b) compared the ILD values of alignments, a value to quantify incongruence among data partitions (Mickevich & Farris, 1981; Wheeler & Hayashi, 1998), obtained by POY under various parameter settings. They obtained the highest congruence between gene and morphology partitions when the gap: transition: transversion ratio was 1:1:1, which is the cost matrix finally employed for POY analysis of deep orthopteroid phylogeny by Terry & Whiting (2005a). In analyzing the effect of the gap and gap-extension cost parameters in POY, Kjer et al. (2007) also found an unambiguous sharp optimal peak at gap cost: gap-extension cost of 1:1. In contrast, under high gap costs (approximately > 10), the ILD values of implied alignments produced by POY became significantly worse (Terry & Whiting, 2005b). These results are suggestive because, when the cost of a new gap is low, gaps can be inserted at preferred sites quite freely. It appears that minimization of incongruence under low gap cost parameters could artificially maximize pseudo-homology within the alignment. In their examination of the ILD values of alignments resulting from Clustal under various parameter settings, Terry & Whiting (2005b) identified a dramatic decrease in congruence under a gap insertion cost of 1 and low transversion weights. Although parameter settings for POY and Clustal cannot be directly compared (Terry & Whiting, 2005b), this parameter area for Clustal is close approximation to the 1:1:1 setting in POY.

In this paper, I test the behaviors of POY and Clustal to clarify the above mentioned predictions. The performances of POY and Clustal have been compared previously by Ogden & Rosenberg (2007) based on substantial simulation data, which showed that Clustal produces a more accurate alignment in most cases. In a different approach, I employ a random-sequence-based test to examine the behaviors of these two programs. Based on the results from the random sequence test, I also evaluated a previously published implied alignment produced using POY.

Materials and Methods

For the test of random sequences, I generated 15 x 1000 bp random sequences with equal base frequencies using MacClade (Maddison & Maddison, 2000). The artificial random sequences were appended to the data (17 taxa) presented in Yoshizawa (2004) (= the real data) and then the entire artificial partition (including two empty taxa) was randomized using the Shuffle option in MacClade. By this means, gaps (about 12% of the total data) were distributed randomly within the data to produce random sequences with length variation (857-907 bp). These were aligned/analyzed using ClustalX version 2 (Larkin et al., 2007) and POY version 4 (Varón et al., in press). For ClustalX, alignments were performed under two parameter settings: 1) the default setting; and 2) gap opening cost = 1, transition weight = 0. These settings were selected because Terry & Whiting (2005b) identified extreme data incongruence under the latter setting, whereas the default setting led to ILD scores not significantly different from the lowest ones found (at gap opening=50, transition weight=0.9). In POY, 10 TBR replicates with default search strategy were performed with gap: transition: transversion = 1:1:1, the best parameter setting identified by Terry & Whiting (2005b), and with 15:1:1 as an approximation to the default parameter setting of Clustal. Analyses were performed on the random data alone and combined with the real data. The alignments produced by POY and Clustal were analyzed using PAUP* 4.0b10 (Swofford, 2002) with TBR branch swapping. If an analysis identifies strong phylogenetic signal from the random sequences, then the alignment under those parameter settings is considered to be problematic because random data should contain no phylogenetic signal beyond what could be expected to occur by chance alone.

Further, I sought to examine the behavior of POY using a previously published data set. Unfortunately, the implied alignments of Terry & Whiting (2005a) were not available from either the given URL or their lab website (final confirmation on Dec. 25, 2009). Thus the implied alignment from a molecular phylogeny of robber flies (Bybee et al., 2004) was downloaded instead (from the website of the same research group: http:// whitinglab.byu.edu: accessed on Nov. 30, 2009). I selected this data set because it was subdivided clearly into conserved and variable regions by the authors. These data contain sequences of 16S, 18S, and 28S rDNA and COII (a total of 4906 characters) subdivided into 55 blocks. Character blocks 11-16 (16S), 31-36 (18S), and 50-54 (28S) correspond to variable regions which were optimized as unfixed blocks using POY at the 1:1:1 parameter setting (Fig. 3, bottom). Based on the implied alignment, I prepared three data sets: a full data set containing both conserved and variable blocks (4906 characters), a variable data set containing the variable data blocks only (blocks 11-16, 31-36, and 50-54; in total 923 characters), and a conserved data set containing the conserved data blocks only (3983 characters). These data sets were analyzed under the parsimony criterion using PAUP* with 100 random starting trees and TBR branch swapping, with the gaps treated as a 5th character state in accordance with Bybee et al. (2004). The extent of gaps in each aligned sequence and the consistency (CI) and retention indexes (RI) of estimated trees were calculated using MacClade. PBS values were calculated using TreeRot v.3 (Sorenson & Franzosa, 2007).

Results and Discussion

Random Data Test

Bootstrap consensus trees (100 replicates) estimated from the real data (MP and NJ) and from the aligned random data (MP) are shown (Figs 1-2). Gaps were treated as a 5th character state, but congruent results were obtained when gaps were treated as missing data. MP and NJ bootstrap trees estimated from the real data as presented in Yoshizawa (2004) were fairly well resolved and almost concordant with one another (Fig.

