Nonhomogeneous Model of Sequence Evolution Indicates Independent Origins of Primary Endosymbionts Within the Enterobacteriales (\(\gamma\)-Proteobacteria)

Joshua T. Herbeck,\(^1\) Patrick H. Degnan, and Jennifer J. Wernegreen

Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts

Standard methods of phylogenetic reconstruction are based on models that assume homogeneity of nucleotide composition among taxa. However, this assumption is often violated in biological data sets. In this study, we examine possible effects of nucleotide heterogeneity among lineages on the phylogenetic reconstruction of a bacterial group that spans a wide range of genomic nucleotide contents: obligately endosymbiotic bacteria and free-living or commensal species in the \(\gamma\)-Proteobacteria. We focus on AT-rich primary endosymbionts to better understand the origins of obligately intracellular lifestyles. Previous phylogenetic analyses of this bacterial group point to the importance of accounting for base compositional variation in estimating relationships, particularly between endosymbiotic and free-living taxa. Here, we develop an approach to compare susceptibility of various phylogenetic reconstruction methods to the effects of nucleotide heterogeneity. First, we identify candidate trees of \(\gamma\)-Proteobacteria groEL and 16S rRNA using approaches that assume homogeneous and stationary base composition, including Bayesian, maximum likelihood, parsimony, and distance methods. We then create permutations of the resulting candidate trees by varying the placement of the AT-rich endosymbiont Buchnera. These permutations are evaluated under the nonhomogeneous and nonstationary maximum likelihood model of Galtier and Gouy, which allows equilibrium base content to vary among examined lineages. Our results show that commonly used phylogenetic methods produce incongruent trees of the Enterobacteriales, and that the placement of Buchnera is especially unstable. However, under a nonhomogeneous model, various groEL and 16S rRNA phylogenies that separate Buchnera from other AT-rich endosymbionts (Blochmannia and Wigglesworthia) have consistently and significantly higher likelihood scores. Blochmannia and Wigglesworthia appear to have evolved from secondary endosymbionts, and represent an origin of primary endosymbiosis that is independent from Buchnera.

This application of a nonhomogeneous model offers a computationally feasible way to test specific phylogenetic hypotheses for taxa with heterogeneous and nonstationary base composition.

Introduction

Phylogenetic inference can be confounded by various evolutionary factors, including unequal substitution rates among sites (Yang 1993), unequal transition and transversion rates (Kimura 1980), and substitution saturation among excessively divergent taxa. In addition, most nucleotide substitution models assume homogeneity of nucleotide composition among taxa (Felsenstein 1988), although this assumption is easily violated in nonsimulated data sets. While it is known that variable base composition among taxa can distort estimations of substitution rates (Tourasse and Li 1999), the effects of variable base composition on phylogenetic analysis is not entirely clear (Mooers and Holmes 2000; Conant and Lewis 2001). Simulation studies (Conant and Lewis 2001; Rosenberg and Kumar 2003) suggest that nucleotide heterogeneity among taxa does not negatively affect parsimony, distance, and likelihood methods. Conant and Lewis (2001) found that only extreme nucleotide bias and long branch lengths can lead parsimony to incorrect phylogenetic inference. Yet, analyses of biological data sets have shown that parsimony (Loomis and Smith 1990; Steel, Lockhart, and Penny 1995), distance (Lockhart et al. 1994; Galtier and Gouy 1995), and maximum likelihood methods (Chang and Campbell 2000) can mistakenly group unrelated species with similar GC contents. Thus, base composition heterogeneity is considered a potential problem for phylogenetic reconstruction.

Several approaches have been proposed to account for variable base content among taxa. Hasegawa (Hasegawa and Hashimoto 1993; Hasegawa et al. 1993) found that nucleotide heterogeneity in 16S rRNA biased the phylogeny of deep diverging eukaryotes and suggested that amino acid sequences are more reliable. However, amino acid composition is also affected by nucleotide compositional bias (Foster, Jermiin, and Hickey 1997; Singer and Hickey 2000). Attempts to correct for this bias include LogDet (Lockhart et al. 1994), a distance method that transforms the substitution matrix to produce additive distances. The LogDet method does not consider rate variation among sites, and, similar to other distance methods, it performs poorly in analyses of taxa with moderate amounts of substitution saturation (Mooers and Holmes 2000). Galtier and Gouy (1995, 1998) have developed a nonhomogeneous Markov model of nucleotide substitution that allows equilibrium base composition to vary among lineages. This method modifies Tamura’s (1992) substitution model of unequal transition and transversion rates and unequal nucleotide content (GC and AT), to include variable substitution rates among sites and variable GC content among branches. This model has been used in a maximum likelihood framework to estimate ancestral GC content of thermophilic organisms (Galtier, Tourasse, and Gouy 1999), and to infer phylogenies of Drosophilidae taxa (Tarrio, Rodriguez-Trelles, and Ayala 2001) and weevil endosymbiotic bacteria (LeFèvre et al. 2004).

Such nonhomogeneous models may be particularly important in inferring phylogenies for bacteria, which

\(^1\) Present address: Department of Microbiology, University of Washington, Seattle, WA 98195.

Key words: nucleotide composition, Buchnera, insect endosymbionts, Enterobacteriales, phylogeny.

