Evolution and Phylogenetic Utility of Alignment Gaps Within Intron Sequences of Three Nuclear Genes in Bumble Bees (*Bombus*)

Atsushi Kawakita, Teiji Sota,† John S. Ascher,‡ Masao Ito,§ Hiroyuki Tanaka,|| and Makoto Kato**

*Graduate School of Human and Environmental Studies and †Department of Zoology, Graduate School of Science, Kyoto University, Kyoto, Japan; ‡Department of Entomology, Comstock Hall, Cornell University, Ithaca, New York; §Sapporo Science and Technology College, Sapporo, Japan; and ||Primate Research Institute, Kyoto University, Inuyama, Japan

To test whether gaps resulting from sequence alignment contain phylogenetic signal concordant with those of base substitutions, we analyzed the occurrence of indel mutations upon a well-resolved, substitution-based tree for three nuclear genes in bumble bees (*Bombus*, Apidae: Bombini). The regions analyzed were exon and intron sequences of long-wavelength rhodopsin (LW Rh), arginine kinase (ArgK), and elongation factor–1α (EF-1α) F2 copy genes. LW Rh intron had only a few uninformative gaps, ArgK intron had relatively long gaps that were easily aligned, and EF-1α intron had many short gaps, resulting in multiple optimal alignments. The unambiguously aligned gaps within ArgK intron sequences showed no homoplasy upon the substitution-based tree, and phylogenetic signals within ambiguously aligned regions of EF-1α intron were highly congruent with those of base substitutions. We further analyzed the contribution of gap characters to phylogenetic reconstruction by incorporating them in parsimony analysis. Inclusion of gap characters consistently improved support for nodes recovered by substitutions, and inclusion of ambiguously aligned regions of EF-1α intron resolved several additional nodes, most of which were apical on the phylogeny. We conclude that gaps are an exceptionally reliable source of phylogenetic information that can be used to corroborate and refine phylogenies hypothesized by base substitutions, at least at lower taxonomic levels. At present, full use of gaps in phylogenetic reconstruction is best achieved in parsimony analysis, pending development of well-justified and generally applicable methods for incorporating indels in explicitly model-based methods.

Introduction

Phylogenetic analysis of nucleotide and amino acid sequence data often requires alignment of homologous sequences that vary in length. As a result, gaps are introduced to the data matrix, representing putative insertion or deletion events. As the products of particular evolutionary processes (mutations), indels are often considered as a class of phylogenetic characters to be incorporated in phylogenetic analysis (Wheeler 1996, 1999; Lutzoni et al. 2000; Simmons and Ochoterena 2000) or to be used to corroborate results derived from base substitutions (van Dijk et al. 1999; Rokas and Holland 2000). However, in most phylogenetic analyses gaps are ignored as missing data, or regions containing gaps are simply excluded from data sets. One reason for discarding gap characters in phylogenetic analyses is that use of gaps as characters is generally confined to the parsimony method, since well-justified and generally applicable methods for incorporating indels have yet to be developed for methods based on explicit models of sequence evolution, such as standard implementations of maximum likelihood (see e.g., Swofford et al. 1996). Another reason to exclude gaps is that their positions are often difficult to determine, especially when analyzing a broad range of taxa and/or highly diverged sequences. Phylogenetic reconstruction is highly sensitive to different alignment options (e.g., gap-to-substitution costs or alignment algorithms), which can lead to very different phylogenetic hypotheses (Giribet and Wheeler 1999; Sanchis et al. 2001). As a novel approach within the parsimony framework, Wheeler (1996) developed a direct optimization method for searching most parsimonious trees without prior sequence alignments. Subsequently, Wheeler (1999) and Lutzoni et al. (2000) independently developed similar methods for accommodating ambiguously aligned sequences without violating positional homology.

Despite recent progress in analyzing gap characters, these have not been widely accepted as phylogenetic markers due, in part, to insufficient empirical study of the quality of gaps as characters. Some authors assume that gaps are less homoplastic and therefore more phylogenetically reliable than base substitutions, since gaps generally occur less frequently (Lloyd and Calder 1991; Giribet and Wheeler 1999; van Dijk et al. 1999). However, others emphasize the potential for gaps to be misleading (Bapteste and Philippe 2002). Despite the need for further critical evaluation of gaps as phylogenetic characters, few studies have focused on investigating homoplasy levels of gaps as compared with base substitutions or assessing the contribution of gaps to phylogenetic resolution and nodal support (Graham et al. 2000; Simmons, Ochoterena, and Carr 2001; Bapteste and Philippe 2002). Additional empirical study of relative homoplasy levels among different types of gap characters and of the degree to which these contribute to phylogenetic reconstruction would facilitate the full and appropriate utilization of information potentially available in variable-length sequence data.

In this study, we tested the potential of gaps as phylogenetic characters by analyzing their occurrence upon a well-resolved, substitution-based tree and assessing their contribution to further resolution among 66 species (23 subgenera) of bumble bees (*Bombus*). *Bombus* is a diverse, monophyletic genus comprising nearly 250 species (38 subgenera) (Williams 1998) and is therefore an ideal group for studying patterns of indel evolution among species and subgenera. We analyzed exon and intron

Key words: arginine kinase, elongation factor–1α, long-wavelength rhodopsin, intron, phylogeny, bumble bee.

E-mail: sota@terra.zool.kyoto-u.ac.jp.

Mol. Biol. Evol. 20(1):87–92. 2003
DOI: 10.1093/molbev/msg007

sequences of three nuclear genes: long-wavelength rhodopsin (LW Rh), arginine kinase (ArgK), and elongation factor-1α F2 copy (EF-1α).

