

Evolutionary Rate Variation in Anthocyanin Pathway Genes

Yingqing Lu and Mark D. Rausher

Department of Biology, Duke University

Over a broad taxonomic range that spans monocots and dicots, upstream enzymes of the anthocyanin pigment pathway have evolved less rapidly than downstream enzymes. In this article we show that this pattern is also evident within the genus *Ipomoea*. Specifically, the most upstream enzyme, chalcone synthase (CHS-D), evolves more slowly than the two most downstream enzymes, ancyocyanidin synthase (ANS) and UDP glucose flavonoid 3-oxy-glucosyltransferase (UFGT). This pattern appears not to be due to variation in mutation rates, because the *CHS-D* gene exhibits higher synonymous substitution rates than the genes for the other two enzymes. Codon-based tests for positive selection suggest that it has been negligible or absent in all three genes. In addition, the mean number of indel-creating events is four times as high in the downstream genes as in *CHS-D*. Unlike the downstream genes, *CHS-D* also exhibits evidence of codon bias. Together, the evidence suggests that the difference in nonsynonymous substitution rates between upstream and downstream genes is due to relaxed constraint on the downstream genes rather than a greater frequency of positively selected substitutions.

Introduction

Proteins vary by orders of magnitude in their rates of evolution (Li, Wu, and Luo 1985), and accounting for this rate variation has been one objective of molecular evolutionary genetics. In some cases, particularly those involving proteins associated with mating and reproduction, it is evident that repeated positive selection is responsible for accelerated rates of evolution (Whitefield, Lovell-Badge, and Goodfellow 1993; Wyckoff, Wang, and Wu 2000; Swanson et al. 2001; Lu 2002), as reflected in greater rates of nonsynonymous substitution than synonymous substitution (i.e., $K_a/K_s > 1$). For most proteins, however, K_a/K_s is less than 1. Although differences in the K_a/K_s ratio among such genes are frequently ascribed to differences in the magnitude of selective constraint (i.e., differences in the strength of purifying selection; Kimura 1977; Li 1997, p. 185), differences in K_a/K_s may alternatively be due to differences in the frequency of positive selection. Although in principle it is possible to distinguish between these explanations for differences in evolutionary rates, this has seldom been attempted. Also, although more and more studies are documenting the occurrence of positive selection in genes with $K_a/K_s < 1$ (McDonald and Kreitman 1991; Yang et al. 2000; Bielawski and Yang 2001; Mondragón-Palomino et al. 2002), few attempts have been made to determine whether differences in the frequency of positive selection contribute to differences in evolutionary rates among such genes. In the report presented here, we describe an initial attempt to distinguish between these two explanations for rate variation among proteins in the anthocyanin biosynthetic pathway.

In a previous analysis of rates of anthocyanin structural gene evolution, Rausher, Miller, and Tiffin (1999) demonstrated that, over a broad taxonomic distance involving comparisons among monocots and dicots, proteins in the upstream portion of the pathway evolved

more slowly than proteins in the downstream portion. One possible explanation for this pattern is that the upstream enzymes are under greater selective constraint than the downstream enzymes. Greater constraint may arise because the upstream enzymes participate in the biosynthesis of a number of different types of flavonoids in addition to anthocyanins (Koes, Quattrocchio, and Mol 1994; Shirley 1996; Sakuda 2000). Flavonoids are believed to serve a number of physiological and ecological functions in plants, including protection from ultraviolet radiation and from natural enemies, facilitating interactions with mycorrhizal symbionts, and mediating pollen–stigma interactions. By contrast, the downstream enzymes are responsible for the production of only anthocyanins (Koes, Quattrocchio, and Mol 1994). It thus seems possible that deleterious mutations affecting enzyme kinetics in upstream genes would have a greater overall effect on fitness than similar mutations on downstream genes. Such mutations in downstream genes would then be more likely to be effectively neutral, causing the downstream genes to experience reduced selective constraint.

Because of the wide taxonomic comparison used in our previous analysis, for many of the genes examined, synonymous substitutions were saturated or close to saturation. Therefore, it was not possible to estimate accurately and compare K_a/K_s ratios for the different anthocyanin structural genes. The observed differences among the genes in nonsynonymous substitution rates could thus have been due to differences in processes (e.g., mutation rates) that affect synonymous and nonsynonymous rates similarly, and not to differences in the degree of selective constraint or the frequency of positive selection. In view of these uncertainties, we have undertaken a new analysis focusing on a taxonomically more restricted set of species within the genus *Ipomoea*. In this analysis, we addressed the following three questions: Is the greater rate of nonsynonymous substitution characteristic of the downstream enzymes detected by Rausher, Miller, and Tiffin (1999) also detectable in comparisons at lower taxonomic levels? If so, can differences in mutation rates be ruled out as the cause of the difference in evolutionary rates between upstream and downstream enzymes? Do downstream enzymes exhibit evidence of more frequent positive selection as a cause of their elevated evolutionary rates?

Key words: anthocyanin pathway, nucleotide substitution rates, positive selection, codon usage, rate variation.

E-mail: mrausher@duke.edu.

Mol. Biol. Evol. 20(11):1844–1853, 2003

DOI: 10.1093/molbev/msg197

Molecular Biology and Evolution, Vol. 20, No. 11,

© Society for Molecular Biology and Evolution 2003; all rights reserved.

Table 1
Summary of Species and Sequences Analyzed in This Study

Gene	Species	Accession Number	Compared Coding Length (bp)	Coding Regions GC Content ^a
<i>CHS</i>	<i>I. hederacea</i>	AY257204 ^b	844	0.770
	<i>I. nil</i>	AB027533	939	0.791
	<i>I. purpurea</i>	AF358659	939	0.777
	<i>I. alba</i>	AY257214 ^b	939	0.774
	<i>I. trifida</i>	AY257205 ^b	939	0.771
	<i>I. batatas</i>	AB037391	939	0.767
<i>ANS</i>	<i>I. hederacea</i>	AY257210	1047	0.687
	<i>I. nil</i>	AB073923	1047	0.687
	<i>I. purpurea</i>	AY257211 ^b	1047	0.678
	<i>I. alba</i>	AY257213 ^b	1044	0.682
	<i>I. trifida</i>	AY257212 ^b	1047	0.699
	<i>I. batatas</i>	BAA75305	1047	0.692
<i>UFGT</i>	<i>I. hederacea</i>	AY257206 ^b	1083	0.596
	<i>I. nil</i>	AY257208 ^b	1083	0.588
	<i>I. purpurea</i>	AY257209 ^b	1083	0.596
	<i>I. alba</i>	AY257215 ^b	1086	0.613
	<i>I. trifida</i>	AY257207 ^b	1068	0.558
	<i>I. batatas</i>	AB038248	1068	0.561

^a GC3 is the GC content of third codon positions.^b New sequence obtained in this study.