E-mail: herbeckt@u.washington.edu.

doi:10.1093/molbev/msi036
Advance Access publication November 3, 2004

Molecular Biology and Evolution vol. 22 no. 3 © Society for Molecular Biology and Evolution 2004; all rights reserved.
show an exceptionally wide range of base compositional biases (ranging from ~25% to 75% GC content; Sueoka 1962). Intracellular bacterial mutualists and pathogens have the lowest known genomic GC contents, and they represent several phylogenetically independent lineages (Moran and Wernegreen 2000; Shigenobu et al. 2000; Charles, Heddi, and Rabbe 2001; Moran 2002). Their AT-richness is most extreme at third codon positions and intergenic spacers, suggesting a strong effect of directional mutational pressure, or biased changes between GC and AT pairs (Muto and Osaka 1987; Sueoka 1988; Sueoka 1992). The Enterobacteriales includes free-living or gut-associated species (Escherichia coli, Salmonella typhimurium, Shigella flexneri, and Yersinia pestis) with moderate base compositions, as well as endosymbionts that form primary (obligate) and secondary (facultative, transient) associations with insects and that are relatively AT-biased (Moran and Telang 1998; Baumann, Moran, and Baumann 2000). The AT-bias of several primary endosymbionts within the Enterobacteriales is quite severe, at ~26% GC (Shigenobu et al. 2000) in the aphid endosymbiont Buchnera, ~22% GC in the tsetse fly endosymbiont Wigglesworthia (Akman et al. 2002), and 27% GC in the ant endosymbiont Blochmannia (Gil et al. 2003).

In addition to their extreme AT-bias, primary endosymbionts are characterized by severe genome reduction. Among Buchnera, Blochmannia, and Wigglesworthia, genome sizes range from 450 kb (Gil et al. 2002) to ~800 kb (Wernegreen, Lazarus, and Degnan 2002) compared to the 6 to 5.3 Mb for Escherichia coli genomes (Berghthorsson and Ochman 1995). Like other intracellular bacteria, these endosymbionts also experience fast rates of sequence evolution (Moran 1996; Woolfit and Bromham 2003), especially at nonsynonymous sites (Clark, Moran, and Baumann 1999; Wernegreen and Moran 1999), deleterious changes at the 16S rRNA gene (Lambert and Moran 1998), and AT-biased amino acid changes (Moran 1996; Clark, Moran, and Baumann 1999; Palacios and Wernegreen 2002; Herbeck, Wall, and Wernegreen 2003; Rispe et al. 2004). Mechanisms driving these shared features of endosymbiont genomes may include a combination of relaxed selection in the host intracellular environment, strong genetic drift resulting in part from the repeated vertical transmission through insect host generations and decreased effective population sizes (Moran 1996; Funk, Wernegreen, and Moran 2001; Abbot and Moran 2002; Mira and Moran 2002; Herbeck et al. 2003), and increased background mutation rates resulting from the loss of DNA repair loci during genome reduction (Mira, Ochman, and Moran 2001).

The phylogenies of the γ-Proteobacteria and Enterobacteriales specifically have received considerable attention, owing in part to their ecological importance, the medical relevance of several species, and the diverse lifestyles this group represents (Lawrence, Ochman, and Hartl 1991; Sproer et al. 1999; Wertz et al. 2003; Canbiäck, Tamas, and Andersson 2004). Of particular interest for comparative genomic studies is the relative position of Buchnera, Blochmannia, and Wigglesworthia, as the full genome sequences of these taxa are now available (Shigenobu et al. 2000; Akman et al. 2002; Tamas et al. 2002; Van Ham et al. 2003; Gil et al. 2003). However, the phylogenetic position of these and other AT-rich endosymbionts has proved difficult to recover and often varies among studies. Published phylogenies that include Buchnera, Wigglesworthia, and Blochmannia often group them as sister taxa or within a clade that includes only endosymbionts (e.g., Schröder et al. 1996; Heddi et al. 1998; Spaulding and von Dohlen 1998; Gil et al. 2003; Lerat, Daubin, and Moran 2003; Woolfit and Bromham 2003; Canbiäck, Tamas, and Andersson 2004). One exception is a phylogenetic study that considered heterogeneity in nucleotide biases in this group and that suggested a paraphyletic relationship among several primary endosymbionts, grouping Buchnera apart from Blochmannia and Wigglesworthia (Charles, Heddi, and Rabbe 2001). In addition, a tree estimated under a nonhomogeneous model (NJ-nh) in Lerat, Daubin, and Moran (2003; fig. 2) positioned Buchnera and Wigglesworthia in separate clades, prompting these authors to consider alternative topologies in which the two endosymbionts were not sister taxa. However, these candidate topologies were rejected as significantly less likely than topologies in which the two endosymbionts are sister taxa, based on an SH test (Shimodaira and Hasegawa 1999) implemented with a homogeneous model.

Our goal in this study is to build upon these previous analyses of the Enterobacteriales by further exploring possible effects of base-compositional biases, and by developing a new approach to accounting for such biases in phylogenetic analysis. We focus primarily on the relationships among select AT-rich primary endosymbionts, to better understand the origins of endosymbiosis and the evolution of their genomic features. Although previous studies have employed a nonhomogeneous model, we use a nonhomogeneous model to systematically evaluate possible candidate trees that were generated by homogeneous methods. First, we develop phylogenetic hypotheses with commonly used methods, including Bayesian inference, maximum likelihood, parsimony, and distance approaches. We then evaluate the resulting phylogenies and multiple permutations of them using the nonhomogeneous maximum likelihood model of Galtier and Gouy (1995).

