Power Analysis of Tests for Loss of Selective Constraint in Cave Crayfish and Nonphotosynthetic Plant Lineages

Jim Leebens-Mack† and Claude dePamphilis†*

*Department of Biology, Colgate University, New York; and †Department of Biology and Institute of Molecular Evolutionary Genetics, Penn State University

Loss of selective constraint on a gene may be expected following changes in the environment or life history that render its function unnecessary. The long-term persistence of protein-coding genes after the loss of known functional necessity can occur by chance or because of selective maintenance of an unknown gene function. The selective maintenance of an alternative gene function is not demonstrated by the failure of statistical tests to reject the hypothesis that there has been no change in the degree of constraint on the evolution of coding genes. Maintenance may be inferred, however, when power analyses of such tests demonstrate that there has been a sufficient number of nucleotide substitutions to detect the loss of selective constraint. Here, we describe a power analysis for tests of loss of constraint on protein-coding genes. The power analysis was applied to loss-of-constraint tests for opsin gene evolution in cave-dwelling crayfish and *rbc*L evolution in nonphotosynthetic parasitic plants. The power of previously applied tests for loss of constraint on cave crayfish opsin genes was insufficient to distinguish between chance retention and selective maintenance of opsin genes. However, the power of codon-based likelihood ratio tests for change in d_N/d_S ($=\omega$) (nonsynonymous to synonymous change) did have sufficient power to detect a loss of constraint on *rbc*L associated with a loss of photosynthesis in most examples but failed to detect such a change in three independent lineages. We conclude that *rbc*L has been selectively maintained in these holoparasitic plant lineages. This conclusion suggests that either these taxa are photosynthetic for at least a part of their life or *rbc*L may have an unknown function in these plants unrelated to photosynthesis.

Introduction

The long-term impact of natural selection acting over many generations is evident in the nucleotide substitution patterns of protein-coding genes. Purifying selection acts to reduce the frequency of amino acid replacements caused by nonsynonymous nucleotide substitutions. As a result, the observed rate of nonsynonymous substitutions per nonsynonymous site (d_N) is typically much lower than the per site rate of synonymous nucleotide substitution (d_S) (e.g., Li, Wu, and Luo 1985; Yang and Nielsen 1998). The long-term strength of purifying selection is measured as the ratio of nonsynonymous to synonymous change ($d_N/d_S = \omega$). For example, in an analysis of 48 mammalian nuclear genes, Yang and Nielsen (1998) estimated rate ratios ranging from 0.017, for a highly conserved ATP synthase gene, to 0.838, for a weakly constrained interleukin gene. In contrast, genes exposed to diversifying selection are expected to have $d_N/d_S > 1.0$ (e.g., Swanson and Vacquier 1995; Messier and Stewart 1997; Yang et al. 2000; Barrier, Robichaux, and Purugganan 2001). These expectations allow one to predict how changes in the selective regime will be reflected in patterns of nucleotide substitutions in phylogenetic analyses of protein-coding genes. Such predictions form the basis of statistical tests for hypothesized variation in selection among lineages in a phylogeny or among coding domains in a gene (reviewed in Yang and Bielawski 2000).

Key words: molecular evolution, statistical power, likelihood ratio test, *rbc*L, opsin, rhodopsin, crayfish, *Orobanchaceae*, *Scrophulariaceae*.

Address for correspondence and reprints: Jim Leebens-Mack, Department of Biology and Life Sciences Consortium, 518 Mueller Laboratory, Penn State University, University Park, Pennsylvania 16802. E-mail: jleebensmack@psu.edu.

Mol. Biol. Evol. 19(8):1292–1302. 2002
© 2002 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

For protein-coding genes that have evolved under purifying selection, the rate of nonsynonymous substitutions and d_N/d_S are expected to increase when a gene is converted to a pseudogene or when changes in the environment, life history, or developmental program render the gene's function obsolete. Obsolete genes may persist as open reading frames for long periods of time because of chance alone (Marshall, Raff EC, and Raff RA 1994). However, before the full-length open reading frame is interrupted by a stop codon, the cessation of purifying selection is expected to become evident in the pattern of nucleotide substitutions. Genes or pseudogenes that are not under the influence of purifying selection should on an average have d_N/d_S rate ratios ($=\omega$) equal to unity. This a priori expectation provides a basis to test for the loss of selectively maintained gene function (Crandall and Hillis 1997; Nei, Zhang, and Yokoyama 1997).

The null hypothesis for such comparative tests is that ω is the same in hypothetically constrained and unconstrained lineages. As in any formal hypothesis test, simply failing to reject the null hypothesis does not prove its veracity. The test may not have sufficient power to detect a change in ω. Given the narrowly defined alternative hypothesis in tests for the loss of selective constraint ($\omega = d_N/d_S = 1$), it is relatively easy to simulate sequence evolution according to the alternative hypothesis and estimate the type-II error rate, β. The type-II error rate is the probability of failing to reject the null hypothesis when the alternative is true. In cases where β is very small, and yet the null hypothesis is not rejected, one must seriously consider the possibility that the null hypothesis is correct. In the context presented here, when a change in the environment, life history, or developmental program that has rendered known gene function obsolete is not associated with a change in ω,

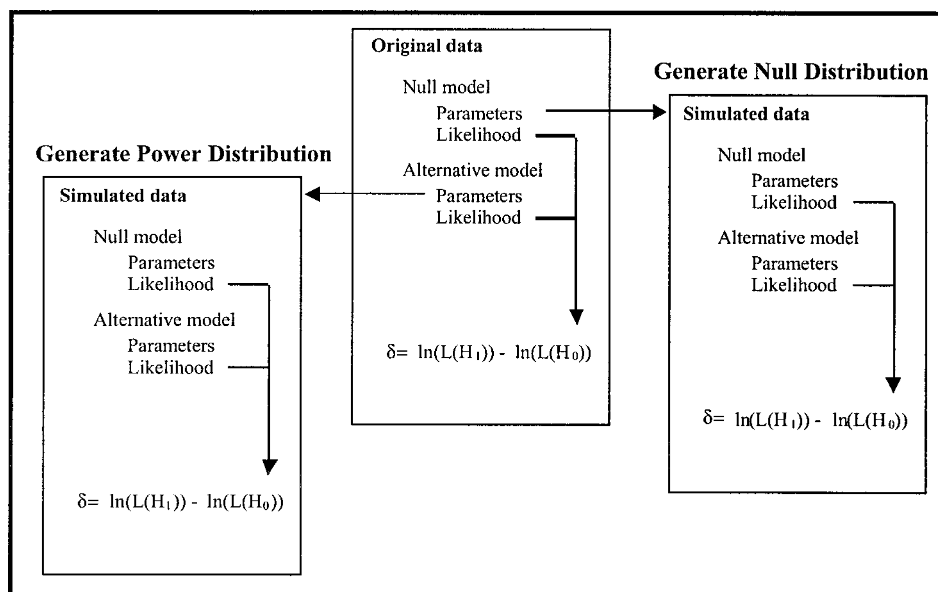

FIG. 1.—Flow diagram for generating test statistic distributions according to the null model for hypothesis testing and according to the alternative hypothesis for power analysis (after Huelsenbeck and Ranala 1997).

selective maintenance of some unknown gene function may be implicated.

Here, we provide a strategy for performing such a power analysis and apply it to the analyses of a photosynthetic gene, *rbc*L, in nonphotosynthetic parasitic plants (Wolfe and dePamphilis 1997, 1998) and a photoreceptor gene coding for an opsin in cave-dwelling crayfish (Crandall and Hillis 1997). In both examples the failure to detect a change in the selective constraint has raised the question of an alternative function for these genes.