Materials and Methods

Species and Genes Analyzed

In this investigation, we limit our analysis to three of the six core anthocyanin structural genes: chalcone synthase D (*CHS-D*), anthocyanidin synthase (*ANS*), and UDP-glucose flavonoid 3-oxy-glucosyltransferase (*UFGT*). *CHS-D* codes for the most upstream enzyme in the pathway, and in the study of Rausher et al. (1999), of the six core genes identified, it exhibited the slowest rate of nonsynonymous substitution. Although it is a member of a small multigene family (Durbin et al. 1995; Huttley et al. 1997; Durbin, McCaig, and Clegg 2000), it is the primary *CHS* gene expressed in most plant tissues in *Ipomoea purpurea* (Durbin et al. 2000). By contrast, *ANS* and *UFGT* code for the two most downstream enzymes in the pathway and exhibit the fastest rate of nonsynonymous substitution. Both are believed to be single-copy genes in *Ipomoea*.

Sequences of *CHS-D*, *ANS*, and *UFGT* were obtained from six species chosen to represent much of the diversity within the genus *Ipomoea*, subject to the constraint that we wished to compare pairs of species separated by a range of genetic distances (table 1). Twelve of the 18 *Ipomoea* sequences compared were newly obtained in this study (see table 1 for accession numbers). *I. hederacea* was sampled from a natural population near Petersburg, Virginia, and *I. purpurea* came from Lee County, N.C. *I. alba* sequences are from a commercial strain. The National Center for Genetic Resources Preservation of the U.S. Department of Agriculture at Colorado provided seeds for *I. nil* (accession number NSL59662), and the Plant Genetic Resources Conservation Unit provided seeds for *I. trifida* (accession numbers PI 561547 and PI561543).

Sequencing

We extracted genomic DNA from young leaves using the Dneasy plant mini kit (Qiagen, Inc.). With genomic

DNA as a substrate, we employed polymerase chain reaction (PCR) to obtain fragments of each gene. For *CHS-D*, we designed the primers D2f (ctcgggtaagtgcgatctctc) and D2r (ccgactactattttcgtatcaact) to flank the second exon. For *ANS*, we paired P32 (caactgttcccagcagggtg) and P33 (gatgatgatcatcaatcaaaa) to obtain a fragment containing most of the two exons as well as the intron; we then used several internal primers (sequences are available upon request) for sequencing. A similar approach was used for *UFGT*: we designed primers A (accacagaggagttcatcatg) and C (aagcaggtgactaattcttggaa) to amplify the genomic fragment including most of its two exons and the intron, and later used A, C, and internal primers B (ctgaaagtaatcccagaatgctc) and D (at[ag]cagttgtggtcgtcgga) for sequencing. The PCR amplifications were carried out at 94°C for 1 min, 52°C for 1 min, 72°C for 3 min, all for 35 cycles on a PerkinElmer 480 thermal cycler. Amplified fragments were cloned into PCR vectors supplied by the TA cloning kit (Invitrogen). Sequencing reactions followed the Big Dye protocol (Applied Biosystems) and were conducted in both directions for all gene fragments. Additional primers used in these reactions included the M13 reverse and M13 forward primers. Sequence data were collected by an ABI 3700 automated sequencer (Applied Biosystems) following the standard protocol.

Comparing Ka and Ks Among Genes

Sequences were initially aligned using MegAlign (DNASTar, Inc), and the resulting alignments were visually adjusted. The best substitution model (F84 with a molecular clock; Felsenstein 1984) for our sequences was determined by comparing the likelihoods of commonly used models using the BaseML feature of PAML 3.13 (Yang 1997). We estimated Ka and Ks using Nei and Gotoh's method as implemented in DnaSP version 3.51 (Rozas and Rozas 1999).

To determine whether the rates of synonymous or nonsynonymous substitutions differed among genes, we compared the slopes of the relationships between Ka or Ks and genetic distance for the three genes. The rationale for this comparison is that, for a given gene, Ka for two species should be proportional to time since speciation. If genes differ in Ka (or Ks), the proportionality constant should differ, and this variation should be reflected in a difference in slope between Ka and time. We used genetic distance as a surrogate for separation time, and because separation times should be the same for all genes, the same measure of genetic distance was used for all genes. Genetic distances were estimated from the branch lengths of the maximum-likelihood phylogeny of the combined sequence of the three genes, using the F84 model with a molecular clock and PAML version 3.13 (Yang 1997). The topology (fig. 1) was first derived from previous studies (Miller, Rausher, and Manos 1999; Huang and Sun 2000; Manos, Miller, and M. D. Rausher. 2001) and is well supported by our new data.

Because pairwise Ka (and Ks) values are not statistically independent, a comparison of slopes using techniques of ordinary least squares (e.g., Searle 1971) is not appropriate. Instead, we used a generalized least squares

approach (Uyenoyama 1995; Martin and Hansen 1997), which essentially provides a comparison of slopes using independent contrasts (Felsenstein 1985; see *Appendix* for an explanation of approach). The analysis reported employed data from nine species pairs: *I. trifida*–*I. batatas*, *I. trifida*–*I. alba*, *I. trifida*–*I. hederacea*, *I. trifida*–*I. nil*, *I. batatas*–*I. hederacea*, *I. alba*–*I. purpurea*, *I. alba*–*I. nil*, *I. purpurea*–*I. nil*, and *I. hederacea*–*I. nil*. Analysis using several other sets of nine species pairs yielded similar results.

Comparing Ka/Ks Among Genes

We compared Ka/Ks ratios for the three genes using two different methods. The first method estimates Ka/Ks from the ratio of slopes of regressions of Ka and Ks on genetic distance obtained from the analysis described in the previous section. The second method uses estimates of the parameter ω generated by the program CodeML of the PAML software package (Yang 1997). For this analysis, a single value of Ka/Ks was obtained for each gene using the model M0 of CodeML. The significance of differences in Ka/Ks between genes was assessed by comparing the likelihood of the model using the estimated value of Ka/Ks to the likelihoods of the same model using Ka/Ks constrained to various values. In particular, we sought to determine whether there existed a value of Ka/Ks such that (1) that value is between the Ka/Ks ratios for the two genes being compared and (2) that value is significantly different from the value for each gene. If such a value is found, it indicates that the confidence intervals (or support regions, *sensu* Edwards 1972) for the Ka/Ks values of the two genes do not overlap, and thus the values are significantly different. Similarly, if a value of Ka/Ks can be found that lies between the values of Ka/Ks for two genes, and if that value is not significantly different from the value for either gene, then a single Ka/Ks is compatible with both genes and the genes can be inferred not to have significantly different Ka/Ks values. Significance was assessed using standard likelihood-ratio statistics (Weir 1990). Results are reported for a model that implements a molecular clock. Results from a model that does not implement a clock were quantitatively almost identical and are therefore not reported.