Our results consistently support topologies that group Buchnera with gut-associated enteric species, separate from a clade that includes Wigglesworthia, Blochmannia, and many secondary (transient) endosymbionts of insects. Although the two genes considered here (groEL and 16S rRNA) produce slightly different topologies of the Enterobacteriales, the phylogenetic independence of Buchnera is well supported. We suggest the combination of homogeneous models to identify candidate trees and nonhomogeneous models to evaluate those topologies offers a robust method to test phylogenetic hypotheses for groups that vary dramatically in base composition.

Materials and Methods

Taxa and Sequences

We obtained sequences of γ-Proteobacteria (including Enterobacteriales and relatives) chaperonin groEL and 16S ribosomal RNA (rRNA) from GenBank. The groEL
data set included 23 γ-Proteobacteria taxa representing nine endosymbionts and 14 free-living bacteria (table 1). We aligned groEL using ClustalW (Chenna et al. 2003) based on amino acid translations of nucleotide sequences and subsequently excluded third positions due to saturation. The groEL alignment was 1,572 bp in total, and 1,048 bp for first and second codon positions alone. The 16S rRNA analysis included 32 taxa (table 2). We used the Ribosomal Database Project II Sequence Aligner (Cole et al. 2003) to align the 16S rRNA sequences, and edited all nucleotide sites and only variable sites in both groEL and 16S rRNA data sets. Both sequence alignments are available in the online resources.

Phylogenetic Analysis

Our approach was to estimate candidate starting trees using standard methods of phylogeny reconstruction that search tree space extensively. We then varied the placement of Buchnera across each starting tree and evaluated all possible permutations under a nonhomogeneous maximum likelihood model of sequence evolution (Galtier and Gouy 1995) (fig. 1).

Starting Phylogenies

We used six phylogenetic approaches to create alternative starting hypotheses of γ-Proteobacteria evolution for the groEL and 16S rRNA data sets: maximum parsimony (MP), GTR maximum likelihood (ML-gtr), Bayesian inference (MB), and Neighbor-Joining under three models of sequence evolution: the Kimura 2-parameter (NJ-k2p), the LogDet transformation, and the nonhomogeneous model T92+Γ+varGC (NJ-nh). The free-living Pseudomonas aeruginosa was designated the root taxon for all trees, after the initial phylogenetic analyses were completed.

Maximum likelihood analyses of both groEL and 16S rRNA were performed in PAUP* 4.0b10 (Swofford 2002) based on a general time reversible (GTR) model with a Γ distribution and a portion of invariable sites estimated from the data, as selected using ModelTest version 3.06 (Posada and Crandall 1998) by the Akaite Information Criterion. Bootstrap support was calculated using 500 replicates, each using 10 random taxon addition replicates with tree bisection and reconnection (TBR) branch swapping. All maximum likelihood heuristic searches and bootstrap analyses were implemented in parallel on a Beowulf cluster utilizing the clusterpaup program (A. G. McArthur, jbpc.mbl.edu/computing). We conducted Bayesian analyses with Mr. Bayes v.3.0 (Ronquist and Huelsenbeck 2003) using the GTR substitution model with base frequencies and substitution rate matrix estimated from the data. For groEL analysis we used 3 million Markov chain Monte Carlo (MCMC) generations with four parallel chains (one cold, three heated), with trees sampled every 50 generations and a 10,000 tree burn-in period. For 16S rRNA analysis, we included 10 million MCMC generations of four chains, with trees sampled every 50 generations, with a 50,000 tree burn-in. Heuristic parsimony analyses included TBR branch-swapping and all characters unordered and equally weighted.

We obtained a NJ-nh topology from a distance matrix estimated using Galtier and Gouy’s (1995) T92+Γ+varGC nonhomogeneous model, implemented in the software Phylowin (Galtier, Gouy, and Gautier 1996). Using PAUP* 4.0b10 (Swofford 2002), we also estimated NJ trees based on the LogDet transformation (Lockhart et al. 1994), a method designed to be robust to variable mutational tendencies among lineages, and the Kimura 2-parameter model (NJ-k2p). Bootstrap support was calculated using 500 replicates for all distance methods.

Permutation of Starting Trees by Varying Position of Buchnera

The five homogeneous analyses and NJ-nh generated alternative phylogenetic hypotheses that we used as starting trees. To focus on the relationship between Buchnera and
other AT-rich primary endosymbionts, we created permutations of these starting trees for all possible placements of the Buchnera clade (lineages of Buchnera from the following aphids: A. pisum, B. pistaciae, and S. graminum). We focused on Buchnera because it is a relatively well-characterized primary endosymbiont, and its placement was the most labile in the starting phylogenies. These permutations were produced in PAUP* by “generating all trees” constrained to the best tree for a given reconstruction method, altered so that the Buchnera clade was a basal polytomy (and thus free to vary in location across the constraint tree). All possible placements of Buchnera generated 234 candidate trees for the groEL data set (39 possible placements of Buchnera on the six starting trees) and 342 candidate trees for the 16S data set (57 possible placements of Buchnera on each of six starting trees).

Evaluation of Alternative Phylogenies under a Nonhomogeneous Model of Nucleotide Substitution

For each starting tree and all permutations of those trees, we used the eval_nh program from the NHML 3.0 package (Galtier, Tourasse, and Gouy 1999) to evaluate the likelihood under Galtier and Gouy’s (1995) T92+Γ+varGC nonhomogeneous model (ML-nh). We then ranked all topologies based on the resulting likelihood scores. We noted the placement of Buchnera relative to the other (AT-rich) primary endosymbionts, particularly Blochmannia and Wigglesworthia, and categorized trees as “polyphyletic” (if Buchnera grouped apart from Blochmannia and Wigglesworthia) or “monophyletic” (if Buchnera grouped with Blochmannia and Wigglesworthia).