Materials and Methods

Phylogenetically based tests for change in tempo or mode of molecular evolution most typically involve comparisons between related lineages. At a minimum, three taxa are required in order to estimate branch lengths. In cases where more than three taxa are included in the analysis, it is best to infer topological relationships using data other than the gene sequences under consideration. In the analyses described here, we chose this approach to circumscribe subsets of taxa that include (1) a clade or lineage where the focal gene is hypothetically obsolete, (2) a sister lineage, and (3) an out-group. For each subset, we obtained maximum likelihood (ML) estimates of the parameters of Goldman and Yang's codon-based model of gene sequence evolution (Goldman and Yang 1994; Yang and Bielawski 2000) and ancestral nucleotide sequences (Yang, Kumar, and Nei 1995). We then tested the null hypothesis that there was no variation in ω among lineages. In cases where we were unable to reject the null hypothesis, we simulated sequence evolution under the loss of functional constraint hypothesis using estimated ancestral sequences, branch lengths, and rate parameters (fig. 1). The simulated data sets were used to perform formal power tests for each analysis.

rbcL Analysis

Parasitic plants have served as model systems for understanding how genes and genomes evolve in new adaptive zones (dePamphilis and Palmer 1990; Wolfe, Morden, and Palmer 1992; dePamphilis 1995). Recent phylogenetic investigations have shown a single origin of parasitism among the parasitic Orobanchaceae-Scrophulariaceae, followed by multiple independent losses of photosynthesis associated with holoparasitism within this group (dePamphilis, Young, and Wolfe 1997; Wolfe and dePamphilis 1998; Young, Steiner, and dePamphilis 1999; Olmstead et al. 2001). This situation offers an ideal opportunity to compare the molecular evolution of plastid genes in distinct nonphotosynthetic lineages (dePamphilis 1995; dePamphilis, Young, and Wolfe 1997; Young and dePamphilis 2000). We performed the following analysis in order to interpret the interesting observation that *rbc*L, the plastid gene coding for the large subunit of RUBISCO, has been retained in many of these distinct nonphotosynthetic lineages.

Sequence Data

Three plastid genes from eight photosynthetic and seven nonphotosynthetic plant species were included in the *rbc*L analysis. The *rbc*L genes sampled from the nonphotosynthetic species were intact, with open reading frames and no apparent amino acid substitutions, relative to photosynthetic seed plants in the region coding for the active site (Kellogg and Juliano 1997; Wolfe and dePamphilis 1998). A combined phylogenetic analysis of plastid-encoded intron maturase (*mat*K) and ribosomal protein (*rps*2) was performed to establish the topological relationships of the species included in the *rbc*L analysis. Neither of these genes is involved in photosynthesis. Most of the *mat*K (Young, Steiner, and dePamphilis 1999), *rps*2 (dePamphilis, Young, and

Wolfe 1997), and *rbc*L (Wolfe and dePamphilis 1997, 1998) sequences were collected for previous studies (table 1). Additional *mat*K sequences were collected using a Beckman CEQ2000 automated sequencer following the manufacture's protocols as modified by Barkman et al. (2000). One-quarter volume sequencing reactions and postsequencing cleanup were performed in 96-well microtiter plates. The methodological details for PCR amplification were as reported by Young, Steiner, and dePamphilis (1999). PCR products were sequenced directly following Qiagen PCR column cleanup. All sequences are available on GenBank (table 1), and the sequence alignments for all three genes have been placed in TreeBASE (http://www.treebase.org/treebase/). The lengths of aligned sequences were 1,320 bp (440 codons) for *rbc*L, 1,632 bp for *mat*K, and 616 bp for *rps*2.

Phylogenetic Analysis

An ML analysis of the combined *rps*2 and *mat*K data set was performed in PAUP* (Swofford 2000) using the Hasegawa-Kashino-Yano model (HKY) of nucleotide evolution (Hasegawa, Kashino, and Yano 1985) with variation in the rates of change across sites modeled as an approximated gamma distribution (Yang 1994). Bootstrap probabilities for each node on the ML tree were estimated after resampling the combined data matrix 100 times (Felsenstein 1985). Bootstrap support values for each clade were also estimated in Neighbor-Joining and maximum parsimony analyses using 500 bootstrap replicates. The HKY plus gamma model described for the ML analysis was also used to estimate the Neighbor-Joining bootstrap support values.

The resulting ML phylogeny was used to identify five pairs of photosynthetic and nonphotosynthetic lineages and an out-group for each pair (figs. 2 and 3). Separate tests for change in constraint on *rbc*L genes in the nonphotosynthetic lineage were performed for each of the five subsets.

Estimation of rbcL *Evolution*

The evolution of *rbc*L was modeled using using the codon-based model of sequence evolution developed by Goldman and Yang (1994) and implemented in Yang's CODEML program (Yang 1999). The probability of nucleotide substitution along a branch of length t is estimated as $P(t) = e^{Qt}$, where the elements of the matrix Q are q_{ij}, the instantaneous substitution rate from codon i to codon j ($i \neq j$):

0 if j is a stop codon
0 if i and j differ at more than a single position
π_j for synonymous transversions
$\kappa\pi_j$ for synonymous transitions
$\omega\pi_j$ for nonsynonymous transversions
$\omega\kappa\pi_j$ for nonsynonymous transitions.

For each of the five comparisons, the estimated parameters included the transition-transversion rate ratio (κ), the per site nonsynonymous-synonymous rate ratio, ω, and the expected equilibrium frequencies for all 61 amino acid–coding codons, π_j. Codon frequencies were estimated from the observed nucleotide frequencies at each of the three codon positions. Ancestral sequences were also estimated in CODEML following the procedure described by Yang, Kumar, and Nei (1995). The probability of stop codons arising, p_{stop}, in the absence of selective maintenance was approximated by allowing $q_{ij} \neq 0$ when j is a stop codon and then summing $p_{ij}(t)$ across the inferred ancestral sequence for all three stop codons. Note that this approximation assumes that ancestral sequences were estimated without error. The probability of stop codons arising in nonphotosynthetic clades, $p_{stop}(\text{clade})$, for *Harveya* and *Orobanche* was approximated as:

$$[1 - p_{stop}(clade)] = \prod_{branches} [1 - p_{stop}(branch)]$$

A likelihood ratio test was used to assess change in ω in nonphotosynthetic plant lineages (Yang 1998; Yang and Nielsen 1998). A simple model, where the rate ratio ω did not vary among branches, was compared with a slightly more complicated model that estimated one-rate ratio for branches leading to photosynthetic taxa (ω_p) and another rate ratio for branches leading to nonphotosynthetic taxa (ω_n). The likelihood ratio was calculated as the difference in log likelihoods: $\delta = \ln L(\text{two-rate ratios}) - \ln L(\text{one-rate ratio})$.

For each of the five analyses, a null distribution for the likelihood ratio was generated using separate Monte-Carlo simulations. The sequence-generating simulation program is available upon request. For each analysis, 100 data sets were simulated using the basal ancestral sequence, model parameters, and branch lengths estimated from the original data sets under the single-rate ratio model (fig. 1). Following the procedures of Goldman (1993, Huelsenbeck and Ranala 1997), the likelihood ratio was calculated for each simulated data set. *P*-values were estimated from the simulated null distributions and compared with *P*-values obtained by assuming that two times the nested likelihood ratio approximates a chi-square distribution for long sequences evolving under the null model (Yang and Bielawski 2000).

In order to assess the power of each test, 100 additional data sets were simulated according to the two-rate ratio model (fig. 1). The likelihood ratio statistic, δ, was calculated for each simulated data set, and the distribution of the resulting δ's was compared with the null distribution. The probability of failing to reject a false null hypothesis, β, was estimated as the proportion of the power distribution with δ less than the upper 5% of the null distribution.