Materials and Methods

A list of exemplar species studied, collection data for each exemplar, and GenBank accession numbers are given in supplementary online materials. Genomic DNA was extracted from thoracic muscles using the standard phenol-chloroform method. We PCR-amplified a total of ~2.4 kb of LW Rh, ArgK, and EF-1α F2 copy genes using the following forward (F) and reverse (R) primers: LW Rh, (F) 5′-AAT TGC TAT TAY GAR ACN TGG GT-3′ and 5′-ATA TGG AGT CCA NGC CAT RAA CCA-3′ (Mardulyn and Cameron 1999); ArgK, (F) 5′-GTT GAC CAA GCY GTY TTG GA-3′ and (R) 5′-CAT GGA AAT AAT ACG RAG RTG-3′ or (F) 5′-GA CAG CAA RTC TCT GCT GAA GAA-3′ and (R) 5′-GGT YTT GGC ATC GTT GTG GTA GAT AC-3′; EF-1α, (F) 5′-GGA CAC AGA GAT TTC ATC AAR AA-3′ and (R) 5′-TTG CAA AGC TTC RTG RTG CAT TT-3′. PCR products were directly sequenced using the above primers.

For initial alignment, we used ClustalX version 1.81 (Jeanmougin et al. 1998) with the default parameter settings. The alignments obtained were then corrected manually for obvious misalignments. Sequence alignment for the introns within LW Rh was trivial and required only three simple gaps, which were parsimony-uninformative. However, alignment for introns within ArgK and EF-1α required gaps of various lengths, which differed markedly in terms of structural characteristics; relatively long gaps that were easily aligned occurred throughout ArgK intron sequences (fig. 1*A*), whereas numerous shorter gaps were required at five particular regions within EF-1α intron sequences, resulting in multiple optimal alignments (fig. 1*B*). This difference in gap characteristics provided an excellent opportunity to study different evolutionary modes of length mutations occurring within the diverse genus *Bombus*.

All phylogenetic analyses were done using PAUP* version 4.0b10 (Swofford 2002). Many intron sequences of the outgroup, *Trigona ventralis*, were highly dissimilar to those of the ingroup. We therefore included only unambiguously determined regions (corresponding to 42% of the intron alignment) for the outgroup in the analysis. The results of partition-homogeneity test (Farris et al. 1994), as implemented in PAUP*, suggested that phylogenetic signals within alignment-unambiguous regions of the three genes were highly congruent ($P > 0.5$ in 999 random partitioning for all data comparisons). We then performed a simultaneous analysis of the data set of substitution characters from all unambiguously aligned regions (with gaps treated as missing) to obtain a robust species relationship—the test phylogeny—using the maximum parsimony (MP) method. We conducted heuristic searches with 100 random addition analyses and tree bisection-reconnection (TBR) branch-swapping (Steepest descent option in effect). In order to assess the robustness of the MP tree to the use of explicit models of sequence evolution, we also performed neighbor joining (NJ) and maximum likelihood (ML) analyses. We used the HKY85 model for distance correction in the NJ analysis. To search for a ML tree, we used the quartet puzzling option as implemented in PAUP* with the HKY85+γ substitution model.

We investigated the influence of gap characters on phylogenetic inference by conducting an additional parsimony analysis with gap information included as coded characters. Alignable gaps of ArgK intron were coded as binary (presence/absence) characters using the method of Simmons and Ochoterena (2000) and added to the data matrix. We weighted base substitutions and indel mutations equally, because there was no a priori reason to differentially weight gap costs relative to base substitutions. For the ambiguously aligned regions of EF-1α intron (fig. 1*B*), we employed the program INAASE (Integration of Ambiguously Aligned Sequences) (Lutzoni et al. 2000) to accommodate these regions in parsimony analysis. Following the criteria detailed in Lutzoni et al. (2000), we delimited five ambiguously aligned regions within the EF-1α intron. However, coding one of these regions (region 4) resulted in more than 32 multistate characters, which exceeds the number that can be handled by PAUP*. We therefore divided the region into two subregions such that sequences within each region are likely homologous (region 4a and region 4b in fig. 1*B*). Several division schemes that we tried all yielded the same number of most-parsimonious trees of the same topology, indicating the validity of this approach.

Results and Discussion

Parsimony analysis of the unambiguously aligned region (2301 characters with 395 informative) resulted in 891 most-parsimonious trees of length 1338 (consistency index excluding uninformative characters [CI] = 0.47, retention index [RI] = 0.76, and rescaled consistency index [RC] = 0.45). The topology obtained (fig. 2) is, with

⟶

FIG. 1.—*A*, Partial alignment of ArgK intron. Parsimony-informative gaps, treated as single indel mutations, are indicated by ===. *B. nevadensis* had an insertion of 323 bp replacing an 111–122 bp sequence present in other species. This sequence was apparently unrelated to the sequences at corresponding positions in other species, so we excluded the sequence and coded it as inapplicable in the alignment (poly-n). Similarly, *B. mendax* and *B. defector* shared identical 46 bp sequences that were ambiguously aligned with respect to other sequences of corresponding position (shaded sequences). Because inclusion of these sequences affected inference of indels in other species, we ignored these sequences in the alignment and instead interpreted their origin as a single insertion event (gap-MD). Only sequences that account for informative gaps are given, and only regions that contain these gaps are shown (separated by vertical bars). Numbers above sequences are positions within aligned intron sequence for the first and last codons of each segment. *B*, Partial alignment of EF-1α showing the entire intron sequences. Ambiguously aligned regions are indicated by ===. Examples of different delimitations that we tried are given for region 4 (a, b); the bottommost scheme was used in the analysis. Only sequences that account for alignment ambiguity are shown.

