J Mol Evol (2004) 59:416–425
DOI: 10.1007/s00239-004-2632-9

JOURNAL OF MOLECULAR EVOLUTION

Molecular Adaptation in Plant Hemoglobin, a Duplicated Gene Involved in Plant–Bacteria Symbiosis

Emilie Guldner, Bernard Godelle, Nicolas Galtier

CNRS UMR 5171— "Génome, Populations, Interactions, Adaptation," Université Montpellier, 2—CC63, Place E. Bataillon, 34095 Montpellier, France

Received: 19 December 2003 / Accepted: 26 March 2004 [Reviewing Editor: Dr. Rasmus Nielsen]

Abstract. The evolutionary history of the hemoglobin gene family in angiosperms is unusual in that it involves two mechanisms known for potentially generating molecular adaptation: gene duplication and among-species interaction. In plants able to achieve symbiosis with nitrogen-fixing bacteria, class 2 hemoglobin is expressed at high concentrations in nodules and appears to be a key factor for the achievement and regulation of the symbiotic exchange. In this study, we make use of codon models of DNA sequence evolution with the goal of determining the nature of the selective forces which have driven the evolution of this gene. Our results suggest that adaptive evolution occurred during the period of time following the duplication event (functional divergence) and that a change in the selective pressures arose in class 2 hemoglobin in relation to the acquisition of a symbiotic function.

Key words: Symbiotic hemoglobin — Legumes — Positive selection — Duplication — Molecular adaptation — Codon models — Covarion

Introduction

The emergence of a new, advantageous function in a DNA or protein sequence must occur trough advantageous mutations fixed by natural selection (called positive selection). Documented instances of adaptive evolution at the molecular level are somewhat rare, however, and mostly restricted to immune system genes (Hughes and Nei 1988, Hughes and Nei 1989) and male reproductive genes (Biermann 1998, Civetta and Singh 1998, Wyckoff et al. 2000, Swanson and Vacquier 2002) in animals. In this paper we study the symbiotic hemoglobin gene of legumes, whose evolutionary history is marked by processes potentially generating positive selection at the molecular level.

In angiosperms, an ancient gene duplication provided two classes of hemoglobins distinct in their function and pattern of expression (Jacobsen-Lyon et al. 1995, Trevaskis et al. 1997, Hunt et al. 2001). In plants able to achieve symbiosis with nitrogen-fixing bacteria, i.e., in most legume species (symbiosis with *Rhizobium* and *Bradyrhizobium*) and in several species from Fagales, Rosales, and Cucurbitales (symbiosis with *Frankia*), class 2 hemoglobins (symbiotic hemoglobins, called leghemoglobins in legumes) are remarkable for being specifically expressed at high concentrations in nodules. In these symbiosis-specific structures, bacteria fix nitrogen thanks to the nitrogenase enzyme, which requires substantial amounts of energy in the form of ATP produced by bacterial respiration. However, oxygen, which is necessary for respiration, readily inhibits the activity of nitrogenase. Class 2 hemoglobin, owing to its extremely fast O_2 association rate and rather slow O_2 dissociation rate, facilitates oxygen diffusion to the symbionts (Appleby 1992) and contributes to the

Correspondence to: Nicolas Galtier; *email:* galtier@univ-montp2.fr

maintenance of an oxygen-free environment around the enzyme.

Thus, hemoglobin allows the proper functioning of the symbiotic exchange. Furthermore, as nitrogenase activity is limited by oxygen supply (Miller et al. 1988), hemoglobin is expected to play a significant role for its regulation in case of nodule environment variations (pH, temperature [Miller et al. 1988]), when plants are stressed, or when nodules senesce. Modifications in the binding properties of hemoglobin have been observed which appear to imply a decrease in the flux of O_2 to the bacteroids (Wagner and Sarath 1987; Jun et al. 1994). Also, the binding of hemoglobin with NO inhibits the oxygen-carrying activity of hemoglobin and appears to have deleterious effects on the functional activity of root nodules (Davies and Puppo 1992). A nitric oxide synthase activity that converts arginin to NO has been reported in roots and nodules (Cueto et al. 1996). Symbiotic hemoglobin, therefore, might be a key factor in a possible conflict of interest between plant and bacteria concerning nodule activity and growth. For this reason, it is a promising candidate to exemplify positive selection at the sequence level in the context of recurrent adaptation to a changing interacting species (Red Queen hypothesis).