Essentially, this approach was an approximate tree-searching method to create phylogenetic hypotheses of interest for subsequent evaluation under the nonhomogeneous model. We chose this approach rather than the tree-searching algorithm available in the NHML program (star_nh) for three reasons: (1) tree-searching under star_nh was prohibitively computationally intensive, given the parameter-rich model and the number of taxa we considered; (2) initial phylogenies based on MP, ML-gtr, Bayesian, and NJ approaches, and permutations of those trees, offered a more comprehensive and stable set of possible topologies than did the branch-swapping method.
implemented by star_nh (unpublished data); (3) because our primary interest was the evolution of AT-rich primary endosymbionts within the Enterobacteriales, we wished to consider all possible placements of Buchnera across reasonable starting trees.

We evaluated the statistical significance of differences in likelihood values among competing phylogenies with Kishino and Hasegawa’s (1989) resampling estimated log likelihood (RELL) procedure implemented by PAUP* 4.0b10. This test entails the bootstrapping of site-specific log-likelihood values estimated under a particular substitution model, in this case Galtier and Gouy’s (1995) nonhomogeneous model (ML-nh). We also performed Mann-Whitney U-tests to identify significant differences in \(-\ln(L)_{nh}\) distributions between trees in which primary endosymbionts are polyphyletic versus monophyletic (as defined above). This allowed us to test whether trees that position Buchnera apart from other endosymbionts have consistently and significantly better likelihood scores than trees that group primary endosymbionts together. We tested the suitability of the ML-nh model for the groEL and 16S rRNA data sets using likelihood ratio tests (Felsenstein 1981) on nested substitution models, using maximum parsimony topologies.

Results

Nucleotide Composition

For both groEL and 16S rRNA, base composition at first and second codon positions is significantly heterogeneous among taxa (although only at variable sites in 16S rRNA) (table 1, table 2). As expected, endosymbiotic bacteria are relatively AT-rich compared to the majority of free-living bacteria. Although the chi-squared test of homogeneity used here ignores potential phylogenetic structure and violates the assumption of independent comparisons, the result suggests that phylogenetic analysis should account for variable GC content among taxa. Consistent with this observed heterogeneity, likelihood ratio tests (table 3) show that ML-nh is a significantly better fit than other evolutionary models for both groEL and 16S rRNA.

Phylogeny of groEL

The inferred phylogenies of groEL (codon positions 1 and 2) vary considerably among the reconstruction methods (fig. 2). The MB and NJ-nh topologies position Buchnera apart from other endosymbionts and as a sister

![Diagram](https://academic.oup.com/mbe/article-abstract/22/3/520/1075915/1075915)
Table 4
Top Log Likelihood Scores, under the Galtier and Gouy Maximum Likelihood Model of Nonhomogeneous Nucleotide Substitution (ML-nh), of 234 Possible groEL Phylogenies

<table>
<thead>
<tr>
<th>Tree Rank</th>
<th>Method for Backbone</th>
<th>ln(L)nh</th>
<th>Buchnera Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>MB</td>
<td>−5139.1474</td>
<td>Polyphyletic</td>
</tr>
<tr>
<td>2.</td>
<td>MB</td>
<td>−5140.9703</td>
<td>Polyphyletic</td>
</tr>
<tr>
<td>3.</td>
<td>MB</td>
<td>−5141.0289</td>
<td>Polyphyletic</td>
</tr>
<tr>
<td>4.</td>
<td>MB</td>
<td>−5142.7799</td>
<td>Polyphyletic</td>
</tr>
<tr>
<td>5.</td>
<td>MB</td>
<td>−5143.1503</td>
<td>Polyphyletic</td>
</tr>
<tr>
<td>6.</td>
<td>MB</td>
<td>−5144.1806</td>
<td>Polyphyletic</td>
</tr>
<tr>
<td>7.</td>
<td>MB</td>
<td>−5145.4126</td>
<td>Polyphyletic</td>
</tr>
<tr>
<td>8.</td>
<td>NJ-nh</td>
<td>−5145.6253</td>
<td>Polyphyletic</td>
</tr>
<tr>
<td>9.</td>
<td>MB</td>
<td>−5145.6942</td>
<td>Polyphyletic</td>
</tr>
<tr>
<td>10.</td>
<td>MB</td>
<td>−5146.0777</td>
<td>Polyphyletic</td>
</tr>
<tr>
<td>11.</td>
<td>MB</td>
<td>−5139.1474</td>
<td>Polyphyletic</td>
</tr>
<tr>
<td>12.</td>
<td>MB</td>
<td>−5148.5264</td>
<td>Polyphyletic</td>
</tr>
<tr>
<td>13.</td>
<td>NJ-nh</td>
<td>−5150.4120</td>
<td>Polyphyletic</td>
</tr>
<tr>
<td>14.</td>
<td>MP</td>
<td>−5152.9485</td>
<td>Polyphyletic</td>
</tr>
<tr>
<td>15.</td>
<td>NJ-nh</td>
<td>−5154.7372</td>
<td>Polyphyletic</td>
</tr>
<tr>
<td>16.</td>
<td>NJ-nh</td>
<td>−5155.5988</td>
<td>Polyphyletic</td>
</tr>
<tr>
<td>17.</td>
<td>ML-gtr</td>
<td>−5157.7527</td>
<td>Polyphyletic</td>
</tr>
<tr>
<td>18.</td>
<td>MP</td>
<td>−5158.3373</td>
<td>Polyphyletic</td>
</tr>
<tr>
<td>19.</td>
<td>ML-gtr</td>
<td>−5159.3627</td>
<td>Polyphyletic</td>
</tr>
<tr>
<td>20.</td>
<td>NJ-nh</td>
<td>−5160.1027</td>
<td>Polyphyletic</td>
</tr>
</tbody>
</table>