Estimates of Codon Use Bias

We computed effective number of codons (ENC; Wright 1990) for the three genes using DNAsp version 3.51 (Rozas and Rozas 1999); ENC is a measure of the degree to which codon usage deviates from equal use of the 61 possible sense codons. Because codon bias is correlated with GC content (Ikemura 1985), and because GC content is higher in *CHS-D* than in *ANS* and *UFGT* (table 1), direct comparisons of the ENCs among the three genes may be confounded by mutation bias (Foster, Eisenstadt, and Cairns 1982). To control for mutation bias, we compared the observed value of ENC for each gene to the value expected for that gene's third-position GC content using the Nc-plot technique of Wright (1990).

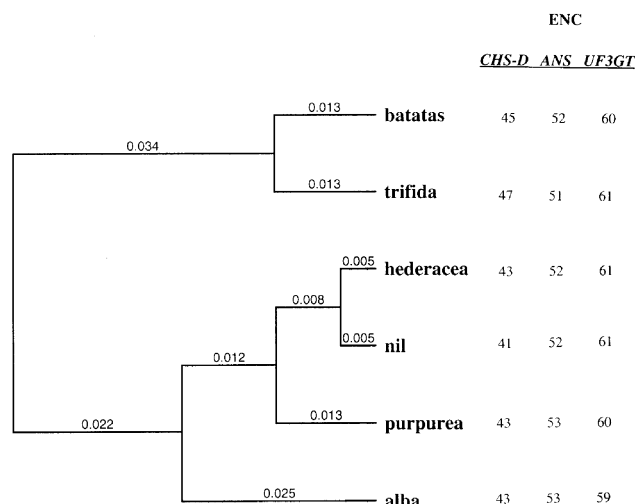


FIG. 1.—The maximum likelihood tree obtained using combined coding sequences of *CHS-D*, *ANS*, and *UFGT* and the F84 (clock) model of nucleotide substitution. Numbers on branches are estimated branch lengths. Numbers to the right of the species name are estimates of the effective number of codons (ENC) for each gene.

Detecting Positive Selection

We employed a codon-based approach to detect amino-acid sites that have undergone positive selection and to determine whether the proportion of such sites differed among the three genes. To implement this approach, we used the program CodeML of the PAML software package (Yang 1997). A comparison of the Ka/Ks ratio was first performed across lineages to verify that lineages did not differ. Subsequently, using the known phylogenetic topology (fig. 1), we determined for each gene separately the number of codons exhibiting evidence of positive selection by comparing models M7 and M8, as recommended by Yang et al. (2000). Model M7 fits codon variation in Ka/Ks with a beta distribution, with Ka/Ks constrained to be equal to or less than one. Model M8, however, includes an additional class of codons that may be positively selected (i.e., with Ka/Ks > 1) and estimates both the proportion of such codons in the gene and their average Ka/Ks ratio. Comparison of the two models by standard likelihood techniques assesses whether the inclusion of positively selected sites in the model provides a significantly better fit to the data (Yang et al. 2000).

Results

Substitution Models

The relative suitabilities of 11 nucleotide substitution models were examined separately for the three genes (table 2). We determined that, for all three genes, the most appropriate substitution model is the Felsenstein 84 (F84) model with a molecular clock (model 5 in table 2), by comparing the log-likelihoods of the various models. For all three genes, the substitution model with the highest likelihood was the HKY model (Hasegawa, Kishino, and Yano 1985) with unequal bases per branch (model 9), followed by the HKY model with a molecular clock and unequal bases per branch (model 8), and the HKY model

Table 2
Comparison of Nucleotide Substitution Models for Each Gene

	Model ^a	#Parameters	<i>CHS-D</i>		<i>ANS</i>		<i>UFGT</i>	
			InL ^b	Alpha ^c	InL	Alpha	InL	Alpha
1	JC69 & clock	6	-2135.63	0.355	-2158.2	0.899	-2409.8	0.575
2	K80 & clock	7	-2128.24	0.343	-2155.78	0.878	-2397.23	0.529
3	F81	11	-2105.56	0.289	-2132.54	0.782	-2403.99	0.563
4	F84	12	-2096.32	0.264	-2130.11	0.741	-2388.55	0.476
5	F84 & clock	7	-2099.61	0.288	-2135.47	0.766	-2390.73	0.508
6	HKY85 & clock	7	-2099.58	0.288	-2130.32	0.742	-2391.34	0.503
7	HKY85	12	-2096.28	0.264	-2134.67	0.765	-2389.21	0.471
8	HKY, clock, & nhomo2	16	-2094.36	0.194	-2132.97	7.286	-2388.55	5.23
9	HKY & nhomo2	21	-2090.89	0.00001	-2127.82	0.00001	-2386.38	4.38
10	TN93 & clock	7	-2099.87	0.292	-2146.81	0.765	-2393.19	0.509
11	REV & clock	8	-2098.83	0.289	-2135.36	0.767	-2389.15	0.519

^a JC69, Jukes-Cantor 1969 model; K80, Kimura 1980 model; F81, Felsenstein 1981 model; F84, Felsenstein 1984 model; HKY85, Hasegawa-Kishino-Yano 1985 model; TN93, Tamura-Nei 1993 model; REV, Yang's 1994 general time-reversible process model; clock, imposing molecular clock; nhomo2, allowing the ratio of transition to transversion to vary among phylogenetic branches. For description of models, see Yang (1997).

^b The natural log of the likelihood of the data under the model.

^c Parameter of the Gamma distribution used to estimate variation in substitution rates among sites (Yang 2000).

without a molecular clock (model 7) (table 2). However, for all three genes, none of these models fit the data significantly better than the HKY model with a molecular clock (model 6; this model is nested within each of models 7–9), given the difference in number of parameters ($P > 0.15$ in 7 of 9 tests, $P > 0.05$ in 2 of 9 tests; log-likelihood ratio test for nested models; Edwards 1972, Weir 1990). The F84 model with a clock (model 5), has the same number of parameters as model 6. Because it has similar likelihood for each of the three genes when examined individually, and because it has a slightly higher likelihood in a combined analysis when the sequences from all three genes are concatenated, it is slightly preferable to the HKY model with a clock. The only other model with a likelihood higher than F84 is the general time-reversible model with a clock (model 11). For all three genes, however, the likelihood is only slightly higher than that of the F84 model, and this difference is not significant, given the extra parameter. All other models (e.g., models 1–4 and 10) have lower likelihoods with the same number of parameters or more parameters than F84 for all three genes, and thus are less suitable.

Because the F84 (clock) model is appropriate for all three genes, we used this model with the combined data (concatenated sequences) from the three genes to estimate branch lengths for the species tree portrayed in figure 1. These branch lengths are also shown in the figure, and they are used to calculate the genetic distances for the analyses in the next section. Analyses similar to those described below were also done using distances calculated with the F84 substitution model without a clock. In all cases, results were qualitatively similar and hence are not presented.