A

```
                                                                           gap-5
                 gap-1            gap-2           gap-3              gap-4  ===     gap-6           gap-7                       gap-8
             1       ===  15 42   ====   55 81 ============         ============   ========          === 144 153                 ===
perplexus    GTATTGAG---CATT GTATTTATTAT-AA AAATATTATTGCAAATTTATTTTTATTCGGTA---ATAA--------CTATAATTATAATAAAT AGT--------------CAG---TAGT
hypnorum     GTATTGAG---GATT GTATTTATTAT-AA AAATATTATTGCAAATTTATTTTTATTCGGTA---ATAA--------CTATAATTATAATAAAT AGT--------------CGG---TAGT
bifarius     GTATTGAGTTAGATT GTATTTATTAT-AA AAATATTATTGCAAATTTATTTTTATTCGGTA---ATAACTATAAAACTATAATTATAATAAAT GGT--------------CAG---TAGT
impatiens    GTATCGAGTTAGATT GTATTTATTAT-AA AAATATTATTGCAAATTTATTTCTATTCGGTA---ATAACTATAAAACTATAACTATAATAAAT GGT--------------CAG---TAGT
melanopygus  GTATTGAG---GATT GTATTTATTATTAA AAATATTATTGCAAATTTATTTTTATTCGGTA---ATAACTATAAAACTATAATTATAATAAAT AGT--------------CAG---TAGT
griseocollis GTATTGAG---GATT GTATTTATTAT-AA AAATATTATCGCAAATTTATTTTT------------TAACTATAAAACTATAATTATAATAAAT AGT--------------CAG---TAGT
crotchii     GTATTGAG---GATT GTATTTATTAT-AA AAATATTATTGCAAATTTATTTTT------------TAACTATAAAACTATAATTATAATAAAT AGT--------------CAG---TAGT
rufocinctus  GTATTGAG---GACT GTATTTATTAT-AA AAATATTATTGCAAATTTATTTCT------------TAACTATAAAACTATAATTATAATAAAT AGT--------------CAG---TAGT
schrencki    GTATTGAG---GATT GTATT----AT-AA AAATATTATTGCAAATTTATTTTTATTCGGTA---GTAATTATAAAAATATAATTATAATAAAT AGT--------------CAG---TAGT
pascuorum    GTATTGAG---GATT GTATT----AT-AA AAATATTATTGCAAATTTATTTTTATTCGGTA---GTAATTATAAAACTATAATTATAATAAAT AGT--------------CAG---TAGT
ashtoni      GTATTGAG---AATT GTATTTATTAT-AA AAATATTATTGCAAATTTATTTTTATTCGGTA---GTAATTATAAAACTATAATTATAATAAAT AGT--------------CAGCAATAGT
bohemicus    GTATTGAG---AATT GTATTTATTAT-AA AAATATTATTGCAAATTTATTTTTATTCGGTA---GTAATTATAAAACTATAATTATAATAAAT AGT--------------CAGCAATAGT
rupestris    GTATTAAG---AATT GTATTTATTAT-AA AAATATTATTGCAAATTTATTTTTATTCGGTA---GTAATTATAAAACTATAATTATAATAAAT AGT--------------CAGCAATAGT
trifasciatus GTATTGAG---GATT GTATTTATTAT-AA AAATATTATTGCAAATTTATTTTTATTCGGTA---GTAATTATAAAACTATAATTATAATAAAT AGT--------------CAG---TAGT
ussurensis   GTATTGAG---GATT GTATTTATTAT-AA AAATATTATTGCAAATTTATTTTTATTCGGTA---GTAATTATAAAACTATAATTATAATAAAT AGT--------------CAG---TAGT
borealis     GTATTGAT---GATT GTATTTATTAT-AA AAATATTATTGCAAATTTATTTTTATTCGGTA---GTAATTATAGAACTATAATTATAATAAAT GGT--------------CGG---TAGT
melanurus    GTATTGAT---GATT GTATTTATTAT-AA AAATATTATTGCAAATTTATTTTTATTCGGTA---GTAATTATAAAACTATAATTATAATAAAT GGT--------------CAG---TAGT
nevadensis   GTATTGAG---AATT GTATTTATnnnnnn nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn nnnnnnnnnnnnnnnnnnnnnnnnnnn
mendax       GTATTGAG---AATT GTATTTATTAT-AA AAA------------TTTATTTTTATTCGGTATATAAAACCATAAAACTATAATTAT---AAAT ATTACTCGTTACTACTG-------AGT
defector     GTATTGAG---AATT GTATTTATTAT-AA AAA------------TTTATTTTTATTCGGTATATAAAACCATAAAACTATAATTAT---AAAT ATTACTCGTTACTACTG-------AGT

                                                                                                                           gap-MD