1, top). When the random data were aligned with the real data (Fig. 1, middle and bottom), both Clustal and POY under different parameter settings recovered pseudo-signals from the random data that are congruent with the real data. This result shows that, under all situations examined here, the actual alignment of randomized regions is affected by the signal in well-aligned regions (Simmons, 2004). Importantly, more pseudo-signal was extracted under lower gap cost settings. Contrary to the behavior identified by the ILD score (Terry & Whiting, 2005b), POY and Clustal behaved similarly for the random data at low gap cost settings, but POY with setting 1:1:1 recovered a much stronger signal: a fully resolved tree, perfect congruence with the real data, and all branches with 100% bootstrap values.

I then generated four additional random data sets and analyzed them by POY at 1:1:1 setting. In all cases, POY resulted in fully resolved and fully supported trees with high congruence with real data (with three minor exceptions: SPI2+3 became paraphyletic with respect to SPI1+COR in data 2; SEX1+2 became paraphyletic with respect to MED in data 3; SPI1 clustered with MAJ in data 5: trees not shown but available online). This result clearly shows that any poorly aligned, homology-ambiguous data may adhere to consistent data partitions when combined and aligned simultaneously under low gap costs, and the effect is stronger for direct optimization (Simmons, 2004).

Trees recovered from the random data analyzed independently are shown in Fig. 2. These trees are not congruent with those obtained from the same random data aligned with real data, but resolution and support for the trees was higher under low gap costs. This shows that the stochastic similarities among random sequences also greatly affect the final alignment under low gap cost. Again, the effect was stronger for direct optimization (all branches received 100% bootstrap support) than for Clustal.

Empirical Data Test

Parsimony analysis of the full robber flies' data set (conserved + variable) yielded a well supported tree (Fig. 3, top-left, black: all branches received ≥84% bootstrap support). However, when the conserved blocks were analyzed independently, the data yielded only a weakly supported tree (Fig. 3, top-right: many deep branches received <50% bootstrap support). In contrast, the variable data blocks analyzed independently yielded a tree that agreed exactly with the tree estimated from the full data set (Fig. 3 top-left, red). Furthermore, bootstrap values show that the variable blocks contain a stronger signal than that of the full data set (Fig. 3). The PBS values also show that the variable blocks generally contain more signal than the conserved blocks, except for some branches strongly supported by the conserved blocks alone. At three branches, the conserved blocks even contain no or sometimes negative signal for the tree from the full data set. There are some differences in topology between trees estimated from the conserved and variable/full blocks, too (indicated by green lines). This shows clearly that the topology from the full data set relies highly on the ambiguously aligned variable blocks, and conserved blocks actually contain little signal in resolving deep branches.

In contrast to the random data, the aligned variable blocks may contain real phylogenetic signal. However, the aligned random sequences shows that POY at the 1:1:1 setting could produce highly congruent alignments even from random sequences, by filling 58.5-60.0% of each sequence with gaps. As shown in Fig. 3 (bottom), an implied alignment of the variable blocks produced by POY also contains a huge numbers of gaps, which occupy 53.4-68.8% of each sequence. The proportion of gaps within the empirical data matrix exceeds even that produced from the random sequences. Therefore, uncertainty of the aligned variable blocks seems to be not significantly different from that of the aligned random sequences. In addition, CI and RI values calculated from each data set show that the conserved blocks contain more homoplasy than is contained in the variable blocks (Fig. 3). It is improbable that the ambiguously aligned regions contain more and clearer phylogenetic signal than the well-aligned regions, and strong signal extracted from such unreliable alignment should be regarded

as artifacts.

Conclusion

The present examination shows that the alignment of highly variable regions under low gap cost setting is problematic for both POY and Clustal because it likely provides artificial maximization of pseudo-homology within the alignment. Especially, the bias for specific topology seems to be more pronounced when using direct optimization (Figs 1-2: Simmons et al., 2008; Wheeler & Giribet, 2009), probably due to repeated reciprocal refinement of the alignment and the tree (Wheeler et al., 2006a).

If the variable regions are aligned together with conserved regions, then the bias could emerge in two ways. When the conserved regions contain only weak phylogenetic signal, stochastic similarities within the variable regions strongly affect the final alignment. This type of bias can emerge even if the variable regions are aligned independently at a low gap cost setting (Fig. 2). As a result, aligned variable sequences may provide a high amount of pseudo-signal which could even mask any weak true signal contained in the conserved regions (Fig. 3). Alternatively, when the conserved regions contain significant phylogenetic signal, the variable regions will play as adherents of the conserved data (Fig. 1: Simmons, 2004) even if the variable regions contain no or even contradicting phylogenetic signal. The adherent bias is especially problematic when data set contains heterogeneous sequences.