Comparing Ka and Ks Among Genes

As is expected with the operation of a molecular clock, the relationship between Ka and genetic distance is linear (fig. 2a). The rate at which Ka increases with genetic distance is lowest for *CHS-D*, the most upstream gene, intermediate for *ANS*, and highest for *UFGT* (fig. 2a), a pattern that is identical with that reported earlier for

nonsynonymous substitutions in these genes across angiosperms (Rausher, Miller, and Tiffin 1999). A generalized least-squares analysis indicates that the slopes of Ka versus genetic distance differ significantly among the three genes ($F_{2,24} = 143.5$, $P < 0.0001$). Pairwise comparisons using a similar analysis reveals that for all three gene pairs, the slopes differ significantly (*CHS-D* vs. *ANS*— $F_{1,16} = 574.6$, $P < 0.0001$; *CHS-D* vs. *UFGT*— $F_{1,16} = 909.9$, $P < 0.0001$; *ANS* vs. *UFGT*— $F_{1,16} = 52.2$, $P < 0.001$).

The relationship between Ks and genetic distance is also linear, and the slope of this relationship differs among the three genes, but in a way unlike the pattern exhibited by Ka (fig. 2b). Specifically, the rate at which Ks increases with genetic distance is greatest for *CHS-D*, while it is lower and similar in magnitude for *ANS* and *UFGT*. Overall, a generalized least-squares analysis indicates that the three genes do not have equal slopes ($F_{2,24} = 1039.5$, $P < 0.0001$). In addition, all pairwise comparisons among genes were highly significant (*CHS-D* vs. *ANS*— $F_{1,16} = 1442.7$, $P < 0.0001$; *CHS-D* vs. *UFGT*— $F_{1,16} = 1009.1$, $P < 0.0001$; *ANS* vs. *UFGT*— $F_{1,16} = 65.7$, $P < 0.001$).

Because the same data were used to estimate Ka, Ks, and genetic distance, the dependent and independent variables used in this analysis are not statistically independent and the calculated significance levels may therefore be inflated. To assess the magnitude of this potential problem, we examined the correlations between genetic distance and pairwise differences in Ka (or Ks) for the three genes. As described in the *Appendix*, this correlation is expected to be zero under the null hypothesis that Ka (or Ks) is equal for the two genes. For this analysis, we again used the generalized least-squares approach by transforming both genetic distance and difference in Ka (or Ks, see *Appendix*).

This analysis produces results qualitatively similar to the results of the previous analysis, although as expected, some comparisons are no longer significant. The comparison between *CHS-D* and *UFGT* yielded a highly significant correlation ($r = 0.90$, $P < 0.005$, $df = 8$), indicating that Ka is significantly lower for *CHS-D* than for *UFGT*.

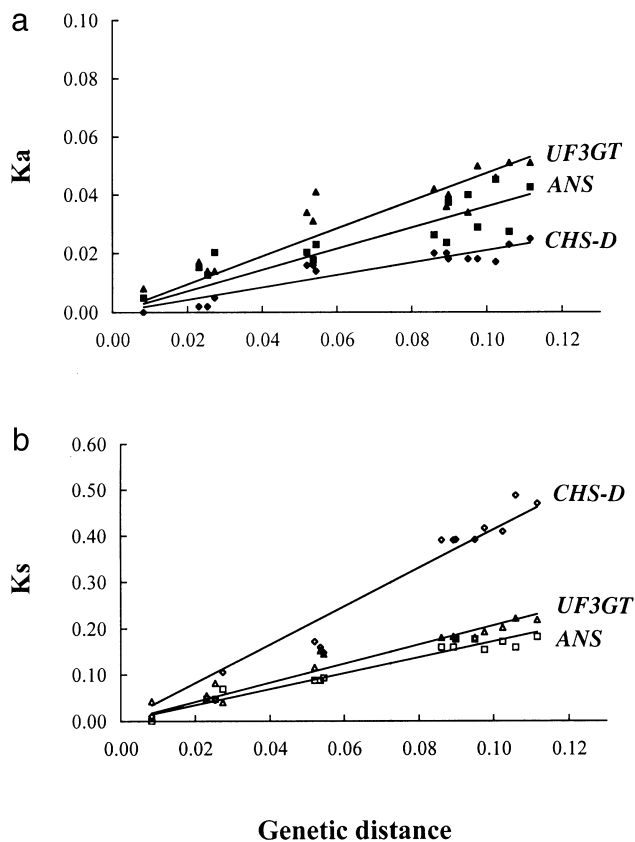


FIG. 2.—Comparisons of Ka and Ks among three anthocyanin genes. Genetic distance between two species is the sum of the branch lengths separating them in figure 1. Each point represents one of the 15 species pairs among the six species examined. (a) Relationship between Ka and genetic distance. (b) Relationship between Ks and genetic distance.

The comparison between *CHS-D* and *ANS* did not yield a significant correlation ($r = 0.28$, $P > 0.1$, $df = 8$). However, we note that for all 15 species pairs, the Ka value for *CHS-D* is less than that for *ANS* (fig. 2), suggesting that Ka is lower for *ANS* and that the nonsignificance of the correlation is likely due to lack of power to detect a difference in the slopes of the relationship between Ka and genetic distance for these two genes. There was also no detectable difference in Ka between *ANS* and *UFGT* ($r = 0.525$, $P > 0.05$). The rate of synonymous substitution, Ks, was significantly greater for *CHS-D* than for *ANS* or *UFGT* ($r = -0.97$ and -0.98 , $P < 0.001$ in both cases), but did not differ between *ANS* and *UFGT* ($r = 0.131$, $P > 0.1$).

Comparing Ka/Ks Among Genes

The Ka/Ks ratios calculated for *ANS* and *UFGT* from the ratio of regression slopes obtained in the previous analysis are similar to each other but are approximately four times the analogous ratio for *CHS-D* (table 3). The small standard errors of the estimates indicate that this difference between *CHS-D* and the downstream genes is statistically significant, while the difference between *ANS* and *UFGT* is not. Moreover, statistical significance of the difference between *CHS-D* and the downstream genes is

Table 3
Rates of Synonymous and Nonsynonymous Substitution in Anthocyanin Genes

Slope	<i>CHS</i> ^c	<i>ANS</i> ^c	<i>UFGT</i> ^c
Ka vs. Genetic distance ^a	0.21 (0.007)	0.41 (0.02)	0.51 (0.02)
Ks vs. Genetic distance ^a	3.72 (0.08)	1.82 (0.05)	2.21 (0.06)
Ka/Ks (ratio of slopes) ^b	0.056 (0.002)	0.232 (0.01)	0.227 (0.01)
Ka/Ks (ω) ^d	0.034 (0.0001)	0.197 (.0001)	0.277 (.0026)

^a In the first two rows, values are coefficients of a regression of Ka and Ks, respectively, on genetic distance.

^b Ka/Ks calculated as the ratio of regression coefficients for Ka and Ks.

^c Standard errors are in parentheses.