Opsin Analysis

Others have taken alternative approaches for assessing variation among evolutionary lineages in the strength of purifying selection acting on protein-coding genes (Crandall and Hillis 1997; Nei, Zhang, and Yokoyama 1997). Here, we use one example to demonstrate how the Monte-Carlo simulation approach described for the likelihood ratio test can be used for any

Table 1
Source of Sequences for Photosynthetic and Nonphotosynthetic Plant Species

Taxon	Photosynthesis	*rps*2	*mat*K	*rbc*L
Nicotiana tabacum	Y	Shinozaki et al. 1986 (CHNTXX)	Shinozaki et al. 1986 (CHNTXX)	Shinozaki et al. 1986 (CHNTXX)
Lindenbergia philippinensis	Y	Young, Steiner, and dePamphilis 1999 (AFO55163)	Young, Steiner, and dePamphilis 1999 (AFO51990)	Olmstead et al. 2001 (AF123664)
Pedicularis foliosa	Y	dePamphilis, Young, and Wolfe 1997 (PFU4874)	Present study (AF489959)	Wolfe and dePamphilis 1998 (AFO26836)
Castilleja lineariifolia	Y	dePamphilis, Young, and Wolfe 1997 (U48739)	Young, Steiner, and dePamphilis 1999 (AFO51981)	Wolfe and dePamphilis 1998 (AFO26823)
Lathraea clandestina	N	dePamphilis, Young, and Wolfe 1997 (U48755)	Young, Steiner, and dePamphilis 1999 (AFO51989)	Wolfe and dePamphilis 1998 (AFO26833)
Melampyrum sylvaticum[a] and *M. lineare*[b]	Y	Young, Steiner, and dePamphilis 1999[a] (AFO55148)	Young, Steiner, and dePamphilis 1999[a] (AFO51991)	Wolfe and dePamphilis 1998[b] (AFO26834)
Orobanche corymbosa	N	dePamphilis, Young, and Wolfe 1997 (U48760)	Young, Steiner, and dePamphilis 1999 (AFO51993)	Wolfe and dePamphilis 1997 (U73969)
Orobanche fasciculate	N	Young, Steiner, and dePamphilis 1999 (AFO55143)	Young, Steiner, and dePamphilis 1999 (AFO51994)	Wolfe and dePamphilis 1997 (U73970)
Alectra orobanchoides	N	dePamphilis, Young, and Wolfe 1997 (U48741)	Present study (AF489960)	Wolfe and dePamphilis 1998 (AFO26819)
Alectra sessiliflora	Y	dePamphilis, Young, and Wolfe 1997 (U48742)	Young, Steiner, and dePamphilis 1999 (AFO51977)	Wolfe and dePamphilis 1998 (AFO26820)
Harveya capensis	N	Young, Steiner, and dePamphilis 1999 (AFO55142)	Present study (AF489961)	Wolfe and dePamphilis 1998 (AFO26829)
Harveya purpurea	N	dePamphilis, Young, and Wolfe 1997 (U48749)	Young, Steiner, and dePamphilis 1999 (AFO51984)	Wolfe and dePamphilis 1998 (AFO26830)
Buchnera floridana[a] and *B. hispida*[b]	Y	dePamphilis, Young, and Wolfe 1997 (U48744)[a]	Present study[b] (AF489962)	Wolfe and dePamphilis 1998 (AFO26822)[a]
Striga asiatica	Y	dePamphilis, Young, and Wolfe 1997 (U48746)	Young, Steiner, and dePamphilis 1999 (AFO52000)	Wolfe and dePamphilis 1998 (AFO26838)
Striga gesnerioides	N	dePamphilis, Young, and Wolfe 1997 (U48747)	Present study (AF489963)	Wolfe and dePamphilis 1998 (AFO26839)

[a,b] Superscripts designate the species from which each gene was sequenced.

loss-of-constraint test performed on protein-coding genes.

Opsins are photosensitive pigments that play a central role in color vision for vertebrates (Yokoyama S and Yokoyama R 1996) and invertebrates (Briscoe and Chittka 2001). Opsins have also been implicated as playing a role in the entrainment of circadian rhythms in vertebrates, including subterranean blind mole rats (*Spalax ehrenbergi*) (Argamaso et al. 1995; David-Gray et al. 1999; Janssen et al. 2000). Crandall and Hillis (1997) tested for loss of selective constraint on opsin genes sampled from three independent cave-dwelling crayfish species: *Cambarus hubrichti*, *Orconectes australis*, and *Procambarus orcinus*. Separate comparative analyses were performed to assess rhodopsin gene evolution in each cave-dwelling lineage relative to a surface-dwelling sister species. Analogous to our *rbc*L investigation, three-taxon data sets included a cave-dwelling species, a surface-dwelling congener, and an outgroup taxon (Crandall and Hillis 1997). Ancestral nucleotide sequences were estimated using CODEML (Yang 1999), and nonsynonymous and synonymous substitutions were inferred for each pair of related cave- and surface-dwelling crayfish lineages. A 2×2 contingency table analysis was then used to test for an association between habitat and the type of inferred nucleotide substitution (nonsynonymous vs. synonymous). When the contingency table analyses and Kolmogorov-Smirnov tests for differences in nucleotide and amino acid substitution matrices all failed to detect a change in the cave-dwelling lineages, Crandall and Hillis (1997) hypothesized that rhodopsin may have a function other than photoreception in lightless cave environments.

In order to assess the power of the contingency table analysis performed by Crandall and Hillis, we simulated opsin gene evolution taking the same approach described for the *rbc*L power analysis. For each three-taxon analysis specified in the original study, sequences

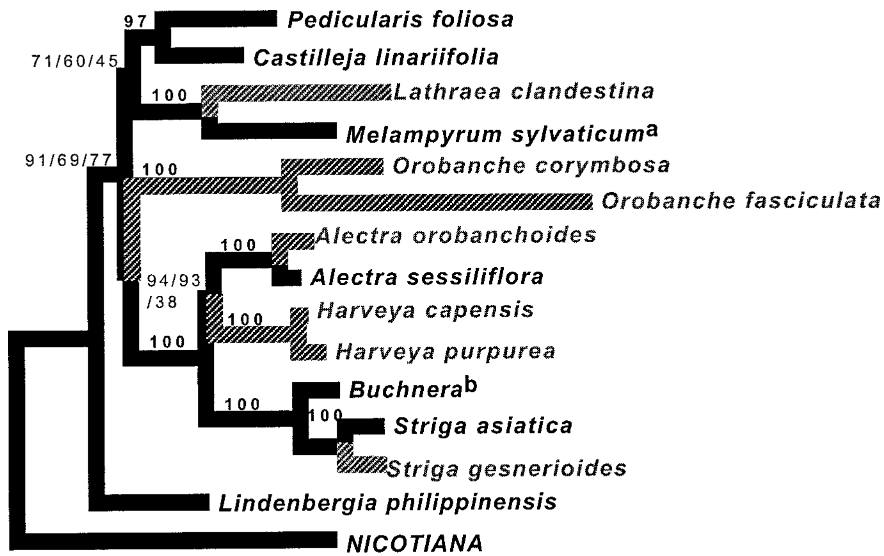

FIG. 2.—The ML phylogram estimated from a combined data set of *mat*K and *rps*2 sequences shows five independent losses of photosynthesis. Bootstrap values (ML-NJ-MP) are shown above each node. One value is shown for the cases where all analyses gave the same result.

were simulated with the per site nonsynonymous-synonymous rate ratio set at $\omega_{cave} = 1.0$ for the cave-dwelling species. Branch lengths, rate ratios for branches leading to the surface-dwelling crayfish, $\omega_{surface}$, and ancestral sequence were estimated in CODEML (Yang 1999).