```

```
                                                                    gap-12
                            gap-9                             gap-11 ===
                              =         gap-10                ============                       gap-13   gap-14              gap-15
                ===================================220 249=========================  282 323   ========    =                   =     369
perplexus    AATTATGATGTTAAAAA--------------GGGAACTAGA CTATAATTACTATT---ACATAGAAATTAGAAG AGAATT--------ATAAAGTATTAATTAATCCTTTATTCTTCCCAG
hypnorum     AATTATGATGTTAAAAA--------------GGGAACTAGA CTATAATTACTATT---ACATAGAAATTAGAAG AGAATT--------ATAAAGTATTAATTAATCCTTTATTCTTCCCAG
bifarius     AGT--------------------------------------- ----------------------------GGAAG AGAATTTAATAAATATAAAGTATTGATTAATCCTTTATTCTTCCCAG
impatiens    AGT--------------------------------------- ----------------------------GGAAG AGAATTTAATAAATATAAAGTATTGATTAATCCTTTATTCTTCCCAG
melanopygus  AGT--------------------------------------- ----------------------------AGAAG AGAATTTAATAAATATAAAGTATTGATTAATCCTTTATTCTTCCCAG
griseocollis AATTATGGTGTTAAAAA--------------GGAAACTAGA CTATTA------------CATAGAAATTAGAAG AGAATTCAATAAATATATAGTATTGATTAATCCTGTATTCTTCCCAG
crotchii     AATTATGATGTTAAAAA--------------GGAAACTAGA CTATTA------------CATAGAAATTAGAAG AGAATTCAATAAATATATAGTATTGATTAATCCTGTATTCTTCCCAG
rufocinctus  AATTACGGTGTTAAAAA--------------GGGAACTAGA CTATAATTGCTATT---ACATAGAAATTAGAAG AGAATTCAATAAATATAAAGTATTGATCAATCCTTTATTCTTCCCAG
schrencki    AATTACAATGTTAAAAA--------------GGGAACTAGA CTATAATTACTATT---ACATAGAAATTAGAAG AGAATTCAATAAATATAAAGTATTGATTAATCCTTTAT-CTT-CCAG
pascuorum    AATTACGATGTTAAAAA--------------GGGAACTAGA CTATAATTACTATT---ACATAGAAATTAGAAG AGAATTCAATAAATATAAAGTATTGATTAATCCTTTAT-CTTCCCAG
ashtoni      AATTACAGTGTTAAAAA--------------GAGAACTAGA CTATAATTACTATT---ACATAGAAATTAGAAG AGAATTCAATAAATATAA-GTATTGATTAATCCTTTAT-CTTCCCAG
bohemicus    AATTACAGTGTTAAAAA--------------GAGAACTAGA CTATAATTACTATT---ACATAGAAATTAGAAG AGAATTCAATAAATATAA-GTATTGATTAATCCTTTAT-CTTCCCAG
rupestris    AATTACGGTGTTAAAAA--------------GAGAACTAGA CTATAATTACTATT---ACATAGAAATTAGAAG AGAATTCAATAAATATAAGGTATTGATTAATCCTTTAT-CTTCCCAG
trifasciatus AATTACGATGATAAAAAA-------------GGGAACTAGA CTATAATTACTGTT---ACATAGAAATTAGAAG AGAATTCAATAAATATAAAGTATTGATTAATCCTTTAT-TTTCCCAG
ussurensis   AATTACGGTGATAAAAAA-------------GGGAACTAGA CTATAATTACTGTT---ACATAGAAATTAGAAG AGAATTCAATAAATATAAAGTATTGATTAATCCTTTAT-CTTCCCAG
borealis     AATTACGGTGTTAAAAA--------------GGGAACTAGA CTATAATTACTATTATTACATAGAAATTAGAAG AGAATTCAATAAATATAAAGTATTTATTAATACTTTAT-TTTCCCAG
melanurus    AATTACGGTGTTAAAAA--------------GGGAACTAGA CTATAATTACTATTATTACATAGAAATTAGAAG AGAATTCAATAAATATAAAGTATTGATTAATATTTTAT-TTTCCCAG
nevadensis   nnnnnnnnnGTTAAAAA--------------GGGAACTACA CTATAATTACTATT---ACACAGAAATTAGAAG AGAATTCAATAAATATAAAGTATTGATTAATCCTTTATTCTTTCCAG
mendax       AATTACTA-GATAAAAG-TAAGATGAGTAAA---AACTAGA CTATAATTACTATT---ACATTGAAATTAAGAG AGAATTCAATAAACATAAAGTATTGATTAATTCTTTATTCTTTCCAG
defector     AATTACTA-GATAAAAG-TAAGATGAGTAAA---AACTAGA CTATAATTACTATT---ACATTGAAATTAAGAG AGAATTCAATAAACATAAAGTATTGATTAATTCTTTATTCTTTCCAG

                  gap-MD (continued)
```