Incongruence among data partitions can be produced not only by contamination but also by evolutionary factors, such as gene introgression (Sota & Vogler, 2001; Bossu & Near, 2009), recombination (Sota & Sasabe, 2006), or lineage sorting (McCracken & Sorenson, 2005). Such real phylogenetic incongruence will be hidden by direct optimization of multiple genes. The ILD or other congruence-based scores (Wheeler et al., 2006b) in particular will mislead in selecting optimal alignment parameters, because parameters providing greater artificial congruence (= low gap cost) would be preferred. In fact, the problematic 1:1:1 setting has been selected as the best parameter setting in a majority of POY analyses (Ogden & Rosenberg, 2007). Despite this, Wheeler & Hayashi (1998) state "Only this method of analysis [= a combination of congruence test and direct optimization] offers the optimality of character congruence and the complementarity of total evidence (p. 188)". Can we consider such a method optimal and complemental ?

In addition, during direct optimization, each nucleotide position is treated as a separate unit, and alignments of base pairs are allowed to change in any preferred way without regarding any independent knowledge about sequence evolution or structure (Rieppel, 2007; Morrison, 2009b). Wheeler & Giribet (2009) state that "Alignments are not an attribute of nature (p. 100)" and "their role can only be as a heuristic tool in the solution of phylogenetic problems (p. 102)". However, if descent with modification exists, then homology is a real phenomenon. Both trees and alignments are ways of representing homology, and their estimations belong to biological issues rather than mathematical optimality problems (Morrison, 2006, 2009a,b). At least so far as the alignment of ribosomal DNA is concerned, structure- (Kjer, 2004; Kjer et al., 2007; Misof et al., 2007; Kjer et al., 2009) or similarity-based approaches (Simmons, 2004) should have primacy over automated congruence-based approach (Rieppel, 2007).

Supporting Information

Additional Supporting Information may be found in the online version of this article under DOI:

File S1. Alignments generated (zip file).

Acknowledgments

I thank Karl Kjer for discussion on the idea. I also thank David Morrison, three anonymous reviewers, and Peter Cranston, the editor of *Systematic Entomology*, for their constructive comments on the manuscript. This study was supported partly by JSPS grant 18770058. Alignments generated are available also at http://psocodea.org/kazu/data/poy.

References

Agolin, M. & D'Haese, C. A. (2009) An application of dynamic homology to morphological characters: direct optimization of setae sequences and phylogeny of the family Odontellidae (Podouromorpha, Collembola). *Cladistics* **25**, 353-385.

Bossu, C. M. & Near, T. J. (2009) Gene trees reveal repeated instances of mitochondrial DNA introgression in orangethroat darters (Percidae: *Etheostoma*). *Systematic Biology* **58**, 114-129.

Bybee, S. M., Taylor, S. D., Nelson, C. R. & Whiting, M. F. (2004) A phylogeny of robber flies (Diptera: Asilidae) at the subfamilial level: molecular evidence. *Molecular Phylogenetics and Evolution* **30**, 789-797.

Damgaard, J., Andersen, N. M. & Meier, R. (2005) Combining molecular and morphological analyses of water strider phylogeny (Hemiptera-Heteroptera, Gerromorpha): effects of alignment and taxon sampling. *Systematic Entomology* **30**, 289-309.

Edgecombe, G. D., Giribet, G. & Wheeler, W. C. (2003) Phylogeny of Henicopidae (Chilopoda: Lithobiomorpha): a combined analysis of morphology and five molecular loci. *Systematic Entomology* **27**, 31-64.

Golubchik, T., Wise, M. J., Easteal, S. & Jermiin, L. S. (2007) Mind the gaps: Evidence of bias in estimates of multiple sequence alignments. *Molecular Biology and Evolution* **24**, 24233-2442.

Hebsgaard, M. B., Andersen, N. M. & Damgaard, J. (2004) Phylogeny of the true water bugs (Nepomorpha: Hemiptera: Heteroptera) based on 16S and 28S rDNA and morphology. *Systematic Entomology* **29**, 488-508.

Jarvis, K. J., Haas, F. & Whiting, M. F. (2005) Phylogeny of earwigs (Insecta: Dermaptera) based on molecular and morphological evidence: reconsidering the classification of Dermaptera. *Systematic Entomology* **30**, 442-453.

Kehlmaier, C. & Assmann, T. (2010) Molecular analysis meets morphology-based systematics - a synthetic approach for Chalarinae (Insecta: Diptera: Pipunclidae). *Systematic Entomology* **35**, 181-195.

Kjer, K. M. (2004) Aligned 18S and insect phylogeny. *Systematic Biology* **53**, 506-514.

Kjer, K. M., Gillespie, J. J. & Ober, K. A. (2007) Opinions on multiple sequence alignment, and an empirical comparison of repeatability and accuracy between POY and structural alignment. *Systematic Biology* **56**, 133-146.

Kjer, K. M., Roshan, U. & Gillespie, J. J. (2009) Structural and evolutionary consideration for multiple sequence alignment of RNA, and the challenges for algorithms that ignore them. Pp. 105-149, in Rosenberg, M. S. (ed), Sequence alignment: Methods, models, concepts, and strategies. University of California Press, Berkeley, CA.

Larkin M. A., Blackshields G., Brown N. P. et al (2007) Clustal W and Clustal X version 2.0. *Bioinfomatics* **23**, 2947-2948.