NOTE.—Also shown is the particular phylogenetic method used to infer the backbone tree, across which the Buchnera clade was placed. The placement of Buchnera is described relative to the Blochmannia-Wigglesworthia clade as polyphyletic (Buchnera is not placed with the Blochmannia-Wigglesworthia clade) or monophyletic for the endosymbionts of interest (Buchnera is placed with the Blochmannia-Wigglesworthia clade). MB = Bayesian analysis, using Mr. Bayes; ML-gtr = maximum likelihood with GTR+I+1 substitution model; MP = maximum parsimony; NJ-nh = neighbor joining with distance matrix estimated using the T92+I+varGC model.

to the enteric bacteria. In addition, consistent with most published phylogenies (e.g., Spaulding and von Dohlen 1998) the MB and NJ-nh phylogenies position the whitefly endosymbiont as a basal lineage. By contrast, the MP and ML-gtr trees group all AT-rich endosymbiont together.

We used eval nh to calculate −ln(L)nh of the data set across 234 candidate groEL phylogenies (table 4). The original MB topology had the highest −ln(L)nh, and permutations of this tree had generally high scores compared to permutations of other starting trees. Permutations of the MP, ML-gtr, and NJ-nh trees have nearly equal distributions of −ln(L)nh scores. The starting trees under LogDet and NJ-k2p, and all permutations of these, have very low −ln(L)nh scores (all ≤5180) (see figures in the Supplementary Material online).

We then compared −ln(L)nh scores of trees developed under a given method, noting the placement of Buchnera as polyphyletic or monophyletic relative to the Blochmannia and Wigglesworthia clade. This comparison determined whether various placements of Buchnera on the starting topologies significantly improved −ln(L)nh. Separation of Buchnera from Blochmannia and Wigglesworthia improved −ln(L)nh scores for MP, ML-gtr, and NJ-nh relative to the original trees (which group all endosymbionts together). Notably, the tree with the best −ln(L)nh (the original MB tree) is polyphyletic. We also explored whether alternative placements of Buchnera had distinct likelihood distributions, regardless of the underlying topology. For MB, MP, ML-gtr, and NJ-nh, the distribution of −ln(L)nh for monophyletic permutations was typically lower than those of polyphyletic permutations of the same starting tree, and it was significantly different for MB and NJ-nh (Mann-Whitney U-test, P < 0.001*). This indicates that, across various starting topologies, trees that position Buchnera apart from other primary endosymbionts have better likelihood scores under the nonhomogeneous model than do trees that group primary endosymbionts together.

Of the 234 groEL topologies considered, those with the top 10 −ln(L)nh scores included nine trees based on
Phylogeny of Enterobacterales with a Nonhomogeneous Model

Bayesian

- mealybug secondary endosymbiont A crah
- mealybug secondary endosymbiont M albiz
- mealybug secondary endosymbiont A grev
- *Blanchninia C pennsylvanica*
- *Blanchninia C floridanus*
- *Buchnera A pinnac*
- *Buchnera S gramin*
- *Buchnera B pistia*
- *Wigleswirthia G brevipalpis*
- mealybug secondary endosymbiont P noth
- mealybug secondary endosymbiont P fusc
- psyllid secondary endosymbiont D nebr
- psyllid secondary endosymbiont B cockerelli
- psyllid secondary endosymbiont P floccose
- *Sodalis glossinidius*
- weevil endosymbiont S organe
- *Brenneria quercina*
- *Citrobacter freundii*
- *Escherichia coli*
- *Shigella flexneri*
- *Salmonella typhimurium*
- *Klebsiella pneumoniae*
- *Proteus vulgaris*
- aphid secondary endosymbiont T type
- *Yersinia pestis*
- Vibrio cholera
- *Haemophilus influenzae*
- Pasturella multocida
- "Pseudomonas aeruginosa"

Maximum Likelihood (ML-gtr)

- mealybug secondary endosymbiont A crah
- mealybug secondary endosymbiont M albiz
- mealybug secondary endosymbiont A grev
- *Blanchninia C pennsylvanica*
- *Blanchninia C floridanus*
- *Buchnera A pinnac*
- *Buchnera S gramin*
- *Buchnera B pistia*
- *Wigleswirthia G brevipalpis*
- mealybug secondary endosymbiont P noth
- mealybug secondary endosymbiont D nebr
- psyllid secondary endosymbiont B cockerelli
- psyllid secondary endosymbiont P floccose
- *Sodalis glossinidius*
- weevil endosymbiont S organe
- *Brenneria quercina*
- *Citrobacter freundii*
- *Escherichia coli*
- *Shigella flexneri*
- *Salmonella typhimurium*
- *Klebsiella pneumoniae*
- *Proteus vulgaris*
- aphid secondary endosymbiont T type
- *Yersinia pestis*
- Vibrio cholera
- *Haemophilus influenzae*
- Pasturella multocida
- "Pseudomonas aeruginosa"