B

```
                             region-1                  region-2        region-3
                     ========================         ============     ==================
mixtus          GCAAG-TCTC----C-GAT---CGAA---TTTCATAGC----TGAA--TTTTTATTC--GTTTT---AC--ATGATGACAACATTATTCGCTCATTCGCGGTTTCCACCGGAGAATTTTC
parthenius      GCAAG-TCTC----C-AAT---CGAA---TTTCATAGC----TGAA--TTTCTATTC--GTTTT--AAC--ATGATGATAACATTATTCGCTCATTCGCGGTTTCCACCGGAGAATTTTC
bifarius        GCAAG-TCTC----C-AAT---CGAA---TTTCATAGC----TGAA--TTTTTATTT--GTTTT---AC--ATGATGATAACATTATTCGCTCATTCGCGGTTTCCACCGGAGAATTTTC
impatiens       GCAAG-TCTC----C-AAT---CGAA---TTTCATAGC----TGAA--TTTTTATTC--GTTTT---AC--ATGATGATAACATTATTCGCTCATTCGCGGTTTCCACCGGAGAATTTCC
crotchii        GCAAG-TTTC----C-GAT---CGAA---TTTCATAGC----TGAA--TTCTTATTT--GTTTT---AC--GTGATGATAACATTATTCGCTCATTCGCGGTTTCCACCGGAGAATTTTC
rufocinctus     GCAAG-TTTC----C-GAT---TGAA---CTTCATAGC----TGAA-TTTTTTATTT--GTCTT-TTAC--ATGATGATAATATTATTTGCTCATTCGCGGTTTCCACCGGAGAATTTTC
wurflenii       GCAAG-TTTCTTTTC-AAT---CGAA---TTTCATGGC----TGAA--TTTTTATTT--GTTTT---AC--ATGATGATAACATTATTCGCTCATTCGCGGTTTCCACCGGAGAATTTTC
lapidarius      GCAGG-TTTC----C-AAT---CGAA---TTTCATAGC----TGAATGTTTTTCTTT--GTTTT---AC--ATGATGATAACGTTATTCGCTCATTCGCGGTTTCCACCGGAGAATTTTC
formosellus     GCAGG-TTTC----C-AAT---CGAA---TTTCATAGC----TGAATTTTTTTATTT--GTTTC---AC--ATGATGATAACATTATTCGCTCATTCGCGGTTTCCACCGGAGAATTTTC
ruderarius      GCAAGTTTTT----C-GATTGTTGAA--TTTTCATAGCCGAA------TTTTTATTT--ATTTC---AA--ATGATGATAAAGTTATTTGCTCATTCGCGGTTTCCACCGGAGAATTTTC
mucidus         GCAAG-TTTC----C-TAT---TGAAGAATTTCATTGT----TGAA--TTTTTATTT--ATTTT---AA--ATGATGATAACGTTATTTTCTCATTCGCGGTTTCCACCGGAGAATTTTC
dahlbomii       GCAAG-TTTC----C-GAT---CGAA---TTTCATAGC----TGAA--TTTTTATTT--GTCACC--AA--ATGATGATAACGTTGTTTGCTCATTCGCGGTTTCCACCGGAGAATTTTC
pensylvanicus   GCAAG-TTTC----C-AGT---CGAA---TTTCATAGCCGAA------TTTTTATTT--ATTTC---AA--ATGATGATAACGTTATTTGCTCATTCGCGGTTTCCACCGGAGAATTTTC
ashtoni         GCAAG-TTTC----CCAAT---CGAA---TTTCATAGG----TGAA--TTTTTATTT--GTTTT---ACGTATGATGATAACATTATTTGCTCATTCGCGGTTTCCACCGGAGAATATTC
sushkini        GCAAG-TTTC----C-AAT---CGAA---TTTCATAGA----TGAA--CTTTTATTTTTGCTTT---AC--ATGATGATAACATTATTCGCTCATTCGCGGTTTCCACCGGAGAATTTTC
trifasciatus    GCAAG-TTTC----C-AGT---CGAA---TTTCACAGC----TGAA--CTTTTATTT--GCTTTTTTAC--ATGATGATAACGTTATTCGCTCATTCGCGGTTTCCACCGGAGAATTTTC
diversus        GCAAG-TTTC----C-AGT---CGAA---TTTCACAGC----TGAA--CTTTCATTT--GCTTTTTTAC--ATGATGATAACGTTATTCGCTCATTCGCGGTTTCCACCGGAGAATTTTC
haemorrhoidalis GCAAG-TTTC----C-AAT---CGAA---TTTCGTAGCCGAATGAACCCTTTTATCAT-GTTCT---AC--ATGATGACAACATTGTTCGCTCATTCGCGGTTTCCACCAGAGAATTTTC
melanurus       GCAAG-TTTC----C-AAT---CGAA--ATTTCATAGC----TGAA--TTTTTATTT--GTTTT---AA--ATGATGATAACATTATTCGCTCATTCGCGGTTTCCACCGGAGAATTTTC
nevadensis      GTAAG-TTTT----C-AAT---CGAA---TTTCACAGT----TCAA--TTTTTATTT--GTTTT---TC--ATGATGATAACATTATTCGCTCATTCGCGGTTTCCACCGGAGAATTTTC
```

```
                                               region-4a             region-4b