Maddison, D. R. & Maddison, W. P. (2000) MacClade 4: Analysis of phylogeny and character evolution. Sinauer Associates, Sunderland, MA.

McCracken, K. G. & Sorenson, M. D. (2005) Is homoplasy or lineage sorting the source of incongruent mtDNA and nuclear gene trees in the stiff tailed ducks (*Nomonyx-Oxyura*)? *Systematic Biology* **54**, 35-55.

Mickevich, M. F. & Farris, J. S. (1981) The implications of congruence in *Menidia*. *Systematic Zoology* **30**, 351-370.

Misof, B., Niehuis, O., Bischoff, I., Rickert, A., Erpenbeck, D. & Staniczek, A. (2007)

Towards an 18S phylogeny of hexapods: Accounting for group specific character covariance in optimized mixed nucleotide/doublet models. *Zoology* **110**, 409-429.
Morrison, D. A. (2006) Multiple sequence alignment for phylogenetic purposes. *Australian Systematic Botany* **19**, 479-539.
Morrison, D. A. (2009a) Why would phylogeneticists ignore computerized sequence alignment? *Systematic Biology* **58**, 150-158.
Morrison, D. A. (2009b) A framework for phylogenetic sequence alignment. *Plant Systematics and Evolution* **282**, 127-149.
Ogden, T. H. & Rosenberg, M. S. (2007) Alignment and topological accuracy of the direct optimization approach via POY and traditional phylogenetics via ClustalW + PAUP*. *Systematic Biology* **56**, 182-193.
Ogden, T. H., Whiting, M. F. & Wheeler, W. C. (2005) Poor taxon sampling, poor character sampling, and non-repeatable analyses of a contrived data set do not provide a more credible estimate of insect phylogeny: a reply to Kjer. *Cladistics* **21**, 295-302.
Rieppel, O. (2007) The nature of parsimony and instrumentalism in systematics. *Journal of Zoological Systematics and Evolutionary Researches* **45**, 177-183.
Robertson, J. A., Mchugh, J. V. & Whiting, M. F. (2004) A molecular phylogenetic analysis of the pleasing fungus beetles (Coleoptera: Erotylidae): evolution of color patterns, gregariousness and mycophagy. *Systematic Entomology* **29**, 173-187.
Robillard, T., Legendre, F., Desutter-Grandcolas, L. & Grandcolas, P. (2006) Phylogenetic analysis and alignment of behavioral sequences by direct optimization. *Cladistics* **22**, 602-622.
Schuh, R. T., Weirauch, C. & Wheeler, W. C. (2009) Phylogenetic relationships within the Cimicomorpha (Hemiptera: Heteroptera): a total-evidence analysis. *Systematic Entomology* **34**, 15-48.
Simmons, M. P. (2004) Independence of alignment and tree search. *Molecular Phylogenetics and Evolution* **31**, 874-879.
Simmons, M. P., Müller, K. F. & Webb, C. T. (2008) The relative sensitivity of different alignment methods and character coding in sensitivity analysis. *Cladistics* **24**, 1039-1050.
Sorenson, M. D. & Franzosa, E. A. (2007) TreeRot, version 3. Boston University, Boston, MA (software distributed by authors available at http://people.bu.edu/msoren/TreeRot.html).
Sota, T. & Sasabe, M. (2006) Utility of nuclear allele networks for the analysis of closely related species in the genus *Carabus*, subgenus *Ohomopterus*. *Systematic Biology* **55**, 329-344.
Sota, T. & Vogler, A. P. (2001) Incongruence of mitochondrial and nuclear gene trees in the carabid beetles *Ohomopterus*. *Systematic Biology* **50**, 39-51.
Swofford, D. L. (2002) PAUP*: Phylogenetic analysis using parsimony (*and other methods), Version 4.0b10. Sinauer Associates, Sunderland, MA.
Terry, M. D. & Whiting, M. F. (2005a) Mantophasmatodea and phylogeny of the lower neopterous insects. *Cladistics* **21**, 240-257.
Terry, M. D. & Whiting, M. F. (2005b) Comparison of two alignment techniques within a single complex data set: POY versus Clustal. *Cladistics* **21**, 272-281.
Varón, A., Vinh, L. S. & Wheeler, W. C. (in press) POY version 4: phylogenetic analysis using dynamic homologies. *Cladistics* **, **-**.
Wheeler, W. C. (2003) Implied alignment: a synapomorphy-based multiple-sequence alignment method and its use in cladogram search. *Cladistics* **19**, 261-268.
Wheeler, W. C. & Giribet, G. (2009) Phylogenetic hypotheses and the utility of multiple sequence alignment. Pp. 95-104 in Rosenberg, M. S. (ed), Sequence alignment: Methods, models, concepts, and strategies. University of California Press, Berkeley, CA.
Wheeler, W. C., Gladstein, D. & De Laet, J. (1996-2003) POY, Version 3.0. American