Parsimony

- mealybug secondary endosymbiont A crah
- mealybug secondary endosymbiont M albiz
- mealybug secondary endosymbiont A grev
- *Blanchninia C pennsylvanica*
- *Blanchninia C floridanus*
- *Buchnera A pinnac*
- *Buchnera S gramin*
- *Buchnera B pistia*
- *Wigleswirthia G brevipalpis*
- mealybug secondary endosymbiont P noth
- mealybug secondary endosymbiont D nebr
- psyllid secondary endosymbiont B cockerelli
- psyllid secondary endosymbiont P floccose
- *Sodalis glossinidius*
- weevil endosymbiont S organe
- *Brenneria quercina*
- *Citrobacter freundii*
- *Escherichia coli*
- *Shigella flexneri*
- *Salmonella typhimurium*
- *Klebsiella pneumoniae*
- *Proteus vulgaris*
- aphid secondary endosymbiont T type
- *Yersinia pestis*
- Vibrio cholera
- *Haemophilus influenzae*
- Pasturella multocida
- "Pseudomonas aeruginosa"

Neighbor-joining, T92 + G + varGC (NJ-nh)
the initial MB phylogeny and one permutation of the NJ-nh topology. Resampling estimated log likelihood analysis shows no statistical support for the single top tree over the next nine topologies, but any one of the top 10 topologies is statistically supported over every 10th tree of the top 100 sampled. (The 10th tree is not significantly better than the 11th, but is significantly better than the 20th, 30th, etc.). The majority of topologies with the greatest nonhomogeneous likelihood place Buchnera separate from Blochmannia and Wigglesworthia, including 24 of the best-scoring 25 and 47 of the best-scoring 50 topologies. The majority-rule consensus of the top 10 trees under the NH model places the Buchnera clade sister to the free-living Erwinia herbicola, while Wigglesworthia and Blochmannia group with the Sitophilus oryzae primary endosymbiont and Sodalis (fig. 3). Consensus of groEL trees with the top 50 

\[-\ln(L)_{\text{nh}}\] scores does not change the placement of Buchnera relative to Blochmannia and Wigglesworthia, and it differs only in the level of resolution given to the placement of Buchnera within the free-living enteric clade (see figure 3 for consensus of top 10 and top 25 trees).

Phylogeny of 16S rRNA

The 16S rRNA analysis shows many of the same trends as groEL. First, the starting topologies estimated under the five homogeneous methods and NJ-nh vary considerably, and methods differ in their placement of Buchnera with or apart from Wigglesworthia and Blochmannia (fig. 4). Interestingly, some methods that group primary endosymbionts together at groEL (e.g., MP, ML-gtr), show polyphylectic relationships at 16S rRNA, and vice versa (e.g., MB). For 16S rRNA, the MB and ML-gtr topologies are identical except for the placement of Buchnera. The endosymbiont of Bemisia tabaci is basal in all 16S rRNA phylogenies, as expected.

Among the various permutations of these starting trees, those based on NJ trees have the lowest 

\[-\ln(L)_{\text{nh}}\] scores (all permutations \leq 9075). Even NJ based on a nonhomogeneous model (NJ-nh) had lower 

\[-\ln(L)_{\text{nh}}\] scores than did trees based on MB, ML-gtr, or MP. New positions of the Buchnera clade generally improve the 

\[-\ln(L)_{\text{nh}}\] scores of the starting tree, with the exception of the ML-gtr tree, for which the original tree has the greatest 

\[-\ln(L)_{\text{nh}}\] and, notably, Buchnera is polyphylectic. The original MB tree is significantly less likely in RELL analysis than the original ML-gtr tree because it groups Buchnera with other endosymbionts. For all methods, separation of Buchnera from other endosymbionts improves 

\[-\ln(L)_{\text{nh}}\] scores as compared to trees that group them together (see figures in the Supplementary Material online). This improvement is reflected in the significant difference in the 

\[-\ln(L)_{\text{nh}}\] distributions between polyphylectic and monophyletic trees (Mann-Whitney U-tests, \(P < 0.001^*\) for each of MB, ML-gtr, MP, and NH-nh).

The top 10 16S rRNA topologies under nonhomogeneous models place Buchnera polyphyletic relative to the other endosymbionts, and classify it as sister to the E. coli group (table 5). As for groEL, the RELL method does not distinguish among the top 10 16S trees, but any one of the top 10 trees is significantly more likely than any of every 10th lower-ranked topology (as described above). The majority-rule consensus of the 10 16S rRNA trees with the top 

\[-\ln(L)_{\text{nh}}\] scores separates Buchnera from Blochmannia, Wigglesworthia, and the other endosymbionts (fig. 5). Also similar to the groEL analysis, the majority of trees with greatest 

\[-\ln(L)_{\text{nh}}\] scores separate Buchnera from the other endosymbionts. These include 25 of the highest 25 trees and 45 of the highest 50.

We also analyzed a 1,280-bp, 30-taxon, 16S rRNA data set that differed from the 32-taxon data set by lacking two endosymbionts and replacing one free-living taxon. Because the results from this 30-taxon data set only corroborate the results and interpretation of the 32-taxon data set described above, we therefore have placed this information in an online supplement.