                                              ================== ==============
                                              =========  ======================                    region-5
                                              ===================     =========          ===========================
mixtus          TAATTTGATGACGGTTAGAGACGTCTGACTAAAATCT-AATCT--TTC-AAAAA-TTATTATTCCTGTCTTGAAA--GGC---A-TAT-GGTAA---TATTATTAAATATGATTGTAG
parthenius      TAATTTGATGACGGTTAGAGACGTCTGACT-AACCCC-GATCT--TTC-AAAAA-TTATTATTCCTGTCTTGAAA--GGC---A-TAT-GGTAA---TATTATTAAATATGATTGTAG
bifarius        TAATTTGATGACGGTTAGAGACGTCTGACT-AAATCT-AATCT--TTC-AAAAA-TTATT---CCTGTCTTGAAA--GGTGGCA-TAT-AGTAA---TATTATTAAATATGATTGTAG
impatiens       TAATTTGATGACGGTTAGAGACGTCTGACT-AAATCT-AATCT--TTC-AAAAA-TTATT---CCTGTCTTGAAA--GGCGGCA-TAT-GGTAA---TATTATTAAATATGATTGTAG
crotchii        TAATTTGATGACGGTTAGAGACGTCTGACTAGAAT-T-AATTT--TTC-AAAAA-TTATT---CCTATTTTGAAA--GGC---A-TAT-GGTAA--CTATTATTAAATATGATTGCAG
rufocinctus     TAATTTGATGACGGTTAGAGATGTCTGACTAAAAT-T-AATTCTTTTC-AAAAA-TTATT---CCTGTTTTGAAA--GGC---A-TAT-AGTAC---TATTATTAAATATGATTGTAG
wurflenii       TAATTTGATGACGGTTAGAGACGTCTGACTAAAAT-T-AATTT--TTC-AACAA-TTATT---CCTGTTTTGAAA--GGC---A-TAT-GGTAC---TATTATTAAATATGATTGTAG
lapidarius      TAATTTGATGACGGCTAGAGACGTCTGATTGAAAT-T-AATCT--TTC-GAAAAATTATT---CCCGCTTTGAAA--GGC---A-TAT-GGTAATACTATTATGAAATATGATTGTAG
formosellus     TAATTTGATGACGGCTAGAGACGTCTGACTAAAAT-T-AATTT--TTC-AAAAAATTATT---CCTGCTTTGAAA--GGC---A-TAT-GGTAC---TATTATAAAATATGATTGTAG
ruderarius      TAATTTGATGACGGTTAGAGACGACTGACTAAAAT-T-AATCT--TTT-AATAA-TTATT---CCTGTCTCGAAA--GGC---A-TATTGGTAA---TATTATTAATTGTGATTGTAG
mucidus         TAATTTGATGACGGTTAGAGACGACTGACTAAAAT-T-AATCT--TTC-AATAA-TTATA---CCTGTTTCGAAA--GGC---A-TATTGGTAA---TATTATTAGTTGTGATTGTAG
dahlbomii       TAATTTGATGACGGTTAGAGACGACTGACTAATGT-T-AATTT--TTC-AACGAATTATT---CCTGTTTCGAAA--GGC---A-TATTGGTAAT--TATTATCAATTGCGATTGTAG
pensylvanicus   TAATTTGATGACGGTTAGAGACGACTGACTAAAAT-T-AATTT--TTC-AATAA-TTATC---CCTGTTTCGAAA--GGC---A-TATTGGTAA---TATTATTAATTGTGATTGTAG
ashtoni         TAATTTGATGACGGTTAGAGACGACTGATTGAAAT-T-AATTTT-TTC-AATAA-TTATT---CCCGTTTTGAAA--GGC---G-TATTGTTAA---TATTGTTAAATATGATTGTAG
sushkini        TAATTTGATGACGGTTAGAGACGTCTGACTAAAAT-T-AATTT--TTCCAATAA-TTATC---CATGTTTTGAAA--GGC---A-CATTGTTAA---TATTATTAAATACGATTGTAG
trifasciatus    TAATTTGATGACGGTTAGAGACGTCTGACTAAAAT-T-AATTTT-TTC-GATAA-TTGTT---CCTGTCTTGAGAGAGGT---AATATTGGTAA---TACTATTAAATACGATTGTAG
diversus        TAATTTGATGACGGTTAGAGACGTCTGACTAAAAT-T-AATTTC-TTC-GATAA-TTATT---CCTGTCTTGAGA--GGC---G-TATTGGTAA---TGTTATTAAATACGATTGTAG
haemorrhoidalis CAATTTGATGACGGTTAGAGACGTCTGACTGAGAT-T-CATTT--TTC-AATAA-TTATC---CCCGTTTCGAAA--GGG--TAATATTGGTAA---TATTATTAAATATGATTGTAG
melanurus       TAATTTGATGACGGTTAGAGACGTCTGACTAAAAT-T-AATTT--TTC-AACAATTTATC---CCTGTCTTGAAA--GGC---A-TATTGGTAA---TATTATTGAACCTGATTGTAG
nevadensis      TAATTTGATGACGGTTAGAGACGTCTGACTAAAAT-TGAATTT--TTT-AAAAA-TTACT---CCTGCTTTGAAAA-AGC---A-TAT-GGTAA---TATTATTAAATATGATTGTAG
```