A standard method for detecting positive selection is the comparison of synonymous (silent *dS*) and nonsynonymous (amino acid replacing *dN*) substitution rates. Positive selection at the protein level must be invoked if the $\omega = dN/dS$ ratio is higher than one. In a preliminary study of the evolutionary history of plant hemoglobin genes, Guldner et al. (2004) reported a higher *dN*/*dS* ratio in class 2 symbiotic genes of legumes than in class 1 genes, suggesting that not all plant hemoglobins have evolved under the same selective pressures. The observed *dN*/*dS* ratio, however, was lower than one, making the interpretation difficult: a relaxing of functional constraints in symbiotic genes, not positive selection, could be the cause of the observed increase in nonsynonymous substitution rate.

In the analysis by Guldner et al. (2004), the *dN*/*dS* ratio was averaged over all codons of the protein, and over lineages. Adaptive changes, however, probably involve a small number of sites during relatively short periods of time (Zhang 2003). Averaging *dN*/*dS* over sites and lineages makes difficult the detection of episodes of adaptive evolution, as the adaptive signal is diluted by the prevalent purifying selection. In this paper, we make use of elaborate models of coding sequence evolution allowing site-specific and lineage-specific variations of the selective pressure (Yang 1998; Yang and Nielsen 2000). A method based on the covarion model (Galtier 2001; Pupko and Galtier 2002) is also used with the specific aim of detecting adaptive episodes at the amino acid level. Our aim is to clarify the role played by hemoglobin during the evolution of the symbiosis between angiosperms and nitrogen-fixing bacteria.

Data Analysis

Data Set

The phylogenetic tree of plant hemoglobin genes (50 sequences) was constructed from amino acid sequences following Guldner et al. (2004) (Fig. 1). A robust rooting of the tree, required for an unambiguous understanding of the molecular evolution of the gene, was obtained thanks to the isolation of two hemoglobin genes in the deeply branching *Euryale ferox* (Nymphaeaceae [Guldner et al. 2004]). A tree topology was first reconstructed from protein sequences using a bayesian analysis (Huelsenbeck and Ronquist 2001) using the JTT + gamma model of amino acid evolution, then little-resolved subtrees were slightly modified by hand to make the tree consistent with the canonical angiosperm phylogeny (Savolainen et al. 2000). Plant hemoglobin genes can be distinguished by their class (class 1 vs. class 2) or by their function (symbiotic vs. nonsymbiotic). With the exception of the gene from *Parasponia* (Ulmaceae), which plays a role in symbiotic and nonsymbiotic tissues, all class 1 genes are nonsymbiotic. Within class 2, genes from legumes (leghemoglobins) and *Casuarina* (Casuarinaceae) are symbiotic, while the others are not.

Codon Models

All statistical analyses were performed on DNA coding sequences using the codeml program in the PAML package (Yang 1997). These analyses involve fitting various models of codon sequence evolution, all based on the model proposed by Goldman and Yang (1994), by the maximum likelihood method. These models include a parameter of interest, ω, which measures the ratio of nonsynonymous-to-synonymous evolutionary rate. Other parameters such as branch lengths and transition/transversion ratio are reestimated separately for each analysis but are considered essentially as nuisance parameters in this study.