Following the procedure of Crandall and Hillis, the number of nonsynonymous and synonymous substitutions along branches leading to the cave- and surface-dwelling species were inferred and used to test for independence between habitat and substitution type for all the 100 simulated data sets. The distribution of *P*-values

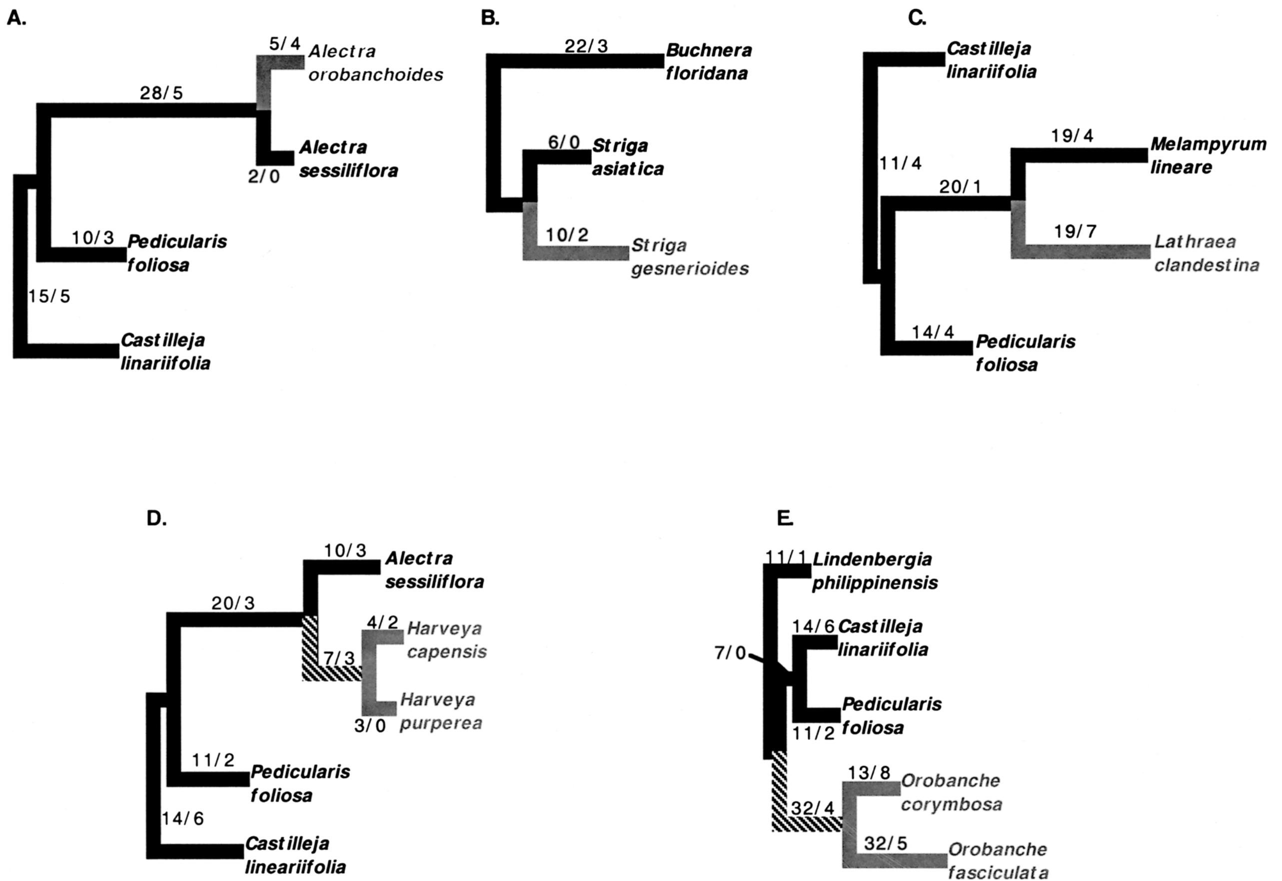

FIG. 3.—The ML *rbc*L-based phylograms for the five separate tests for loss of constraint with loss of photosynthesis. The inferred number of substitutions (nucleotide–amino acid) assuming a single-rate ratio, ω, is shown above each branch.

Table 2
Likelihoods, Nonsynonymous-Synonymous Rate Ratios (ω), and Test Statistics for Assessing Loss of Constraint on *rbc*L Genes in Nonphotosynthetic Holoparasites

Nonphotosynthetic Taxon	Probability of Stop Codon	One-Rate ln(L)	One-Rate ω	Two-Rate ln(L)	Two-Rate ω_p, ω_n	δ (P, $P\chi^2$)	β
Alectra orobanchoides	0.1149	2,116.11351	0.1432	2,113.94263	0.1194, 1.0817	**2.17088 (0.05, 0.037)**	0.77
Striga gesnerioides	0.1697	1,960.77034	0.0455	1,960.64188	0.0383, 0.0682	0.12846 (0.74, 0.61)	0.057
Lathraea clandestina	0.3481	2,230.52375	0.0991	2,228.61045	0.0727, 0.2326	**1.91330 (0.04, 0.05)**	<0.01
Harveya							
Early loss	0.2788	2,180.75529	0.1306	2,180.37383	0.1151, 0.2039	0.38146 (0.40, 0.38)	0.15
Late loss	0.1457	2,180.75529	0.1306	2,180.53859	0.1231, 0.2117	0.2167 (0.57, 0.51)	0.27
Harveya capensis[a]	0.0852	2,180.75529	0.1306	2,179.05989	0.1150, 0.8605	**1.69540 (0.035, 0.065)**	—
Orobanche							
Early loss	0.7831	2,413.90739	0.0816	2,413.87481	0.0756, 0.0851	0.032574 (0.85, 0.80)	≪0.01
Late loss	0.6140	2,413.90739	0.0816	2,413.24938	0.0658, 0.1103	0.65801 (0.25, 0.25)	≪0.01
Orobanche corymbosa[a]	0.2296	2,413.90739	0.0816	2,408.82077	0.06, 0.4941	**5.08661 (<0.01, 0.001)**	—

NOTE:—The probability of stop codons arising in *rbc*L is estimated on branches between maximum likelihood reconstructions of ancestral sequences (Yang, Kumar, and Nei 1995) and holoparasitic terminal taxa. Cases where the one-rate ratio model was rejected are in boldface.

[a] Tests for change in constraint on branches leading to *H. capensis* and *O. corymbosa* were a posteriori. The reported *P*-values are therefore illegitimate and included only for heuristic purposes.

obtained for Fisher's exact test was calculated for each analysis. Whereas the original paper reported the probability of the inferred contingency table values for nonsynonymouns and synonymous substitutions in cave- and surface-dwelling crayfish under the null hypothesis, we calculated *P*-values for the more appropriate one-tailed test. The *P*-values for the one-tailed test are influenced not only by the probability of the observed contingency table values but also by the probabilities of all other contingency tables that would suggest a stronger association while maintaining equal row and column totals. The power of each analysis, β, was estimated in two ways. First, β was estimated as the frequency of simulated data sets, where $P_{\text{one-tailed}} > 0.05$ (fig. 1). As a comparison, β was also estimated using a more standard power analysis for Fisher's exact test (Cassagrande, Pike, and Smith 1978) implemented in DSTPLAN (Brown et al. 2000).

Results and Discussion

Holoparasite *rbcL* Analysis

The history of plastid diversification among the sampled taxa was well resolved in the *rps*2-*mat*K phylogeny (fig. 2). The generally high bootstrap values allowed for unambiguous circumscription of five subsets that included a nonphotosynthetic lineage, closely related photosynthetic lineage, and an out-group lineage (fig. 3). The nonphotosynthetic lineage in each subset represents an independent origin of holoparasitism (fig. 2, dePamphilis, Young, and Wolfe 1997; Wolfe and dePamphilis 1998; Young, Steiner, and dePamphilis 1999). Independent tests for loss of selective constraint on *rbc*L were performed on each subset as described previously.