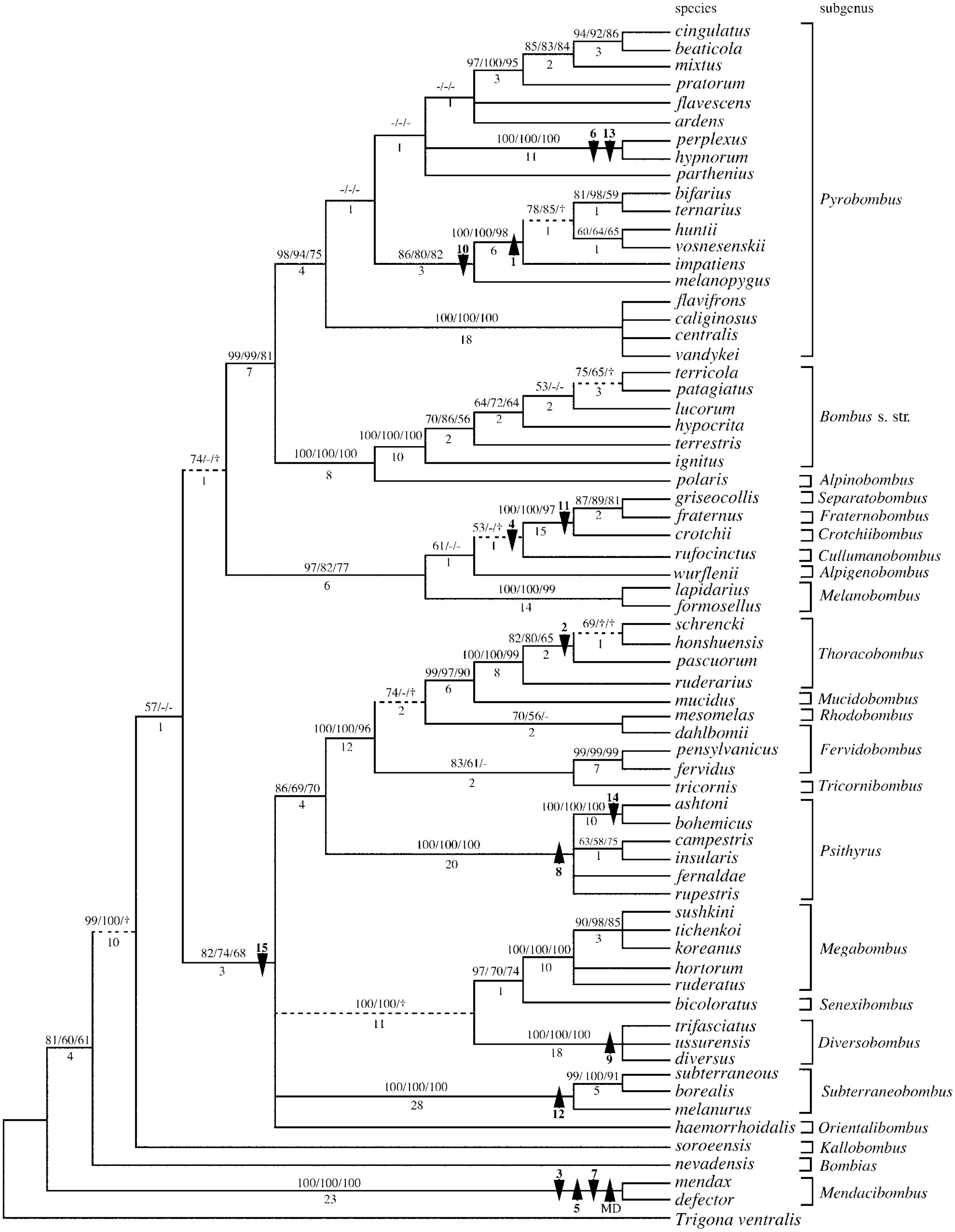

FIG. 2.—Strict consensus of 891 most-parsimonious (MP) trees based upon base substitutions of the unambiguously aligned regions. The 16 unambiguously aligned indels within ArgK intron were mapped onto specific branches of the tree without homoplasy. Upward and downward arrowheads represent insertions and deletions respectively. Numbers on arrowheads correspond to the gap codes in Fig. 1*A*. "MD" represents shaded sequences in Fig. 1*A*. Branches that collapse in the NJ and/or ML trees are presented as dotted lines. Nodal support is assessed by bootstrap values (above branches; MP/NJ/ML respectively; shown only when >50%) and branch support indices (below branches).

few exceptions, consistent with traditional subgeneric classifications of bumble bees based on adult morphological characters (Williams 1998). The topologies of NJ and ML trees are similar to that of MP tree (see supplementary online materials).

Overall, we inferred 16 parsimony-informative gaps within the ArgK intron (fig. 1*A*). All gaps were successfully assigned to specific branches of the MP tree without homoplasy (fig. 2), suggesting that the signal of gaps is concordant with that of base substitutions. This result holds for the NJ and ML trees with one exception. On the MP tree gap-4 was a synapomorphy uniting the species *B. griseocollis*, *B. fraternus*, *B. crotchii*, and *B. rufocinctus*, but monophyly of these four species was not recovered on the ML tree due to the sister relationship of *B. rufocinctus* to a species that lacks gap-4, *B. wurflenii* (fig. 2). Support in the ML analysis for paraphyly of the four species with gap-4 was ambiguous, as demonstrated by the extremely short lengths of the branches concerned. Therefore congruence of gap-4 with the phylogeny is not strongly rejected.

Analysis of the data matrix including the 16 characters coding ArgK gaps ("GAP" characters throughout) resulted in the same shortest trees (1354 steps: CI = 0.48; RI = 0.77; RC = 0.46) as did the data matrix including only base substitutions. Inclusion of GAP characters (16 of 411 informative characters) did not contribute to resolution, but bootstrap values for the nodes concerned increased from >53% to >72%, and the total branch support (Bremer 1994) increased from 323 to 335.

The six multistate characters coding ambiguous regions of EF-1α ("INAASE characters" throughout) mapped onto the substitution-based tree with a total of 187 steps and a CI of 0.77 (RI = 0.87; RC = 0.67). This indicates that the INAASE characters contain rich information consistent with the phylogenetic signal of base substitutions. Simultaneous analysis of an expanded data matrix, including the 16 GAP characters and the six INAASE characters resulted in four most-parsimonious trees of length 1535 (CI = 0.53; RI = 0.78; RC = 0.48) (supplementary online data) with lower homoplasy levels than the analysis based on substitutions and GAP characters. The six INAASE characters collectively had CI = 0.80, RI = 0.89, and RC = 0.71. Although the alignment ambiguous EF-1α intron regions accounted for only ~5% of all sites in the original data matrix, INAASE characters added 181 steps to the tree (12% of the total steps) and contributed substantially to phylogenetic resolution, resulting in a drastic decrease in the number of shortest trees (from 891 to 4) and recovery of 62 of 64 possible nodes in the strict consensus tree (an addition of 9 nodes). INAASE characters also significantly improved nodal support, increasing average bootstrap values by 3.9% (based on nodes in common on the strict consensus trees), and summed branch support by 55 (from 335 to 390).

Our analyses showed that gaps and ambiguously aligned regions of nuclear intron sequences contain useful phylogenetic signal concordant with that of base substitutions. Unambiguously aligned gaps exhibited minimal homoplasy and were consistently congruent with the substitution-based tree, whether derived from MP, NJ, or ML method. Thus, our results reinforce several earlier suggestions that gaps are phylogenetically reliable characters (Lloyd and Calder 1991; van Ham et al. 1994; Graham et al. 2000; Simmons, Ochoterena, and Carr 2001). For example, Lloyd and Calder (1991) found that all seven informative gaps longer than 1 bp within ψη-globin pseudogene of primates could be assigned to specific branches of substitution-based tree without homoplasy. Similarly, van Ham et al. (1994) showed that 13 of 15 informative gaps longer than 1 bp could be mapped on substitution-based *trnL-trnF* intergenic spacer gene tree in Crassulaceae plants without homoplasy. However, in these studies single-nucleotide indels were often homoplastic or ambiguously aligned. Multinucleotide indels are probably more reliable than single-nucleotide indels because the former are less frequent than single-nucleotide indels in many data sets (Saitou and Ueda 1994; van Ham et al. 1994; Graham et al. 2000) and because homoplasies by parallel and back mutations can occur only when they match exactly in length and position (and sequence for insertions) with the corresponding indels (Lloyd and Calder 1991).

On the other hand, the results of several other studies contrast with those described above. For example, in Bapteste and Philippe's (2002) analysis of eukaryotic phylogeny, the observed pattern of amino acid indels within enolase and impdh genes likely resulted from recombination and lateral gene transfer as well as from convergence and reversal. In other cases, apparently identical indels seem to have originated independently in distantly related taxa (van Ham et al. 1994; Philippe and Laurent 1998; Graham et al. 2000). However, these results do not necessarily indicate that indels are generally of lesser utility than base substitutions, because homoplasy can be found in any class of molecular characters, especially when dealing with phylogeny on a large scale (e.g., across phyla). Considering these results together with ours, we conclude that indels can be highly reliable characters, especially at lower taxonomic levels, but recognize that gaps, like all classes of phylogenetic characters, are not devoid of homoplasy. It may therefore be inadvisable to identify higher monophyletic groups based solely on a single indel (or a few indels) (Bapteste and Philippe 2002).