Models of Variable Selective Pressures Among Lineages

We first applied models that allow for different *dN*/*dS* ratios among evolutionary lineages (Yang 1998; Bielawski and Yang 2003). Arbitrarily chosen subtrees are assigned a distinct ω parameter, which can be estimated from the data. Models differ by the number and nature of such subtrees. We assessed the

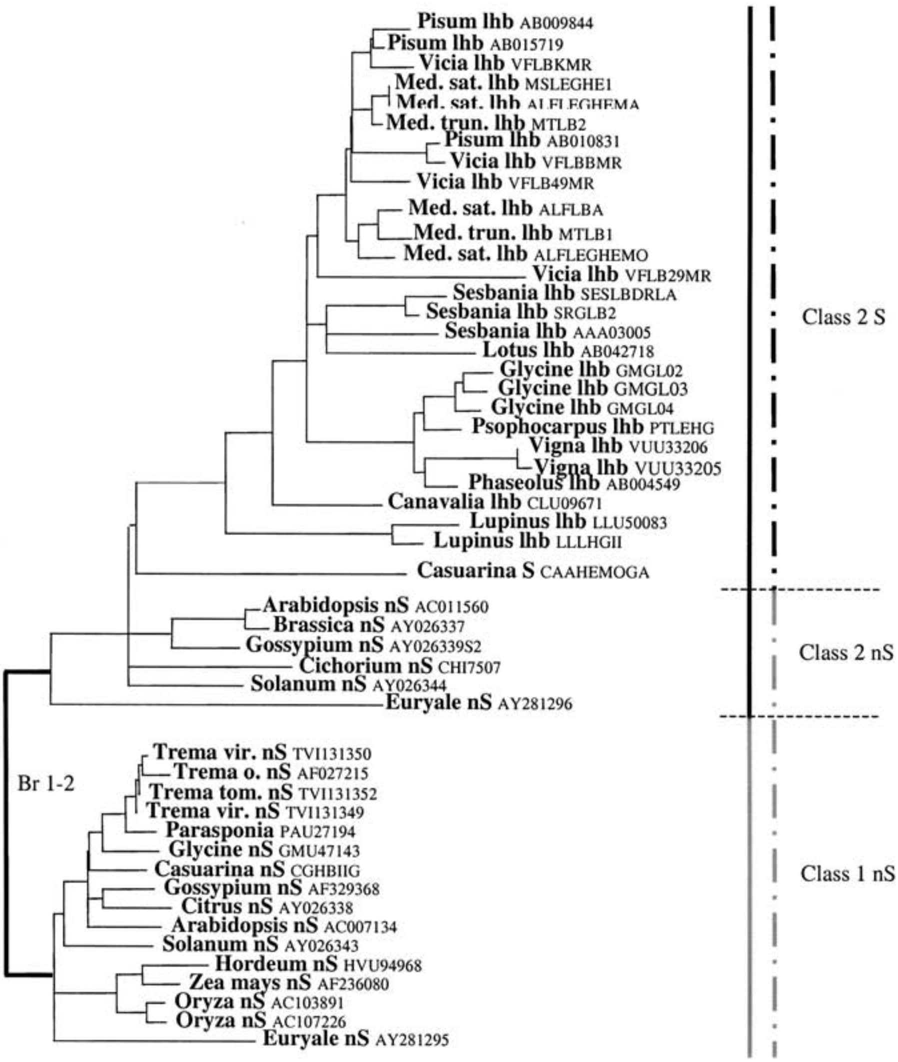

Fig. 1. Hemoglobin gene phylogeny. Fifty species, 132 amino acid sites. Branch lengths were estimated by fitting PAM distances between amino acid sequences to a modified hemoglobin tree topology (see Guldner et al. 2004). These genes can be distinguished by their class (gray line, class 1 genes: black line, class 2 genes) or by their function (dotted gray line, nonsymbiotic genes; dotted black line, symbiotic genes). Br1-2 designates the branch immediately following the duplication event. 1hb, leghemoglobins; S, symbiotic gene; nS, nonsymbiotic gene.

relevance of various evolutionary hypotheses by comparing the maximum likelihood of competing models through likelihood ratio tests (LRT).