Discussion

We have focused on the evolutionary relationships of AT-rich primary endosymbionts, using groEL and 16S rRNA genes and a nonhomogeneous substitution model that accounts for variable AT-bias among taxa. Our goal in this study was to develop a approach that takes advantage of (1) the tree-searching algorithms implemented by several standard phylogenetic methods, and (2) the utility of nonhomogeneous models to evaluate relationships among taxa with widely different base compositions. We first demonstrated that variable base composition strongly affects estimates of relationships among free-living and endosymbiotic Enterobacteriales, as the likelihood ratio tests show ML-nh is the best fit to both groEL and 16S rRNA data sets. Our main result is that a nonhomogeneous likelihood model supports the separation of Buchnera from Blochmannia and Wigglesworthia. Three lines of evidence presented here support that conclusion. First, homogeneous models gave incongruent topologies and sometimes placed Buchnera, Blochmannia, and Wigglesworthia as monophyletic, as found in previous studies. However, the nonhomogeneous likelihood scores \((-\ln(L)_{\text{nh}}\) of such “monophyletic” trees was always improved in a permutation that separated Buchnera from Blochmannia and Wigglesworthia. Second, across all standard phylogenetic methods, various permutations that separate Buchnera had overall better \(-\ln(L)_{\text{nh}}\) an improvement that was often statistically significant. Third, for both 16S rRNA and groEL, the consensus trees of phylogenies with the best 

\[-\ln(L)_{\text{nh}}\] scores separate Buchnera from a clade that includes Wigglesworthia, Blochmannia, and other insect endosymbionts.

Discussion.

We have focused on the evolutionary relationships of AT-rich primary endosymbionts, using groEL and 16S rRNA genes and a nonhomogeneous substitution model that accounts for variable AT-bias among taxa. Our goal in this study was to develop a approach that takes advantage of (1) the tree-searching algorithms implemented by several standard phylogenetic methods, and (2) the utility of nonhomogeneous models to evaluate relationships among taxa with widely different base compositions. We first demonstrated that variable base composition strongly affects estimates of relationships among free-living and endosymbiotic Enterobacteriales, as the likelihood ratio tests show ML-nh is the best fit to both groEL and 16S rRNA data sets. Our main result is that a nonhomogeneous likelihood model supports the separation of Buchnera from Blochmannia and Wigglesworthia. Three lines of evidence presented here support that conclusion. First, homogeneous models gave incongruent topologies and sometimes placed Buchnera, Blochmannia, and Wigglesworthia as monophyletic, as found in previous studies. However, the nonhomogeneous likelihood scores \((-\ln(L)_{\text{nh}}\) of such “monophyletic” trees was always improved in a permutation that separated Buchnera from Blochmannia and Wigglesworthia. Second, across all standard phylogenetic methods, various permutations that separate Buchnera had overall better \(-\ln(L)_{\text{nh}}\) an improvement that was often statistically significant. Third, for both 16S rRNA and groEL, the consensus trees of phylogenies with the best 

\[-\ln(L)_{\text{nh}}\] scores separate Buchnera from a clade that includes Wigglesworthia, Blochmannia, and other insect endosymbionts.

Discussion.

We have focused on the evolutionary relationships of AT-rich primary endosymbionts, using groEL and 16S rRNA genes and a nonhomogeneous substitution model that accounts for variable AT-bias among taxa. Our goal in this study was to develop a approach that takes advantage of (1) the tree-searching algorithms implemented by several standard phylogenetic methods, and (2) the utility of nonhomogeneous models to evaluate relationships among taxa with widely different base compositions. We first demonstrated that variable base composition strongly affects estimates of relationships among free-living and endosymbiotic Enterobacteriales, as the likelihood ratio tests show ML-nh is the best fit to both groEL and 16S rRNA data sets. Our main result is that a nonhomogeneous likelihood model supports the separation of Buchnera from Blochmannia and Wigglesworthia. Three lines of evidence presented here support that conclusion. First, homogeneous models gave incongruent topologies and sometimes placed Buchnera, Blochmannia, and Wigglesworthia as monophyletic, as found in previous studies. However, the nonhomogeneous likelihood scores \((-\ln(L)_{\text{nh}}\) of such “monophyletic” trees was always improved in a permutation that separated Buchnera from Blochmannia and Wigglesworthia. Second, across all standard phylogenetic methods, various permutations that separate Buchnera had overall better \(-\ln(L)_{\text{nh}}\) an improvement that was often statistically significant. Third, for both 16S rRNA and groEL, the consensus trees of phylogenies with the best 

\[-\ln(L)_{\text{nh}}\] scores separate Buchnera from a clade that includes Wigglesworthia, Blochmannia, and other insect endosymbionts.
Bayesian analysis, using Mr. Bayes; ML-gtr $= 100$. Varying placements of original phylogenies (figs. 2 and 4) is also reflected by the Gil et al. 2003; Canbäck, Tamas, and Andersson 2004). (Schroeder et al. 1996; Charles, Heddi, and Rahbe 2001; the Blochmannia taxon sets. The lability of placement of consensus trees show slight differences (e.g., in the ML-gtr (for groEL the original phylogenies estimated by MB (for 16S) and trees under the nonhomogeneous model are most similar to data sets with base composition variation. The consensus standard phylogenetic methods may be most robust for Buchnera backbone tree. The placement of Buchnera is described relative to the Blochmannia-Wiggleworthia clade as polyphyletic (Buchnera is not placed with the Blochmannia-Wiggleworthia clade) or monophyletic for the endosymbionts of interest (Buchnera is placed within the Blochmannia-Wiggleworthia clade). MB = Bayesian analysis, using Mr. Bayes; ML-gtr = maximum likelihood with GTR + $I + I$ substitution model; MP = maximum parsimony.

The results of this study also shed light on which standard phylogenetic methods may be most robust for data sets with base composition variation. The consensus trees under the nonhomogeneous model are most similar to the original phylogenies estimated by MB (for 16S) and ML-gtr (for groEL). This observed robustness of MB and ML-gtr is consistent with a recent simulation study showing that a maximum-likelihood approach is best able to handle heterogeneity (Rosenberg and Kumar 2003). Interestingly, NJ-nh separated Buchnera from Blochmannia and Wiggleworthia at both groEL and 16S rRNA, suggesting that this method successfully accounts for base compositional variation. However, other aspects of this tree were apparently quite poor, as the NJ-nh tree (and permutations thereof) had lower $- \ln(L)_{nh}$ than phylogenies based on most other methods. This result suggests that implementation of a nonhomogeneous model should not be limited to only an NJ approach.