Recent methodological progress in handling gap characters in phylogenetic analyses affords the opportunity to incorporate this useful phylogenetic information derived from sequence data. However, these new methods are confined to the parsimony optimality criterion, and a well-justified, statistically-robust, generally applicable, and widely accepted method for incorporating indels within explicitly model-based phylogenetic methods such as standard implementations of ML is still lacking. The development of explicit models of indel evolution should not be particularly difficult, but selection of parameters to be included and statistical interpretation of resulting analyses will be far from straightforward (see Farris 1999; Sanderson and Kim 2000). Even within the context of parsimony analysis, the difficulty of assuming gap-to-substitution costs or weighting gaps of different lengths has impeded extensive use of the methods. However, gaps are strong indicators of common descent, and full utilization of indel information to corroborate and refine phylogenies inferred primarily from substitution data would surely improve the accuracy and efficiency of phylogeny estimation for many data sets (e.g., Rokas and Holland 2000; Danforth 2002).

Acknowledgment

Supported by a Grant-in-Aid from the Japan Society for Promotion of Science (no. 11304056).

Literature Cited

Bapteste, E., and H. Philippe. 2002. The potential value of indels as phylogenetic markers: position of trichomonads as a case study. Mol. Biol. Evol. **19**:972–977.

Bremer, K. 1994. Branch support and tree stability. Cladistics **10**:295–304.

Danforth, B. N. 2002. Evolution of sociality in a primitively eusocial lineage of bees. Proc. Natl. Acad. Sci. **99**:286–290.

Farris, J. S. 1999. Likelihood and inconsistency. Cladistics **15**:199–204.

Farris, J. S., M. Källersjö, A. G. Kluge, and C. Bult. 1994. Testing significance of incongruence. Cladistics **10**:315–320.

Giribet, G., and W. C. Wheeler. 1999. On gaps. Mol. Phylogenet. Evol. **13**:132–143.

Graham, S. W., P. A. Reeves, A. C. E. Burns, and R. G. Olmstead. 2000. Microstructural changes in noncoding chloroplast DNA: interpretation, evolution and utility of indels and inversions in basal angiosperm phylogenetic inference. Int. J. Plant Sci. **161**:S83–S96.

Jeanmougin, F., J. D. Thompson, M. Gouy, D. G. Higgins, and T. J. Gibson. 1998. Multiple sequence alignment with ClustalX. Trends Biochem. Sci. **23**:403–405.

Lloyd, D. G., and V. L. Calder. 1991. Multi-residue gaps, a class of molecular characters with exceptional reliability for phylogenetic analyses. J. Evol. Biol. **4**:9–21.

Lutzoni, F., P. Wagner, V. Reeb, and S. Zoller. 2000. Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. Syst. Biol. **49**:628–651.

Mardulyn, P., and S. A. Cameron. 1999. The major opsin in bees (Insecta: Hymenoptera): a promising nuclear gene for higher level phylogenetics. Mol. Phylogenet. Evol. **12**:168–176.

Philippe, H., and J. Laurent. 1998. How good are deep phylogenetic trees? Curr. Opin. Genet. Dev. **8**:616–623.

Rokas, A., and P. W. H. Holland. 2000. Rare genomic changes as a tool for phylogenetics. Trend Ecol. Evol. **15**:454–459.

Saitou, N., and S. Ueda. 1994. Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. Mol. Biol. Evol. **11**:504–512.

Sanchis, A., J. M. Michelena, A. Latorre, D. L. J. Quicke, U. Gardenfors, and R. Belshaw. 2001. The phylogenetic analysis of variable-length sequence data: elongation factor–1α introns in European populations of the parasitoid wasp genus Pauesia (Hymenoptera: Braconidae: Aphidiinae). Mol. Biol. Evol. **18**:1117–1131.

Sanderson, M. J., and J. Kim. 2000. Parametric phylogenetics? Syst. Biol. **49**:817–829.

Simmons, M. P., and H. Ochoterena. 2000. Gaps as characters in sequence-based phylogenetic analyses. Syst. Biol. **49**: 369–381.

Simmons, M. P., H. Ochoterena, and T. G. Carr. 2001. Incorporation, relative homoplasy, and effect of gap characters in sequence-based phylogenetic analyses. Syst. Biol. **50**:454–462.

Swofford, D. L. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4.0. Sinauer, Sunderland, Mass.

Swofford, D. L., G. J. Olsen, P. L. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pp. 407–514 *in* D. M. Hillis, C. Moritz, and B. K. Mabble, eds. Molecular systematics. Sinauer, Sunderland, Mass.

van Dijk, M. A., E. Paradis, F. Catzeflis, and W. W. de Jong. 1999. The virtues of gaps: Xenarthran (Edentate) monophyly supported by a unique deletion in αA-crystallin. Syst. Biol. **48**:94–106.

van Ham, C. H. J., H. t Hart, T. H. M. Mes, and J. M. Sandbrink. 1994. Molecular evolution of noncoding regions of the chloroplast genome in the Crassulaceae and related species. Curr. Genet. **25**:558–566.

Wheeler, W. C. 1996. Optimization alignment: the end of multiple sequence alignment in phylogenetics? Cladistics **12**:1–9.

———. 1999. Fixed character states and the optimization of molecular sequence data. Cladistics **15**:379–385.

Williams, P. H. 1998. An annotated checklist of bumble bees with an analysis of patterns of description (Hymenoptera: Apidae, Bombini). Bull. Br. Mus. Nat. Hist. (Ent.) **67**:79–152.

Ross Crozier, Associate Editor

Accepted September 9, 2002