For two of the five subsets, the likelihood ratio test resulted in the rejection of the null hypothesis that ω was equal in the nonphotosynthetic and photosynthetic lineages (table 2). Note that for all the likelihood ratio tests, the *P*-values derived from our simulation–based null distribution were very similar to those based on the chi-square null distribution (table 2). The rate ratios differed significantly between *Alectra orobanchoides*, a holoparasite, and *A. sessiliflora*, a photosynthetic hemiparasite (table 2). In this example, the power of the likelihood ratio test was suggested by the significant difference detected in the d_N/d_S rate ratio along a branch with only five inferred nucleotide substitutions (fig. 3). However, the probability of failing to reject the null hypothesis when $\omega_n = 1.0$ on a phylogeny with similar ancestral sequence, branch lengths, and transition bias was quite high ($\beta = 0.77$, fig. 4*A*).

The test for loss of constraint was substantially more powerful in the other four analyses (table 2, figs. 4 and 5), and a second significant change in ω was found on the branch leading to *Lathraea clandestine* (table 2). In contrast, no change in constraint on *rbc*L sequence evolution was detected in the lineages leading to *Striga gesnerioides*, *Harveya* or *Orobanche* (table 2). The results of power analyses suggested that we should have detected a change in ω, assuming that these lineages lost photosynthetic function immediately after each had diverged from their respective photosynthetic sister clades (fig. 3).

Given these results, we might hypothesize that RUBISCO plays an important role in a function other than photosynthesis in some but not all nonphotosynthetic plants. An alternative explanation, however, is that photosynthesis has been lost only recently in these lineages, and our power estimates are inflated by the assumption that all of the observed substitutions occurred after the loss of photosynthesis. Although we have no way to infer the time at which photosynthesis was lost along the branches ending in single holoparasitic species (*A. orobanchoides*, *S. gesnerioides*, and *L. clandestina*), in the *Harveya* and *Orobanche* analyses we can invoke a parsimony argument to assert that photosynthesis had been lost before the holoparasitic sister species diverged from each other (fig. 3*D* and *E*). In both these cases, two power analyses were performed: (1) assuming loss

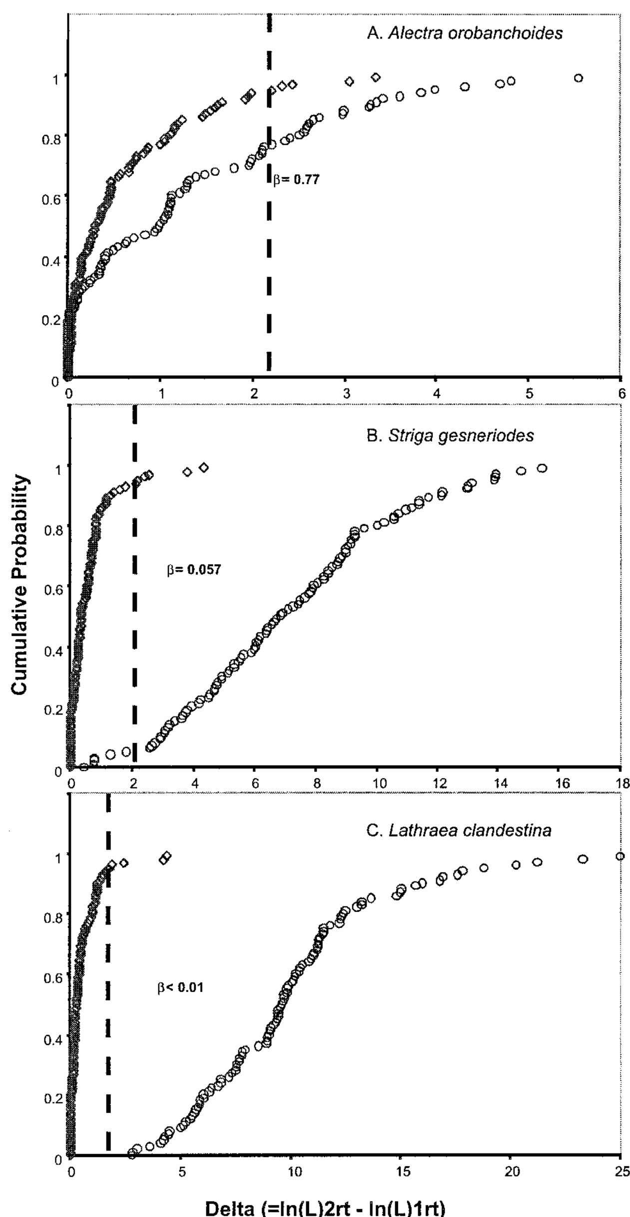

FIG. 4.—The cumulative probability distributions of the test statistic δ estimated from data sets ◇ simulated under the null model (one ω) and ○ simulated under the alternative model ($\omega_p \neq \omega_n = 1.0$). The dashed line represents the upper 5% of the null distribution.

of photosynthesis occurred with the origin of the holoparasitic clade (early), and (2) assuming loss of photosynthesis occurred at the time the sister species diverged (late).

In the *Harveya* analysis, the tests for change in ω with loss of photosynthesis was more powerful when photosynthesis was assumed to be lost early in the history of the holoparasitic clades (table 2, fig. 5). This makes sense because there is more time for change in constraint to affect the number of observed synonymous and nonsynonymous substitutions. The branches leading to both *Orobanche* species were quite long (fig. 3*E*), and there was no overlap observed between δ distributions estimated from data simulated under the null hypothesis ($\omega_p = \omega_n$) and the alternate hypothesis ($\omega_n = 1$) (fig. 5). Despite the power of the likelihood ratio tests in the *Harveya* and *Orobanche* analyses, no change in ω was detected when the *rbc*L genes from holoparasitic sister species were assumed to be evolving under the same degree of constraint. Even under the null model, however, the inferred ratio of amino acid to nucleotide substitutions was much higher for *H. capensis* relative to *H. purpurea*, and for *O. corymbosa* relative to *O. fasciculata* (fig. 3). On the basis of this observation, we estimated the likelihood ratios for change in ω along the branches leading to *H. capensis* and *O. corymbosa* (table 2). For each of these analyses, the *P*-values reported in table 2 would lead us to reject the null hypothesis if we had proposed a priori that the *rbc*L genes in *H. capensis* and *O. corymbosa* were evolving differently from the *rbc*L genes in their respective sister species. However, because we posed this hypothesis only after inspecting the inferred substitution patterns, the reported *P*-values are inflated, and should be interpreted with caution.

With this caveat in mind, the combined results of the likelihood ratio tests and power analyses clearly demonstrate that *rbc*L evolution has been constrained in the lineages leading to the holoparasitic species *S. gesnerioides*, *H. purpurea*, and *O. fasciculata*. These species are achlorophyllous and thought to be totally dependent on their hosts for carbohydrates. Although weak *rbc*L gene expression has been detected in *L. clandestina* (Bricaud, Thalouarn, and Renaudin 1986; Delavault, Sakanyan, and Thalouarn 1995; Lusson, Delavault, and Thalouarn 1998), there are no *rbc*L gene expression data for *S. gesnerioides*, *H. purpurea*, and *O. fasciculata*. Intact *rbc*L transcription promoters have been sequenced from both *O. fasciculata* and *O. corymbosa* (Wolfe and dePamphilis 1997). Interestingly, the stem-loop structure deduced for the 3′ untranscribed portion of the *rbc*L gene in photosynthetic plants, thought to be important in termination, is conserved in *O. corymbosa* but degenerate in *O. fasciculate* (Wolfe and dePamphilis 1997).

Although the *rbc*L genes in some of the holoparasitic lineages included in this investigation are apparently evolving under relaxed selective constraint, it is not surprising that the *rbc*L reading frames are still intact. Given the inferred ancestral sequences and the relatively short branches along which we are assuming loss of gene function, the probability of stop codons arising in the holoparasitic lineages was inferred to be quite low in most cases (table 2). As others have pointed out, unexpressed or obsolete genes may persist as open reading frames for some time even in the absence of selective maintenance (Marshall, Raff EC, and Raff RA 1994). It is also worth noting that the probability of the origin of a stop codon is not perfectly proportional to branch length. The probability is strongly influenced by the frequency of codons in the ancestral gene that are within one nucleotide substitution of a stop codon. As a consequence, the stop codon probabilities reported in table 2 may be inaccurate if the inferred ancestral sequences differ from the true ancestral sequences.