Comparison with Previously Published Phylogenies

Our results do not provide a single best Enterobacteriales phylogeny, as the groEL and 16S rRNA consensus trees show slight differences (e.g., in the placement of Erwinia spp.) that may reflect the different taxon sets. The lability of Buchnera observed across our original phylogenies (figs. 2 and 4) is also reflected by the varying placements of Buchnera across published trees (Schröder et al. 1996; Charles, Heddi, and Rahbe 2001; Gil et al. 2003; Canbäck, Tamas, and Andersson 2004).

Despite the variable position of Buchnera in our starting trees, evaluation of these placements (and many more, in permutations of these trees) under the nonhomogeneous models supports the grouping of Buchnera with a clade of enteric bacteria, apart from an endosymbiont clade that includes Blochmannia and Wiggleworthia.

The fact that our results conflict with those of several previous studies warrants a comparison of the phylogenetic models and taxon sampling employed. First, previous studies suggesting monophyly of Buchnera, Wiggleworthia, and/or Blochmannia often employ homogeneous models. Similarly, we also found that the best trees of some homogeneous models group these endosymbionts closely together (e.g., ML-gtr and MP for 16S rRNA, and MP and MB for groEL). Previous studies that implemented nonhomogeneous models have typically used NJ-nh, and they either produced trees that separated Buchnera and Wiggleworthia but were statistically rejected by other methods (Lerat, Daubin, and Moran 2003), or they produced a tree in which these two endosymbionts were sister taxa (Canbäck, Tamas, and Andersson 2004). In contrast, our use of nonhomogeneous model is likelihood-based rather than exclusively NJ-based.

Second, several previous analyses have taken advantage of the numerous loci available in full genome sequences, but generally they included fewer taxa in the analysis (16 or fewer). The inclusion of numerous genes or entire genomes has obvious benefits for studies of...
Implications for the Evolution of Endosymbiosis

Our results have two implications for the evolution of endosymbiosis within the Enterobacteriales. First, *Blochmannia* and *Wigglesworthia* apparently represent an origin of primary endosymbiosis that is independent from *Buchnera*. Genome sequence data may also support this independent transition to primary endosymbiosis, as the three fully sequenced endosymbiont genomes share only \( \sim 50\% \) of their genes, or \( \sim 70\% \) for any pairwise comparisons between genera (Gil et al. 2003). Because transition to primary endosymbiosis is thought to impose immediate, severe genome reduction through large genome deletion events, such endosymbionts may rapidly become constrained to their particular host association (Moran and Mira 2001; Van Ham et al. 2003). This genome reduction may impose severe constraints on extracellular existence, and it may limit switching among hosts with different nutritional physiologies. The phylogenies presented here cannot distinguish whether *Blochmannia* and *Wigglesworthia* acquired the primary endosymbiotic lifestyle independently of each other. However, given that these two genomes share just \( \sim 70\% \) of their genes and are highly specialized to the nutritional physiology of their respective hosts, independent acquisitions of obligate endosymbiosis is the more likely possibility.

Second, *Blochmannia* and *Wigglesworthia* are part of a diverse clade consisting of secondary endosymbionts of insects, suggesting that primary endosymbionts may evolve from secondaries. Koga, Tsuchida, and Fukatsu (2003) provide experimental evidence that secondary endosymbionts may move into the symbiotic niche of primaries. Specifically, they showed that a facultative endosymbiotic \( \gamma \)-Proteobacterium infected the cytoplasm of bacteriocytes of aphid hosts from which *Buchnera* had been eliminated, and that it thus compensated for the essential roles of *Buchnera*. A second example of the potential for primary endosymbionts to evolve from secondary endosymbionts is that of *Sodalis glossinidius* and the *Sitophilus oryzae* primary endosymbiont. The close phylogenetic association of this secondary endosymbiont of tsetse flies and this obligate mutualist of weevils shown here (figs. 3 and 5) is further corroborated by their shared maintenance and expression of a type III secretion system, which likely was acquired prior to their divergence (Dale et al. 2002). Although these endosymbionts are associated with distinct insect hosts, *Sodalis* is still capable of horizontal transmission (Aksoy, Chen, and Hypsa 1997; Dale and Maudlin 1999). Understanding specific routes by which diverse endosymbioses are established will require more extensive taxon sampling of facultative and primary endosymbiont lineages; however, the current study suggests that primary endosymbiosis has originated more often than previously thought, and it may represent the end of an evolutionary spectrum between the facultative and obligate intracellular lifestyles.

Acknowledgments

The authors thank N. Galtier and M. Gouy for providing the NML program for implementation of nonhomogeneous models. We are grateful to F. Rodriguez-Trelles for invaluable assistance applying these models and sharing modified source codes. We thank A.G. McArthur for providing clusterpaup and other phylogenetic analysis programs; D. M. Hillis for helpful discussion of taxon sampling and related issues; and two anonymous reviewers for useful comments on an earlier version of this manuscript. This work was supported by grants to J.J.W. from the National Institutes of Health (NIH R01 GM62626–01), the National Science Foundation (DEB 0089455), and the Josephine Bay Paul and C. Michael Paul Foundation.

Literature Cited


Manolo Gouy, Associate Editor

Accepted October 4, 2004