Museum of Natural History, New York, NY.
Wheeler, W. C. & Hayashi, C. Y. (1998) The phylogeny of the extant chelicerate orders. *Cladistics* **14**, 173-192.
Wheeler, W. C., Aagesen, L., Arango, C. P., Faivovich, J., Grant, T., D'Haese, C., Janies, D., Smith, W. L., Varon, A. & Giribet, G. (2006a) Dynamic homology and phylogenetic systematics: a unified approach using POY. American Museum of Natural History, New York, NY.
Wheeler, W. C., Ramírez, M. J., Aagesen, L. & Schulmeister, S. (2006b) Partition-free congruence analysis: implications for sensitivity analysis. *Cladistics* **22**, 256-263.
Whiting, M. F. (2002) Phylogeny of the holometabolous insect orders: molecular evidence. *Zoologica Scripta* **31**, 3-15.
Whiting, M. F. & Whiting, A. S. (2004) Is wing recurrence *really* impossible?: a reply to Trueman *et al. Systematic Entomology* **29**, 140-141.
Xie, Q., Tian, X., Qin, Y. & Bu, W. (2009) Phylogenetic comparison of local length plasticity of the small subunit of nuclear rDNA among all hexapoda orders and the impact of hyper-length-variation on alignment. *Molecular Phylogenetics and Evolution* **59**, 310-316.
Yoshizawa, K. (2004) Molecular phylogeny of major lineages of *Trichadenotecnum* and a review of diagnostic morphological characters (Psocoptera: Psocidae). *Systematic Entomology* **29**, 383-394.
Yoshizawa, K. & Johnson, K. P. (2005) Aligned 18S for Zoraptera (Insecta): Phylogenetic position and molecular evolution. *Molecular Phylogenetics and Evolution* **37**, 572-580.

Accepted January 5th 2010

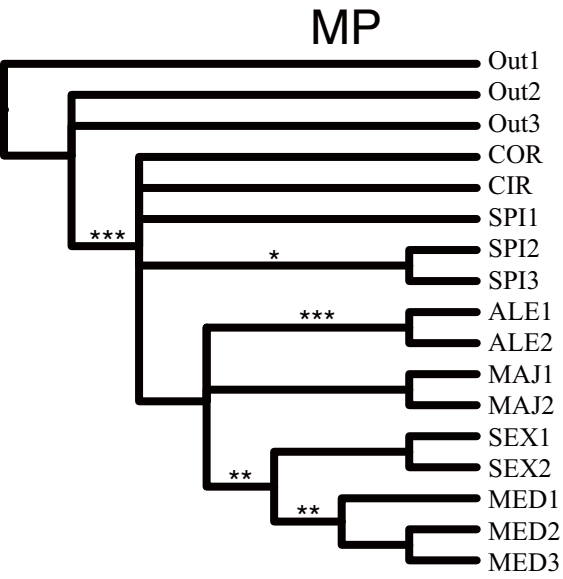
Figure caption

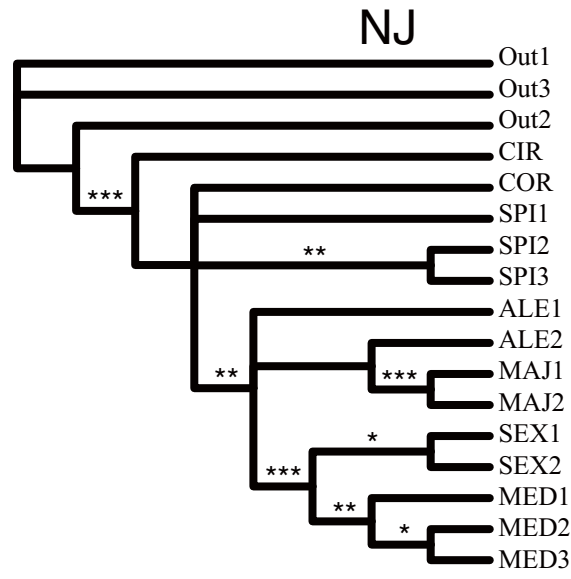
Fig. 1. Bootstrap consensus trees estimated from the real data (Yoshizawa, 2004) and random data alignments (aligned with real data) generated by different parameter strategies (see text for detail). Terminal taxa are labeled according to Table 1. Broken lines indicate branches which disagree between those estimated from the true and random data. Bootstrap support for branches is indicated by asterisks: *** = 100%; ** ≥ 90%; * ≥ 70%; none ≥ 50%.

Fig. 2. Bootstrap consensus tree estimated from data matrices generated from random data aligned independently. See Fig. 1 for detail.