Because length mutations were not considered, the probability that an open reading frame would be trun-

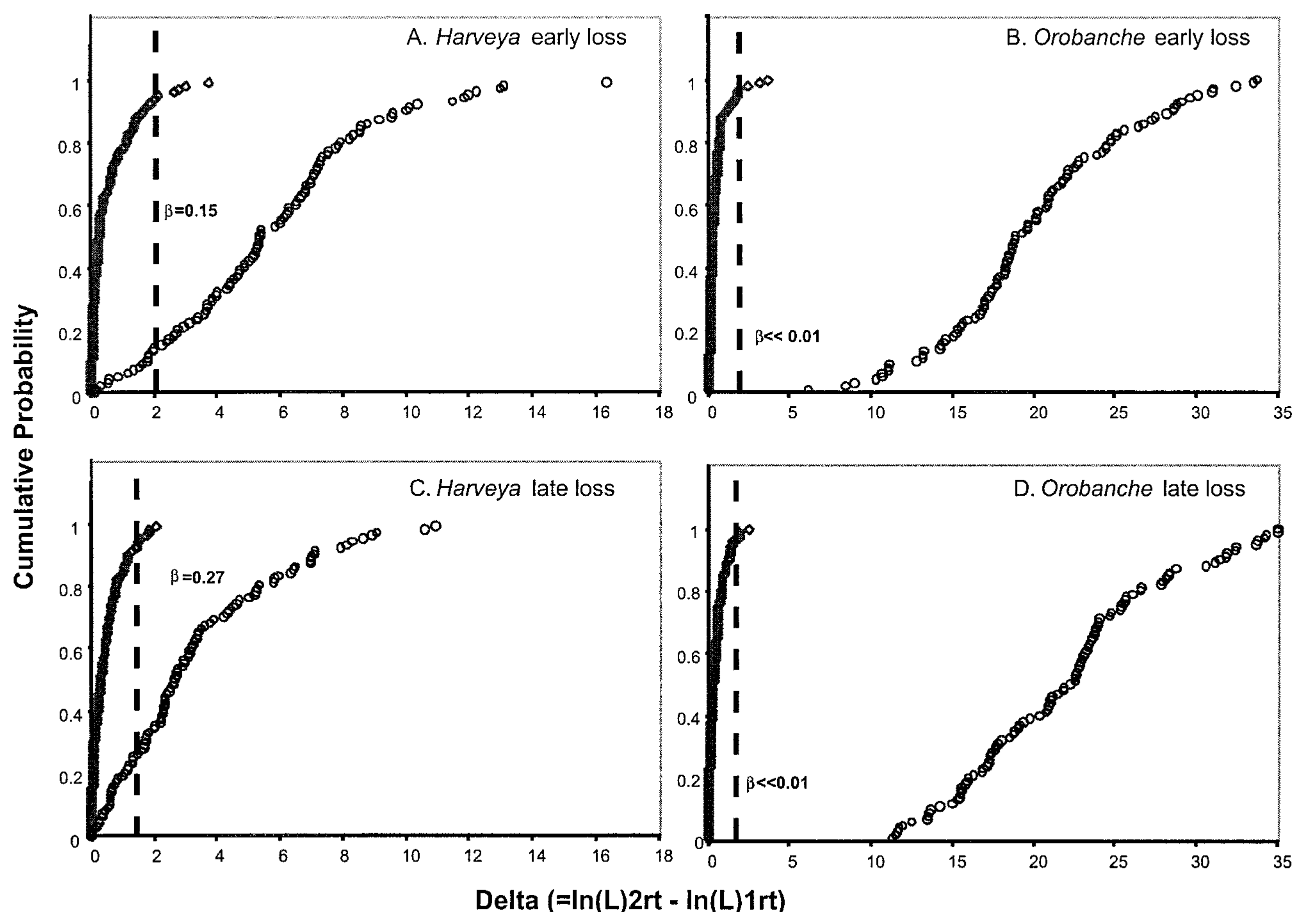

FIG. 5.—Analyses of loss constraint on *rbc*L in Harveya and Orobanche were performed assuming (*A* and *B*) early and (*C* and *D*) late loss of photosynthesis in either clade. The cumulative probability distributions of the test statistic δ estimated from data sets ◇ simulated under the null model (one ω) and ○ simulated under the alternative model ($\omega_P \neq \omega_n = 1.0$). The dashed line represents the upper 5% of the null distribution.

cated in the absence of selective constraint is even greater than the reported stop codon probabilities (table 2). Analyses of the noncoding regions of the plastid genome suggest that length mutations, especially those involving nucleotide repeats, may be as common as base changes (e.g., Golenberg et al. 1993). With this in mind, intact functional *rbc*L sequences can be taken as additional circumstantial evidence that the large subunit of RUBISCO has been selectively maintained in some holoparasitic lineages.

Crayfish Opsin Analysis

Although the contingency test for change in d_N/d_S used by Crandall and Hillis in their analysis of opsin genes in cave-dwelling crayfish (Crandall and Hillis 1997) differed from the likelihood ratio test described previously, the power of the test was also assessed with data sets simulated under the alternative hypothesis that hypothetically unconstrained sequences were evolving with d_N equal to d_S. For all the three contingency tests performed in the original article, β was found to be greater than 65% (fig. 6). The more standard power analysis of Cassagrande, Pike, and Smith gave very similar estimates of β (fig. 6). The results of the contingency analyses, therefore, are not suggestive of selectively maintained function for opsin in cave-dwelling crayfish.

Detecting Change in Function Versus Loss of Constraint

An adaptive change in gene function has been inferred as positive selection for increase in the ratio of nonsynonymous to synonymous nucleotide substitutions (e.g., Messier and Stewart 1997; Barrier, Robichaux, and Purugganan 2001). Although the null hypothesis in tests for positive selection is identical to the null hypothesis for the loss of function tests described previously, the alternative hypotheses are different. The alternative hypothesis for change of function test is $\omega > 1.0$. However, the inferred rate ratios may be much less than 1.0 when adaptive change has occurred over a short period of time or adaptations have involved just a few key amino acid substitutions. Depending on the context of a particular test, an estimated rate ratio less than unity may be interpreted as adaptive evolution (Yang 1998) or loss of function (table 2). The context is set by the ecological or biochemical information that motivated the a priori test for change in the ratio of nonsynonymous to synonymous change. When adaptive evolution is suspected,

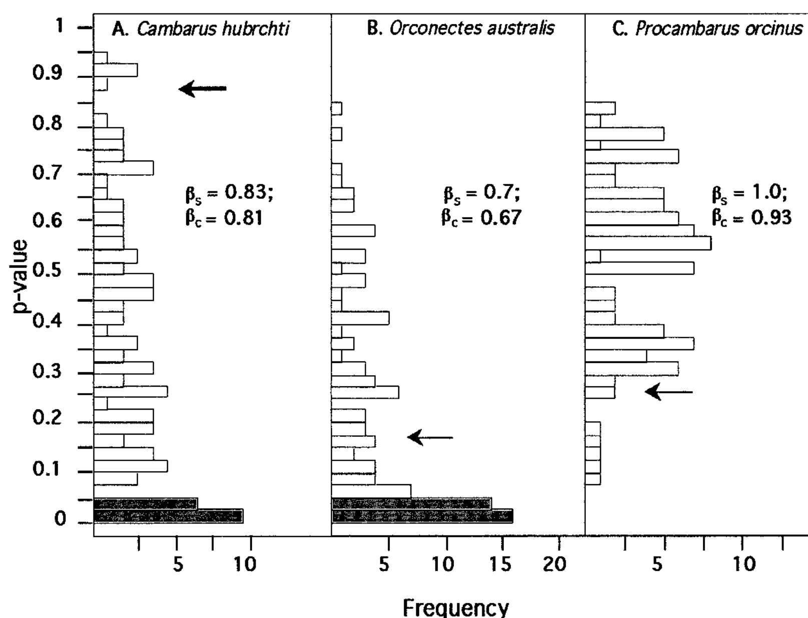

FIG. 6.—The distribution of *P*-values for Fisher's Exact Test performed on 100 data sets simulated according to the alternative hypothesis ($\omega_{surface} \neq \omega_{cave} = 1.0$). The dark bars represent replicates where the null hypothesis was rejected. Type-II error rates were estimated as proportion of simulated data sets with $P > 0.05$ (β_s) and using the method of Cassagrande et al. (β_c). Arrows represent *P*-values obtained for one-tailed Fisher's Exact Test applied to the original data (Crandall and Hillis 1997).

the power analyses described here to test for the loss of function could easily be modified for tests for positive selection on entire genes or specific domains.