^d Ka/Ks calculated as the parameter ω using PAML model M0.

indicated by the fact that *CHS-D* has both a significantly lower Ka and a significantly higher Ks, compared to the other genes, as described above.

Ka/Ks ratios calculated using PAML are similar to those described above (table 3). Log-likelihood ratio comparisons indicate that the estimated Ka/Ks of 0.034 for *CHS-D* is significantly greater than a value of 0.11 ($\chi^2_1 = 34$, $P < 0.001$), whereas the estimated Ka/Ks ratios for *ANS* (0.197) and *UFGT* (0.277) are significantly greater than 0.11 ($\chi^2_1 = 8.0$ and 27.8 , $P < 0.01$ and 0.001 , respectively). Because there exists a value of Ka/Ks that is between the value for *CHS-D* and the values for *ANS* and *UFGT*, and because that value is not consistent with the Ka/Ks of any of the genes, it can be inferred that Ka/Ks is significantly lower for *CHS-D* than for the other two genes. By contrast, neither the estimated Ka/Ks ratio for *ANS* nor the ratio for *UFGT* differs significantly from a value of 0.237, indicating that a single value is consistent with both estimates and that they are thus not significantly different. Both the regression and PAML approaches thus yield the same conclusion: Ka/Ks is substantially lower for *CHS-D* than for *ANS* or *UFGT*.

Codon Use Bias

The three genes examined here differed substantially in degree of codon bias, as measured by the ENC. The gene with the lowest rate of nonsynonymous substitution, *CHS-D*, exhibited substantial codon bias, with a mean ENC across all six species of 43.7, indicating that in this gene 17 codons are effectively unused. By contrast, the gene with the highest nonsynonymous substitution rate, *UFGT*, exhibited little bias (mean ENC = 60.2); finally, *ANS*, with an intermediate nonsynonymous substitution rate, also exhibited an intermediate degree of codon bias (mean ENC = 52.0). Within each gene, there is very little variation in codon bias across species (fig. 1), indicating that the magnitude of codon bias has undergone little change during most of the diversification of the genus *Ipomoea*. This lack of evolutionary change suggests that degree of codon bias has reached a different evolutionary equilibrium for each gene, or that the rate of evolutionary change in ENC is very slow.

The three genes also differ in GC content at the third codon position, with *CHS-D* having the highest GC content and *UFGT* having the lowest (table 1). One possible explanation for this difference is that CHS is under more intense selection for codon-use bias, because in many

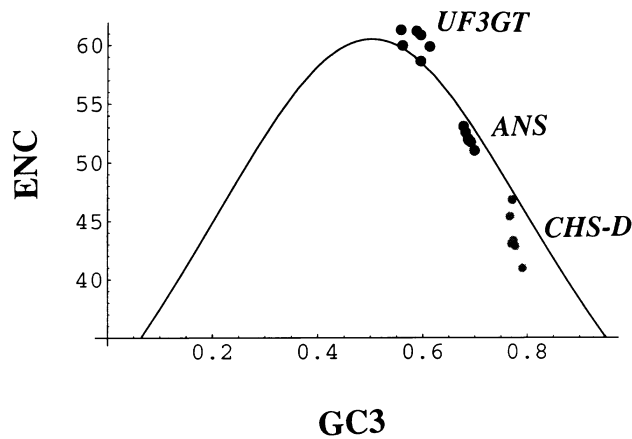


FIG. 3.—Codon bias in anthocyanin genes. The predicted relationship between GC percentage at the third (synonymous) codon position (GC3) and ENC under the assumption of no selection on codon usage is given by the solid curves. Points are observed values of GC3 and ENC for the six species examined.

species most preferred codons end in G and C (Fennoy and Bailey-Seres 1993; Chiapello et al. 1998). Nevertheless, because differences in GC content can be caused by processes other than selection for codon bias (e.g., mutation bias), and because differences in GC content can by themselves cause differences in ENC even in the absence of selection for codon bias, we asked whether ENC for each gene deviated from expectation based simply on the gene's GC content. In figure 3, the solid line portrays this expectation as a function of third-position GC content (Wright 1990). For *UF3GT*, the points for all six species lie on or above this expectation. For *ANS*, the points for all six species lie just below the expectation line, and for *CHS-D*, the points for all species lie below the expectation and several lie substantially below that level. This pattern is consistent with historical selection for codon bias in *CHS-D*, perhaps weaker selection for codon bias in *ANS*, and the absence of such selection in *UF3GT*.

Detecting Positive Selection

Although the lower Ka/Ks ratio for *CHS-D* is consistent with the hypothesis that this gene is subject to more intense purifying selection (selective constraint) than *ANS* or *UF3GT*, it is also consistent with an alternative hypothesis: the degree of purifying selection is similar for all three genes, but *ANS* and *UF3GT* have experienced more frequent episodes of positive selection than has *CHS-D*. However, this alternative hypothesis is not supported by our codon-based analysis of selection. For neither *CHS-D* nor *UF3GT* is there any indication of positive selection having acted. Analysis of these genes using models M7 and M8 of Yang et al. (2000) gives no indication that a class of positively selected codons exists for either gene (table 4). For *CHS-D*, for example, the estimated frequency of such additional codons is about 0.26, but the estimated Ka/Ks ratio for this class of codons is only 0.10; positively selected codons should have a ratio greater than one. Similarly, for *UF3GT*, the estimated frequency is

approximately $\frac{1}{2}$, but the estimated Ka/Ks for this class of codons is 0.546, again substantially less than 1.

By contrast, we did detect some evidence of positive selection in *ANS*. Comparison of models M7 and M8 detects a class of sites that have undergone positive selection (table 4). Although the estimated average Ka/Ks ratio for these sites is 30.94, there are only two such sites among the 349 codons examined (table 4). Although there is some uncertainty associated with the estimated number of selected sites, the standard error of this estimate is only 1 site, which suggests that positive selection has occurred at, at most, four sites (mean + 2 S.E.). This small proportion of positively selected sites does not seem likely to account for the fourfold to sixfold greater Ka/Ks in *ANS* than in *CHS-D*.

Discussion

The Rausher, Miller, and Tiffin (1999) analysis of rates of evolution of anthocyanin pathway genes demonstrated that, over a broad taxonomic scale encompassing both monocots and dicots, genes coding for upstream enzymes had lower rates of nonsynonymous substitution than genes coding for downstream enzymes, suggesting that upstream genes are under greater selective constraint. The present analysis supplements this finding by revealing additional details of this pattern.

Although in this study we examined only three of the six core pathway enzymes, we found the pattern of nonsynonymous rate variation across the genus *Ipomoea* to be consistent with the pattern documented by Rausher, Miller, and Tiffin (1999). Specifically, in both studies, the most upstream gene in the pathway, *CHS-D*, had the lowest nonsynonymous rate, while the most downstream, *UF3GT*, had the highest rate. And in both studies, the rate for *ANS* was intermediate between those of the other two genes. This similarity indicates that the pattern exhibited across angiosperms is also detectable at a much finer taxonomic scale.