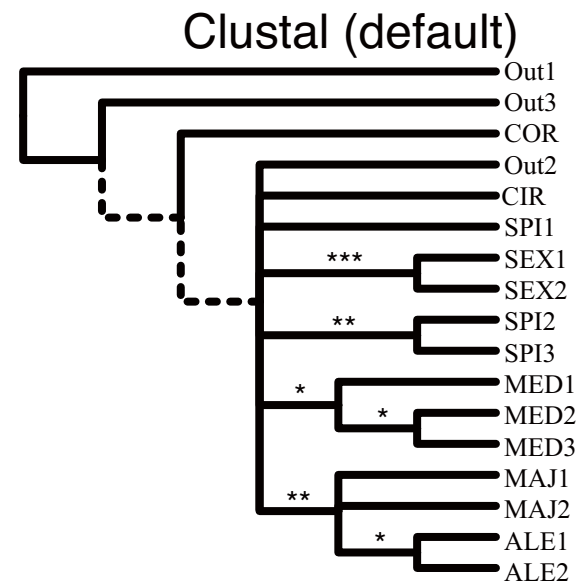
Fig. 3. (Top) Parsimonious trees estimated from the data presented in Bybee et al. (2004). Trees estimated from the full data set (left-black), conserved blocks (right-blue), and variable blocks (left-red) are shown. Numbers without parentheses are bootstrap support values (100 replicates) higher than 50% (black from the full data set and red from the variable blocks). Numbers in parentheses are partitioned Bremer support values from conserved blocks (blue) and variable blocks (red). Taxon labels are according to the data prepared by Bybee et al. (2004). Differences in topologies are indicated by green lines.
(Bottom) Part of the implied alignment presented in Bybee et al. (2004). Different nucleotides are indicated by different colors (red: A; blue: T; green: C; yellow: G). White areas indicate gaps or missing data.