The analyses of protein-coding genes described here are obviously simplistic. A single-rate ratio parameter, ω, is estimated and modeled for the entire gene. Studies of *rbc*L evolution have clearly demonstrated variation in amino acid substitution rates among structural domains (Kellogg and Juliano 1997; Wolfe and dePamphilis 1998). Recent advances in the modeling of amino acid substitution patterns among the structural domains exhibited in globular proteins (Goldman, Thorne, and Jones 1998) and transmembrane proteins (Liò and Goldman 1999) suggest that the estimation of variation in evolutionary rates among gene sequences and phylogenies would be enhanced by including information on the secondary structure. Even in instances when protein structure is not known, analyses allowing for among site variation in ω could help elucidate the function of particular sequence domains (Yang et al. 2000; Yang and Bielawski 2000). The likelihood ratio tests available in PAML for assessing variation in ω across a protein-coding gene (Yang et al. 2000) have been shown to be quite powerful (Anisimova, Bielawski, and Yang 2001). The number of parameters estimated in such an analysis, however, could be quite large in tests of variation in ω across a coding gene and among branches on a phylogeny. Failure to detect lineage-specific adaptive or diversifying selection in such analyses would be difficult to interpret without an accompanying power analysis.

Tests for variation in ω among domains and branches may be especially weak when branch lengths are short, the variation in selection among domains is subtle, or domain size is limited to a few codons. For example, most of the variation in the spectral sensitivity of vertebrate opsins can be explained by the amino acid composition at just five sites (Yokoyama and Radlwimmer 2001). Moreover, adaptive substitutions at one or two of these sites have resulted in spectral tuning that has maximized light sensitivity in extreme environments (Yokoyama 1997). The two functional opsin genes known from the Comoran coelacanth (*Latimeria chalumnae*) have evolved to produce pigments with maximal sensitivity (λ_{max}) in the very narrow range of the light spectrum that reaches its deep-sea habitat (Yokoyama et al. 1999). Functional cone (David-Gray et al. 1999) and rod (Janssen et al. 2000) opsin genes have been identified for the subterranean blind mole rat *S. ehrenbergi*. As with the coelacanth opsins, λ_{max} for the cone opsin shows an adaptive shift to the red portion of the spectrum relative to λ_{max} for homologous opsins in other mammals. Although experiments and anatomical studies have shown that *S. ehrenbergi* is indeed blind, multiple lines of evidence suggest that the reduced, subcutaneous eye and its photosensitive pigments have been coopted to aid in light-mediated entrainment of circadian rhythms (Argamaso et al. 1995; David-Gray et al. 1999; Janssen et al. 2000). In these examples, adaptive evolution has been demonstrated by linking change in amino acid sequence with change in protein function (Yokoyama S and Yokoyama R 1996; David-Gray et al. 1999; Yokoyama et al. 1999). A change in the ratio of nonsynonymous to synonymous nucleotide substitutions has not been demonstrated in these cases.

Conclusions

Comparative approaches in phylogenetics offer a solid framework for testing hypotheses concerning the

evolutionary processes responsible for the observed patterns of diversity (e.g., dePamphilis 1995). However, a recent change in the evolutionary process may not have detectable influence on patterns of diversity for some time. Power analysis provides a means of assessing whether the failure to detect a change in the evolutionary process is caused by the absence of change or insufficient data. Given the complexity of the processes being inferred in studies of molecular evolution, the power of tests for change in constraint or rate are likely to depend not only on the degree of variation among sequences but also on the phylogenetic relationships among the genes being analyzed (Bromham et al. 2000). The Monte-Carlo simulation offers a flexible means to generate distributions for hypothesis testing and power analyses that are tailored to each data set. The generality of the likelihood ratio test extends to the assessment of its power for specific applications (fig. 1). For example, the same approach described here for analyzing changes in evolution at the molecular level has been used to assess the power of a likelihood ratio test for simultaneous Pleistocene speciation events in the grasshopper genus *Melanoplus* (Knowles 2000).

Testing the fit of gene sequence data to explicit evolutionary models can certainly increase our understanding of gene function. The results of such tests, however, must be interpreted with caution. In some instances power analysis can guide the interpretation of statistical tests that have failed to detect change or reject a null hypothesis. In earlier articles, Wolfe and dePamphilis (1997, 1998) found no evidence of reduced evolutionary constraint on *rbc*L genes of the nonphotosynthetic plants included in their analysis. Three hypotheses were offered to explain their surprising result: (1) the plants are photosynthetic for at least a portion of their life, (2) the carboxylase or oxygenase activity of the peptide encoded by *rbc*L serves a critical function in some nonphotosynthetic plants, or (3) the open reading frame for *rbc*L is simply persisting by chance, and it will eventually be interrupted by a stop codon or length mutation. The power analyses performed here suggest that we can reject the third hypothesis for the persistence of *rbc*L in *S. gesnerioides*, *Harveya purpurea*, and *Orobanche fasciculata*. Physiological and gene expression studies are necessary to test the remaining two hypotheses.

Acknowledgments

We thank Hiroshi Akashi, Todd Barkman, Olle Pellmyr, Allan Strand, Andi Wolfe, Ziheng Yang, Ned Young, and two anonymous reviewers for their helpful comments on our manuscript. We thank Lena Landherr Sheaffer for sequencing additional *mat*K genes. We especially thank Ziheng Yang for helping us use routines from his CODEML program for use in our sequence generating simulation program. Keith Crandall and David Hillis kindly provided us with their crayfish opsin sequences. This work was supported through NSF grants to J.L.-M. and C.deP.

LITERATURE CITED

Anisimova, M., J. P. Bielawski, and Z Yang. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. **18**:1585–1592.

Argamaso, S. S., A. C. Froehloch, M. A. McCall, E. Nevo, I. Provencio, and R. G. Foster. 1995. Photopigments and circadian systems of vertebrates. Biophys. Chem. **56**:3–11.

Barkman, T. J., G. Chenery, J. R. McNeal, J. Lyons-Weiler, W. J. Elisens, G. Moore, A. D. Wolfe, and C. W. dePamphilis. 2000. Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. Proc. Natl. Acad. Sci. USA **97**:13166–13171.

Barrier, M., R. H. Robichaux, and M. D. Purugganan. 2001. Accelerated regulatory gene evolution in an adaptive radiation. Proc. Natl. Acad. Sci. USA **98**:10208–10213.

Bricaud, C. H., P. A. Thalouarn, and S. Renaudin. 1986. Ribulose 1,5-bisphosphate carboxylase activity in the holoparasite *Lathraea clandestina* L. J. Plant Physiol. **125**: 367–370.

Briscoe, A. D., and L. Chittka. 2001. The evolution of color vision in insects. Annu. Rev. Entomol. **46**:471–510.