One possible explanation for this difference in nonsynonymous substitution rates is that, for some reason, downstream genes have higher mutation rates than upstream genes. Although there is no a priori reason to expect this hypothesis to be true, Rausher, Miller, and Tiffin (1999) could not rule it out because nonsynonymous substitutions were nearly saturated for the distantly related species they examined. It was therefore impossible to determine whether synonymous substitution rates were lower for upstream genes, as this hypothesis would predict. Because of the narrower taxonomic focus of the present study, we were able to obtain reliable estimates of synonymous substitution rates, and thus resolve this issue. Like the nonsynonymous rates, the synonymous rates also differ among the three genes examined. Somewhat surprisingly, however, the upstream gene (*CHS-D*) had a substantially higher rate of synonymous substitution than the two downstream genes (fig. 2). The higher rates of amino-acid substitution in the downstream genes cannot, therefore, be due to an elevated mutation rate.

Our failure to detect more than minimal positive selection on any of the three genes examined suggests that the increased rate of nonsynonymous substitution in the

Table 4
Test for Positively Selected Amino Acid Sites

Gene	Model ^a	Parameters estimated	lnL ^b	Chi-sq. test ^c	Ka/Ks ^d
<i>CHSD</i>	M7	$p = 0.323, q = 8.337$	-1865.83		0.035
	M8	$p_0 = 0.736, p = 0.011, q = 8.330$ ($p_1 = 0.264$), $\omega = 0.103$	-1865.77	$P > 0.5$	0.035
<i>ANS</i>	M7	$p = 0.011, q = 0.038$	-1998.66		0.205
	M8	$p_0 = 0.994, p = 0.104, q = 0.481$ ($p_1 = 0.006$), $\omega = 30.9$	-1986.24	$P < 0.001$	0.355
<i>UFGT</i>	M7	$p = 0.413, q = 1.025$	-2284.16		0.286
	M8	$p_0 = 0.479, p = 0.001, q = 1.580$ ($p_1 = 0.521$), $\omega = 0.546$	-2283.92	$P > 0.1$	0.284

^a Model M7 (Yang et al. 2000) fits Ka/Ks ratios at individual sites to a beta distribution with parameters p and q ; a total of 12 parameters are estimated under this model. Model M8 assumes a proportion p_0 of sites follow a beta distribution, while a proportion p_1 ($= 1 - p_0$) sites have a common Ka/Ks equal to ω , which may be > 1 ; a total of 14 parameters are estimated under the model.

^b The natural log of the likelihood of the data given the model.

^c Significance level of the likelihood ratio χ^2 statistic for comparing the two models.

^d Average Ka/Ks ratio across sites, as estimated from the model.

downstream enzymes is more likely to be due to relaxed constraint on these enzymes than to increased rates of positive selection. This interpretation is also consistent with indel patterns and with the differences among genes in the degree of codon bias. If downstream genes were under reduced selective constraint, they would be expected to tolerate more indels than upstream genes. Corresponding to this expectation, the most downstream gene, *UFGT*, exhibited seven identifiable insertion/deletion events during the divergence of the species examined. By contrast, the most upstream gene, *CHS-D*, exhibited only one indel. Interpretation of this pattern is complicated by the fact that the downstream gene *ANS* also exhibited only 1 indel, but when averaged together, the two downstream genes still exhibited four times as many indels per gene as *CHS-D*.

The magnitude of codon bias in a gene is also often taken to reflect the degree to which that gene is subject to selective constraint, for several reasons. First, codon bias is believed to reflect selection for translation efficiency in highly expressed genes (Bennetzen and Hall 1982; Sharp and Li 1987; Powell and Moriyama 1997). Because highly expressed genes are likely to be more "critical" to an organism, they are also likely to be subject to greater selective constraint. Second, Akashi (1994) demonstrated that highly functional and evolutionarily conserved amino acids within genes had greater codon bias than less conserved amino acids, presumably because of selection for translational accuracy. This result suggests that genes with a greater fraction of constrained amino acid sites will also exhibit a greater average degree of codon bias. Finally, Comeran and Kreitman (1998) demonstrate an excess of doubly substituted codons in *Drosophila* and argue that the best explanation for this effect is relaxed selection on the codons involved, which would tend to decrease codon bias. These arguments suggest that our finding of reduced codon bias in downstream genes is due to relaxed constraint in these genes.

One potentially puzzling result of our analyses is that the gene with the greatest codon bias (lowest ENC), *CHS-D*, is also the gene with the highest rate of synonymous substitution, contrary to the pattern reported for other

genes (e.g., Fitch and Strausbaugh 1993; Zhang, Vision, and Gaut. 2002; but see also Kusumi et al. 2002). It might be expected that at codon-bias equilibrium, more intense selection for codon bias would decrease synonymous substitution rates. This expectation is reasonable for genes that have a similar underlying synonymous mutation rate, because once codon bias reaches equilibrium, selection to maintain bias is essentially a form of purifying selection. Consequently, the stronger the selection for bias, the greater the selective constraint, and the lower the expected substitution rate. However, if for *CHS-D* the underlying synonymous mutation rate is substantially higher than for *ANS* or *UFGT*, then the synonymous substitution rate for *CHS-D* will be elevated compared to the other genes, and that gene's turnover rate of preferred and nonpreferred codons will increase, but the equilibrium codon bias should not be affected. We thus suggest that the association of high codon bias with high rates of non-synonymous substitution in *CHS-D* can be explained as resulting from a relatively elevated mutation rate at this locus.

In conclusion, it appears that the three independent types of evidence obtained in this study are consistent with the hypothesis that amino acid substitutions occur more frequently in downstream anthocyanin genes because of reduced constraint, rather than as a result of enhanced positive selection. One caveat must be added to this conclusion, however. Betancourt and Presgraves (2002) recently demonstrated in *Drosophila* that strongly selected amino acid sites exhibit reduced codon bias, presumably as a result of selective interference. If this effect were strong in the *Ipomoea* genes examined here, that would imply that the downstream genes experience a greater frequency of positively selected substitutions than the upstream genes. This expectation is inconsistent, however, with our failure to detect any evidence of substantial positive selection on the downstream genes.

Appendix

In this appendix, we describe our generalized least-squares approach to comparing the Ka and Ks among

genes. We assume a molecular clock (see text for justification) and examine the relationship between K_a (or K_s) and genetic distance, asking whether the slope of the relationship differs for different genes. We describe here the analysis for K_a . The analysis for K_s is completely analogous.