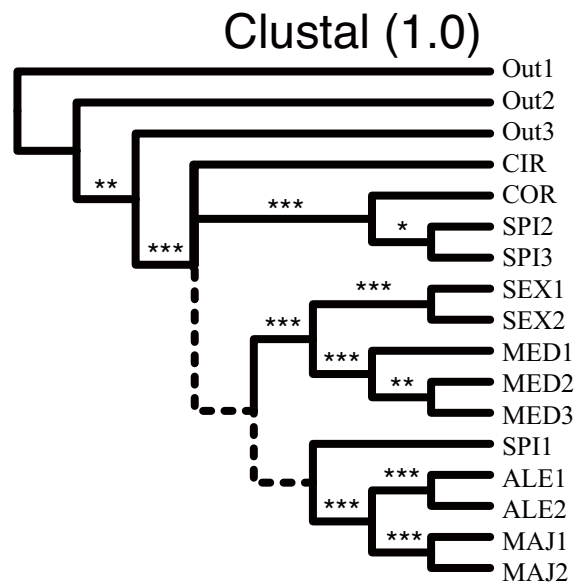
real data

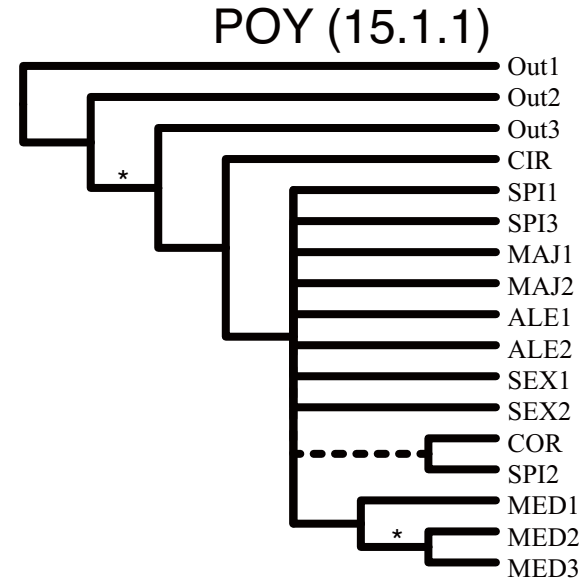


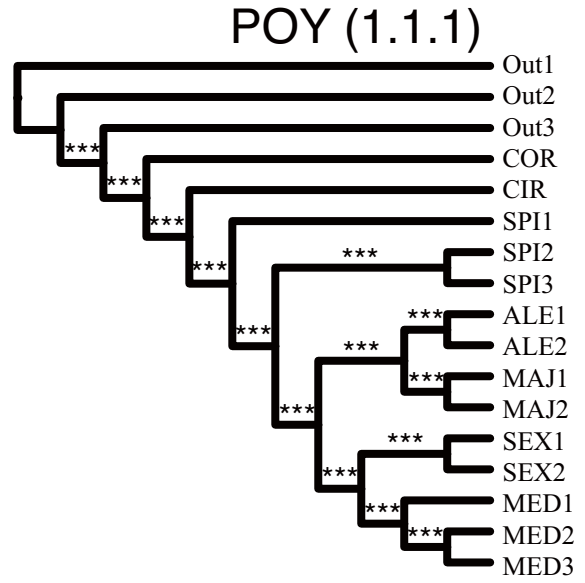


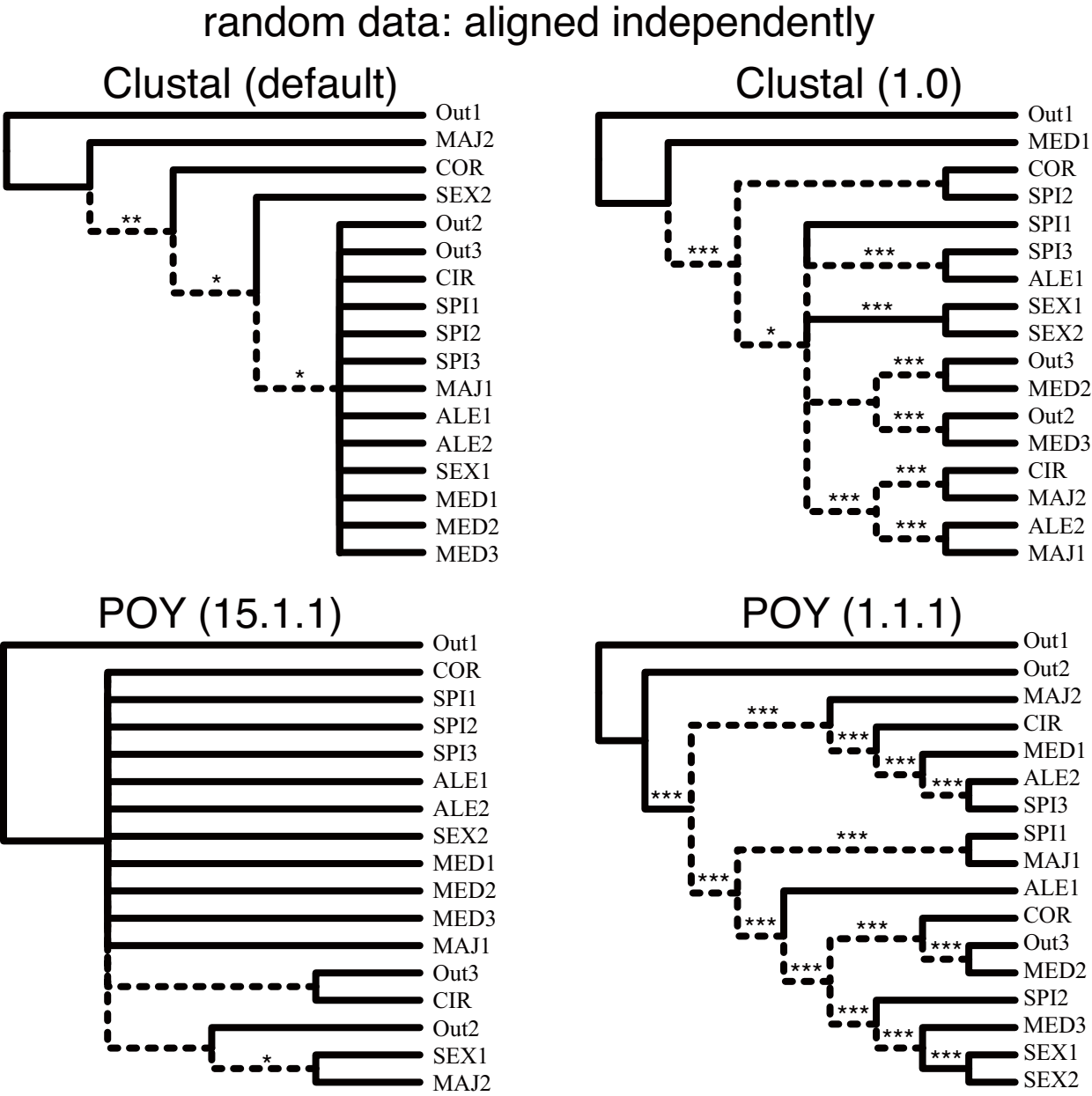
random data: aligned with real data







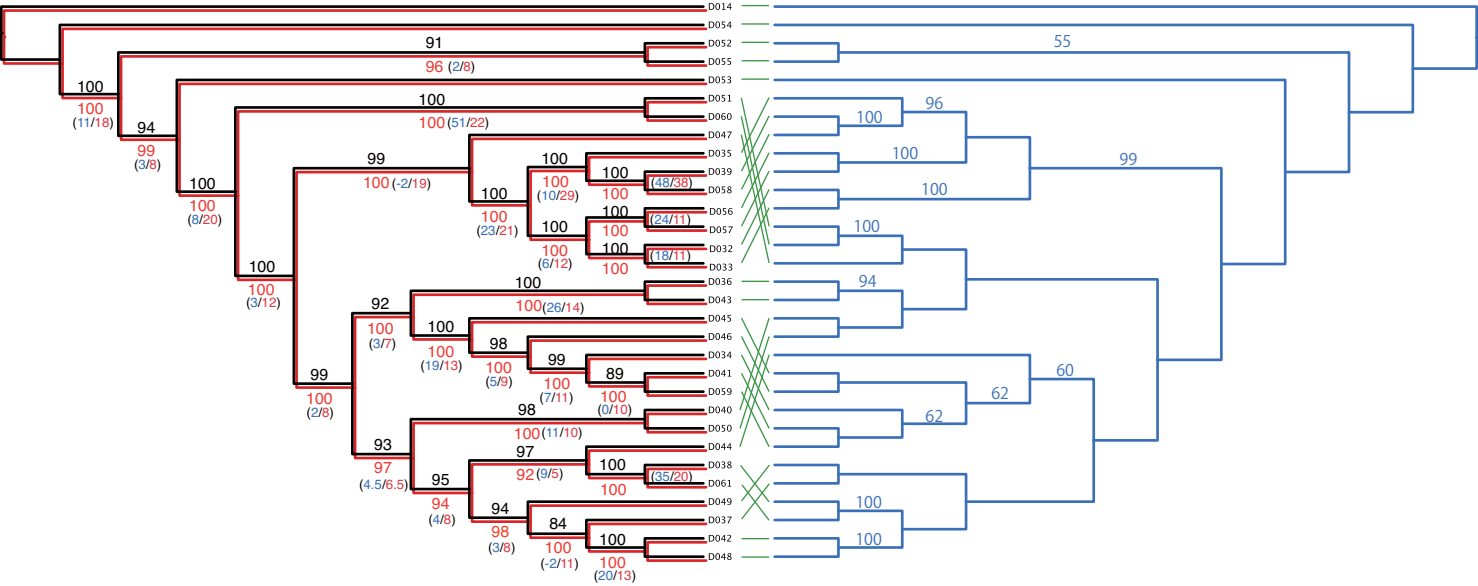


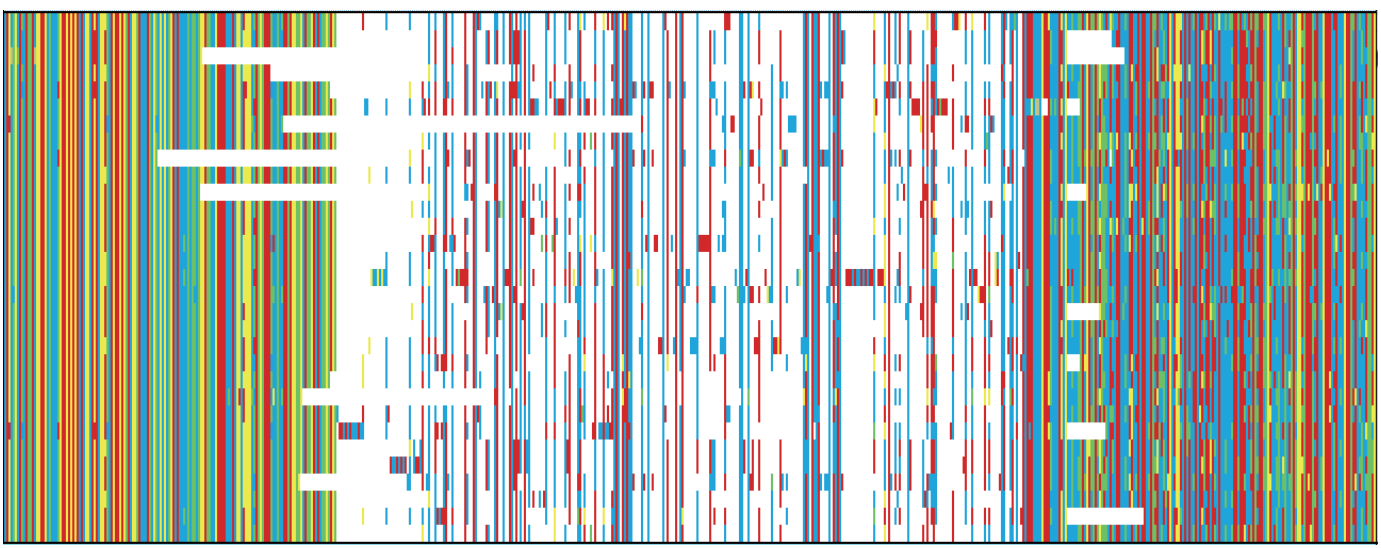

random data: aligned independently
Clustal (default)
Out1
MAJ2
COR
SEX2
Out2
Out3
CIR
SPI1
SPI2
SPI3
MAJ1
ALE1
ALE2
SEX1
MED1
MED2
MED3
**
*
*
Clustal (1.0)
Out1
MED1
COR
SPI2
SPI1
SPI3
ALE1
SEX1
SEX2
Out3
MED2
Out2
MED3
CIR
MAJ2
ALE2
MAJ1
***
***
***
*
***
***
***
***
***
POY (15.1.1)
Out1
COR
SPI1
SPI2
SPI3
ALE1
ALE2
SEX2
MED1
MED2
MED3
MAJ1
Out3
CIR
Out2
SEX1
MAJ2
*
POY (1.1.1)
Out1
Out2
MAJ2
CIR
MED1
ALE2
SPI3
SPI1
MAJ1
ALE1
COR
Out3
MED2
SPI2
MED3
SEX1
SEX2
***
***
***
***
***
***
***
***
***
***
***
***
***

MP tree from full data set (CI=.45, RI=.44)

MP tree from variable blocks (CI=.67, RI=.67)

MP tree from conserved blocks (CI=.35, RI=.33)





28S conserved block (character 48–49)

28S variable block (character 50–54)

COII conserved block (character 55)