Bromham, L., D. Penny, A. Rambaut, and M. D. Hendy. 2000. The power of relative rates tests depends on the data. J. Mol. Evol. **50**:296–301.

Brown, B. W., C. Brauner, A. Chan, D. Gutierrez, J. Herson, J. Lovato, J. Polsley, K. Russell, and J. Venier. 2000. DSTPLAN: calculations for sample sizes and related problems. Version 4.2. (http://odin.mdacc.tmc.edu/anonftp/page_2.html). University of Texas, M. D. Anderson Cancer Center.

Cassagrande, J. T., M. C. Pike, and P. G. Smith. 1978. An improved approximate formula for calculating sample sizes for comparing two binomial distributions. Biometrics **34**: 483–486.

Crandall, K. A., and D. M. Hillis. 1997. Rhodopsin evolution in the dark. Nature **387**:667–668.

David-Gray, Z. K., H. M. Cooper, J. W. H. Janssen, E. Nevo, and R. G. Foster. 1999. Spectral tuning of a circadian photopigment in a subterranean 'blind' mammal (*Spalax ehrenbergi*). FEBS Lett. **461**:343–347.

Delavault, P. M., V. Sakanyan, and P. A. Thalouarn. 1995. Divergent evolution of two plastid genes, *rbc*L and *atp*B, in a non-photosynthetic parasitic plant. Plant Mol. Biol. **29**:1071–1079.

dePamphilis, C. W. 1995. Genes and genomes. Pp. 177–205 *in* Malcolm C. Press and Jonathan D. Graves, eds. Parasitic plants. Chapman and Hall, London.

dePamphilis, C. W., and J. D. Palmer. 1990. Loss of photosynthetic and chlororespiratory genes from the plastid genome of a parasitic flowering plant. Nature **348**:337–339.

dePamphilis, C. W., N. D. Young, and A. D. Wolfe. 1997. Evolution of plastid gene *rps*2 in a lineage of hemiparasitic and holoparasitic plants: many losses of photosynthesis and complex patterns of rate variation. Proc. Natl. Acad. Sci. USA **94**:7367–7372.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.

Goldman, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. **36**:182–198.

Goldman, N., J. L. Thorne, and D. T. Jones. 1998. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. Genetics **149**: 445–458.

GOLDMAN, N., and Z. YANG. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

GOLENBERG, E. M., M. T. CLEGG, M. L. DURBIN, J. DOEBLEY, and D. P. MA. 1993. Evolution of a noncoding region of the chloroplast genome. Mol. Phylogenet. Evol. **2**:52–64.

HASEGAWA, M., H. KASHINO, and T. YANO. 1985. Dating the human-ape split by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

HUELSENBECK, J. P., and B. RANALA. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science **276**:227–232.

JANSSEN, J. W. H., P. H. M. BOVEE-GEURTS, Z. P. A. PEETERS, J. K. BOWMAKER, H. M. COOPER, Z. K. DAVID-GRAY, E. NEVO, and W. J. DEGRIP. 2000. A fully functional rod visual pigment in a blind mammal: a case for adaptive functional reorganization? J. Biol. Chem. **275**:38674–38679.

KELLOGG, E. A., and N. D. JULIANO. 1997. The structure and function of RuBisCo and their implications for systematic studies. Am. J. Bot. **84**:413–428.

KNOWLES, L. L. 2000. Tests of Pleistocene speciation in montane grasshoppers (genus *Melanoplus*) from the sky islands of western North America. Evolution **54**:1337–1348.

LI, W.-H., C.-I. WU, and C.-C. LUO. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitutions considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2**:150–174.

LIÒ, P., and N. GOLDMAN. 1999. Using protein structural information in evolutionary inference: transmembrane proteins. Mol. Biol. Evol. **16**:1679–1710.

LUSSON, N. A., P. M. DELAVAULT, and P. A. THALOUARN. 1998. The *rbc*L gene from the non-photosynthetic parasite *Lathraea clandestina* is not transcribed by plastid-encoded RNA polymerase. Curr. Genet. **34**:212–215.

MARSHALL, C. R., E. C. RAFF, and R. A. RAFF. 1994. Dollo's law and the death and resurrection of genes. Proc. Natl. Acad. Sci. USA **91**:12283–12287.

MESSIER, W., and C.-B. STEWART. 1997. Episodic adaptive evolution of primate lysozymes. Nature **385**:151–154.

NEI, M., J. ZHANG, and S. YOKOYAMA. 1997. Color vision of ancestral organisms of higher primates. Mol. Biol. Evol. **14**: 611–618.

OLMSTEAD, R. G., C. W. DEPAMPHILIS, A. D. WOLFE, N. D. YOUNG, W. ELISENS, and P. REEVES. 2001. Disintegration of the Scrophulariaceae. Am. J. Bot. **88**:348–361.

SHINOZAKI, K., M. OHME, M. TANAKA et al. (23 co-authors). 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expressions. Embo. J. **5**:2043–2049.

SWANSON, W. J., and V. D. VACQUIER. 1995. Extraordinary divergence and positive Darwinian selection in a fusagenic protein coating the acrosomal process of abalone spermatozoa. Proc. Natl. Acad. Sci. **92**:4957–4961.

SWOFFORD, D. L. 2000. PAUP*: phylogenetic analysis using parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland, Mass.

WOLFE, A. D., and C. W. DEPAMPHILIS. 1997. Alternative paths of evolution for the photosynthetic gene *rbc*L in four nonphotosynthetic species of Orobanche. Plant Mol. Biol. **33**: 965–977.

———. 1998. The effect of relaxed functional constraints on the photosynthetic gene *rbc*L in photosynthetic and nonphotosynthetic parasitic plants. Mol. Biol. Evol. **15**:1243–1258.

WOLFE, K. H., C. W. MORDEN, and J. D. PALMER. 1992. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc. Natl. Acad. Sci. USA **89**:10648–10652.

YANG, Z. 1994. Maximum Likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**:306–314.

———. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. **15**:568–573.

———. 1999. PAML: phylogenetic analysis by maximum likelihood. (http://abacus.gene.ucl.ac.uk/software/paml.html). University College London, London.

YANG, Z., and J. P. BIELAWSKI. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. **15**:496–503.

YANG, Z., S. KUMAR, and M. NEI. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141**:1641–1650.

YANG, Z., and R. NIELSEN. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. **46**:409–418.

YANG, Z., R. NIELSEN, N. GOLDMAN, and A. KRABBE-PEDERSEN. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431–449.

YOKOYAMA, S. 1997. Molecular genetic basis of adaptive selection: examples from color vision in vertebrates. Annu. Rev. Genet. **31**:311–332.

YOKOYAMA, S., and F. B. RADLWIMMER. 2001. The molecular genetics of red and green vision in vertebrates. Genetics **158**:1697–1710.

YOKOYAMA, S., and R. YOKOYAMA. 1996. Adaptive evolution of photoreceptors and visual pigments in vertebrates. Annu. Rev. Ecol. Syst. **27**:543–567.

YOKOYAMA, S., H. ZHANG, F. B. RADLWIMMER, and N. S. BLOW. 1999. Adaptive evolution of color vision of the Comoran coelacanth (*Latimeria chalumnae*). Proc. Natl. Acad. Sci. USA **96**:6279–6284.

YOUNG, N. D., and C. W. DEPAMPHILIS. 2000. Purifying selection detected in the plastid *matK* and flanking ribozyme regions within a group II intron of non-photosynthetic plants. Mol. Biol. Evol. **17**:1933–1941.

YOUNG, N. D., K. D. STEINER, and C. W. DEPAMPHILIS. 1999. The evolution of parasitism in Scrophulariaceae/Orobanchaceae: plastid gene sequences refute an evolutionary transition series. Ann. Mo Bot. Gard. **86**:876–893.

ELIZABETH KELLOGG, reviewing editor

Accepted April 2, 2002