Under the generalized least-squares framework, the statistical model for the relationship between K_a and genetic distance, D , is given by

$$E[\mathbf{Ka}] = \mathbf{D}\mathbf{b},$$

where $E[\]$ is the expectation operator, \mathbf{Ka} is the vector of pairwise \mathbf{Ka} values, \mathbf{D} is the corresponding vector of pairwise genetic differences, and \mathbf{b} is the slope of \mathbf{Ka} on genetic distance. The expectation of the variance-covariance matrix of \mathbf{Ka} is assumed to be

$$\text{Var}[\mathbf{Ka}] = \Sigma.$$

As shown below, if all 15 species pairs are included in the analysis, Σ is singular. However, there are sets of nine species pairs for which covariances are linearly independent and which thus yield a Σ matrix that is non-singular and positive-definite. The remaining analysis pertains to such a subset of species pairs.

Let \mathbf{C} be the matrix of normalized eigenvectors of Σ and let $\mathbf{\Lambda}$ be a diagonal matrix of the corresponding eigenvalues. Then, because Σ is positive-definite,

$$\mathbf{C}^T \Sigma \mathbf{C} = \mathbf{\Lambda},$$

and all of the diagonal elements of $\mathbf{\Lambda}$ are positive, implying that $\mathbf{\Lambda}^{-1/2}$ exists. Therefore,

$$\Sigma = [\mathbf{C}^T]^{-1} \mathbf{\Lambda} \mathbf{C}^{-1}.$$

If we let $\mathbf{A} = \mathbf{C} \mathbf{\Lambda}^{-1/2}$, then

$$\mathbf{A}^T \mathbf{\Lambda} \mathbf{A} = \mathbf{I},$$

where \mathbf{I} is the identity matrix. Consequently, applying the transformations

$$\mathbf{Ka}' = \mathbf{A}^T \mathbf{Ka}$$

and

$$\mathbf{D}' = \mathbf{A}^T \mathbf{D},$$

the original statistical model becomes

$$E[\mathbf{Ka}'] = \mathbf{D}'\mathbf{b}$$

with

$$\text{Var}[\mathbf{Ka}'] = \mathbf{I}\sigma^2,$$

where σ^2 is a constant.

Because this transformed model meets the independence assumptions of ordinary least squares (Searle 1971), the transformed variables \mathbf{Ka}' and \mathbf{D}' can then be used with standard analysis of covariance approaches to test for differences in slope among genes. In particular, we employed the test for homogeneous slopes described by Timm (1975) to test whether \mathbf{b} differed among the three genes examined.

This approach requires an estimate of the expectation of the original variance-covariance matrix Σ . In this matrix, the diagonal element Σ_{ii} is the expected variance of \mathbf{Ka} for species pair i , while the off-diagonal element Σ_{ij} is the expected covariance for \mathbf{Ka} between species pairs i and j . These expected values may be determined as follows: Under the assumption of constant (Poisson) substitution rates (molecular clock), the expected number of substitutions per site, s_k , along any branch k of a phylogeny is proportional to the length of the branch, l_k ; i.e.,

$$s_k = \alpha l_k,$$

where α is the substitution rate constant. Therefore, the expected total number of substitutions per site between species pair i , which is just K_a , is given by

$$\text{Exp}[\mathbf{Ka}_i] = \Sigma_k s_k = \alpha \Sigma_k l_k,$$

where the summation is over all branches k connecting the two species. Under the Poisson assumption, the variance in total number of substitutions separating two species is equal to the mean; i.e.,

$$\text{Var}[\mathbf{Ka}_i] = \text{Var}[\Sigma_k s_k] = \alpha \Sigma_k l_k,$$

i.e., the expected variance in K_a for two species is proportional to the sum of the lengths of the branches separating the two species. Thus, for two pairs of species, i and j , the expected covariance in K_a is, in similar fashion,

$$\text{Cov}[\mathbf{Ka}_i, \mathbf{Ka}_j] = \text{Cov}[\Sigma_k s_k, \Sigma_k s_k] = \text{Var}[\Sigma_k s_k] = \alpha \Sigma_k l_k,$$

where in this case the summation is taken over all branches k that are common to species pair i and species pair j .

Because $\text{Var}[\mathbf{Ka}_i]$ represents the i th diagonal element of Σ and $\text{Cov}[\mathbf{Ka}_i, \mathbf{Ka}_j]$ represents the i,j th off-diagonal element, the expected value of Σ is

$$\Sigma = \alpha \mathbf{L},$$

where element L_{ij} of the matrix \mathbf{L} is the total length of the branches in common between species pairs i and j .

A complication arises in applying this analysis if the data used to estimate genetic distances are not independent of the data used to estimate K_a and K_s , as is true in this study. This lack of independence is manifested in an expected covariance between, say, \mathbf{Ka}_{ij} and genetic distance, D_i , for species pair i and gene j . With sequences of approximately the same length, n , this covariance is

$$\text{Cov}(\mathbf{Ka}_{ij}, D_i) = \text{Cov}(N_{ij}/n\theta, \Sigma_k(N_{ik} + S_{ik})),$$

where θ is the proportion of sites that are nonsynonymous, N_{ij} is the number of nonsynonymous substitutions in gene j between species pair i , and S_{ik} is the number of synonymous substitutions in gene j between species pair i . Because N_{ij} is uncorrelated with S_{ik} for all j and k , and because N_{ij} is uncorrelated with N_{ik} for j not equal to k , this covariance simplifies to

$$\text{Cov}(\mathbf{Ka}_{ij}, D_i) = (1/n\theta)\text{Cov}(N_{ij}, N_{ij}) = (1/n\theta)\text{Var}(N_{ij}).$$

As described above, $\text{Var}(N_{ij}) = \gamma_j l_i$, where γ_j is the rate of nonsynonymous substitution at gene j . Consequently,

$$\text{Cov}(K_{a_{ij}}, D_i) = (1/n\theta)\gamma_j l_i.$$

One way to remove this dependence is to examine the relationship between genetic distance and pairwise between-species differences in K_a . For example, for genes 1 and 2, this pairwise difference is

$$\delta_{i,12} = K_{a_{i1}} - K_{a_{i2}},$$

and the covariance between this difference and genetic distance is then

$$\begin{aligned} \text{Cov}(\delta_{i,12}, D_i) &= \text{Cov}(K_{a_{i1}} - K_{a_{i2}}, D_i) \\ &= (1/n\theta)\gamma_1 l_i - (1/n\theta)\gamma_2 l_i. \end{aligned}$$

Under the null hypothesis that the rates of non-synonymous substitutions are equal for genes 1 and 2 (i.e., that K_a is the same for the two genes), $\gamma_1 = \gamma_2$ and $\text{Cov}(\delta_{i,12}, D_i) = 0$. This null hypothesis may thus be rejected if there is a significant correlation between $\delta_{i,12}$ and genetic distance. A similar argument, substituting $(1 - \theta)$ for θ , applies to the covariance between genetic distance and the difference in K_s between two genes.

Acknowledgments

We thank the centers of plant genetic resource units of USDA for seeds of *I. nil* and *I. trifida*, Joel Kniskern for providing *I. purpurea* seeds, and Rick Miller for providing genomic DNA of *Ipomoea purpurea*. The work was supported by National Science Foundation grant MCB 0110596.

Literature Cited

- Akashi, H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* **136**:927–935.
- Bennetzen, J. L., and B. D. Hall. 1982. Codon selection in yeast. *J. Biol. Chem.* **257**:3026–3031.
- Betancourt, A. J., and D. C. Presgraves. 2002. Linkage limits the power of natural selection in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **99**:13616–13620.
- Bielawski, J. P., and Z. Yang. 2001. Positive and negative selection in the DAZ gene family. *Mol. Biol. Evol.* **18**:523–529.
- Chiapello, H., F. Lisacek, M. Caboche, and A. Henaut. 1998. Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. *Gene* **209**:GC1–GC38.
- Cameron, J. M., and M. Kreitman. 1998. The correlation between synonymous and nonsynonymous substitutions in *Drosophila*: mutation, selection, or relaxed constraints. *Genetics* **150**:767–775.
- Durbin, M. L., G. H. Learn, G. Huttley, and M. T. Clegg. 1995. Evolution of the chalcone synthase gene family in the genus *Ipomoea*. *Proc. Natl. Acad. Sci. USA* **92**:3338–3342.
- Durbin, M. L., B. McCaig, and M. T. Clegg. 2000. Molecular evolution of the chalcone synthase multigene family in the morning glory genome. *Plant Mol. Biol.* **42**:79–92.
- Edwards, A. F. W. 1972. *Likelihood*. Cambridge University Press, Cambridge.
- Felsenstein, J. 1984. DNAML in PHYLIP 2.6. University of Washington, Seattle.
- . 1985. Phylogenies and the comparative method. *Am. Nat.* **125**:1–15.
- Fennoy, S. L., and J. Bailey-Seres. 1993. Synonymous codon usage in *Zea mays* L. nuclear genes is varied by levels of C and G-ending codons. *Nucleic Acids Res.* **21**:5294–5300.
- Fitch, D. H. A., and L. D. Strusbaugh. 1993. Low codon bias and high rates of synonymous substitution in *Drosophila hydei* and *Drosophila melanogaster* histone genes. *Mol. Biol. Evol.* **10**:397–413.
- Foster, P. L., E. Eisenstadt, and J. Cairns. 1982. Random components in mutagenesis. *Nature* **299**:365–367.
- Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**:160–174.
- Huang, J. C., and M. Sun. 2000. Genetic diversity and relationships of sweetpotato and its wild relatives in *Ipomoea* series *Batatas* (Convolvulaceae) as revealed by inter-simple sequence repeat (ISSR) and restriction analysis of chloroplast DNA. *Theor. Appl. Genet.* **100**:1050–1060.
- Huttley, G. A., M. L. Durbin, D. E. Glover, and M. T. Clegg. 1997. Nucleotide polymorphism in the chalcone synthase—A locus and evolution of the chalcone synthase multigene family of common morning glory *Ipomoea purpurea*. *Mol. Ecol.* **6**:549–558.
- Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2**:13–34.
- Kimura, M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**:275–276.
- Koes, R. E., F. Quattrocchio, and J. N. M. Mol. 1994. The flavonoid biosynthetic pathway in plants: function and evolution. *BioEssays* **16**:123–132.
- Kusumi, J., Y. Tsumura, H. Yoshimaru, and H. Tachida. 2002. Molecular evolution of nuclear genes in Cupressaceae, a group of conifer trees. *Mol. Biol. Evol.* **19**:736–747.
- Li, W.-H. 1997. *Molecular evolution*. Sinauer Associates, Sunderland, Mass.
- Li, W.-H., C.-I. Wu, and C.-C. Luo. 1985. A new method of estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2**:150–174.
- Lu, Y. 2002. Molecular evolution at the self-incompatibility locus of *Physalis longifolia* (Solanaceae). *J. Mol. Evol.* **54**:784–793.
- Manos, P. S., R. E. Miller, and M. D. Rausher. 2001. Phylogenetic analysis of *Ipomoea*, *Argyrea*, *Stictocardia*, and *Turbina* suggests a generalized model of morphological evolution in morning glories. *Syst. Bot.* **26**:585–602.
- Martin, E. P., and T. F. Hansen. 1997. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am. Nat.* **149**:646–667.
- McDonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the *ADH* locus in *Drosophila*. *Nature* **351**:652–654.
- Miller, R. E., M. D. Rausher, and P. S. Manos. 1999. Phylogenetic systematics of *Ipomoea* (Convolvulaceae) based on ITS and Waxy sequences. *Syst. Bot.* **24**:209–227.
- Mondragón-Palomino, M., B. C. Meyers, R. W. Michelmore, and B. S. Gaut. 2002. Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. *Genome Res.* **12**:1305–1315.
- Powell, J. R., and E. N. Moriyama. 1997. Evolution of codon usage bias in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **94**:7784–7790.

- Rausher, M. D., R. E. Miller, and P. Tiffin. 1999. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. *Mol. Biol. Evol.* **16**:266–274.
- Rozas, J., and R. Rozas. 1999. DnaSP version3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**:174–175.
- Sakuda, M. 2000. Transcriptional control of chalcone synthase by environmental stimuli. *J. Plant Res.* **113**:327–333.
- Searle, S. R. 1971. *Linear models*. Wiley, New York.
- Sharp, P. M., and W.-H. Li. 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol. Biol. Evol.* **4**:222–230.
- Shirley, B. W. 1996. Flavonoid biosynthesis: “new” functions for an “old” pathway. *Trends Plant Sci.* **1**:377–382.
- Swanson, W. J., Z. Yang, M. F. Wolfner, and C. F. Aquadro. 2001. Positive Darwinian selection derives the evolution of several female reproductive proteins in mammals. *Proc. Natl. Acad. Sci. USA* **98**:2509–2514.
- Timm, N. H. 1975. *Multivariate analysis with applications in education and psychology*. Books/Cole Publishing, Monterey, Calif.
- Uyenoyama, M. K. 1995. A generalized least-squares estimate for the origin of sporophytic self-incompatibility. *Genetics* **139**:975–992.
- Weir, B. S. 1990. *Genetic fate analysis*. Sinauer Associates, Sunderland, Mass.
- Whitefield, L., R. Lovell-Badge, and P. Goodfellow. 1993. Rapid sequence evolution of the sex-determining gene SRY. *Nature* **364**:713–715.
- Wright, F. 1990. The “effective number of codons” used in a gene. *Gene* **87**:23–29.
- Wyckoff, G. J., W. Wang, and C.-I. Wu. 2000. Rapid evolution of male reproductive genes in the descent of man. *Nature* **403**:304–309.
- Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**: 555–556.
- Yang, Z., R. Nielsen, N. Goldman, and A. K. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**:431–449.
- Zhang, L., T. J. Vision, and B. S. Gaut. 2002. Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. *Mol. Biol. Evol.* **19**:1464–1473.

Brandon Gaut, Associate Editor

Accepted June 11, 2003