By defining the appropriate subtrees, we tested whether hemoglobins of class 1 have evolved under similar selection pressures as hemoglobins of class 2 and whether the acquisition of symbiotic function by class 2 genes was accompanied by a change in selection pressures. The four-ratio model combines the two hypotheses in assuming a different ω ratio for nonsymbiotic class 1 genes, nonsymbiotic class 2 genes, and symbiotic class 2 genes, respectively. In addition, some of the models assume a distinct ω ratio for the branch immediately following the duplication event (noted Br1-2), which allows testing for a functional divergence postduplication.

Results are given Table 1. Each column in this table is for a distinct model. The first four lines give the assumptions of each model, that is, which subtrees were assigned a distinct ω parameter. For example, the three-ratio model assumes a specific ω for the branch separating class 1 from class 2 hemoglobins, one for the class 1, and one for the class 2, irrespective of their symbiotic status. Line 5 gives the maximum likelihood under the considered model, and the next lines give LRT values corresponding to comparisons with competing models.

Let us first consider models defined in order to know if the change in selection pressures depends on the class and/or the function of genes. The LRT comparing the four-ratio model with the three-ratio model is highly significant (LR = 28.13, χ^2 1 df, $p < 10^{-6}$), which means that within the class 2 genes, the nonsymbiotic and the symbiotic genes have not evolved under the same selection pressures. The LRT comparing the four-ratio model with the three-ratio-S model, how-

Table 1. Codon models of variable selective pressures among lineages

	One-ratio	Two-ratio	Three-ratio	Four-ratio	Two-ratio-L	Two-ratio-S	Three-ratio-S
ω_0		Br1-2	Br1-2	Br1-2	nS + *Cas*	nS	Br1-2
ω_1		C1 + C2	C1	C1	lhb	lhb + *Cas*	ns
ω_2			C2	C2 nS			S
ω_3				C2S			
Ln*L*	−11,688.939	−11.687.043	−11,677.69	−11,663.634	−11.666.386	−11,663.954	−11,663.694
LRT/one-ratio		3.79	22.48	50.6	45.1	49.97	50.49
LRT/two-ratios			18.69	46.82			46.7
LRT/three-ratios				28.13			
LRT/two-ratios-L				5.5			5.38
LRT/two-ratios-S				0.64			0.52
LRT/three-ratios-S				0.12			

Note. Analyses are applied on (a) the whole data set (53 sequences, 207 codons, (b) class 2 genes (35 sequences, 153 codons), and (c) non-symbiotic genes of both classes (22 sequences, 211 codons). Br1-2 refers to the branch just following the duplication event (i.e., which separates the two groups of class 1 and class 2 genes). C1, class 1 genes; C2 class 2 genes; nS, nonsymbiotic genes; S, symbiotic genes; lhb, leghemoglobins; *Cas*, symbiotic gene of *Casuarina*. The four-ratio model assumes a different ω ratio for Br1-2 (ω_0), the genes of class 1 (ω_1), the nonsymbiotic genes of class 2 (ω_2), and the symbiotic genes of class 2 (ω_3).

ever, is not significant (LR = 0.12), showing that it is not relevant to assume different ω ratios for class 1 genes and class 2 nonysymbiotic genes. Thus, all nonsymbiotic genes, whatever their class, appear to be constrained by similar selective pressures, and these selective pressures are different from those driving the evolution of symbiotic class 2 genes. The acquisition of symbiosis, not the duplication, appears to have significantly modified the evolutionary dynamics of hemoglobin in plants. The estimated ω parameter, however, is much lower than one for both nonsymbiotic (0.16) and symbiotic (0.33) genes (four-ratio model; other models yielded very close estimates).

The class 2 gene of *Casuarina* is the only nonlegume symbiotic gene of the data set. *Casuarina* is involved in symbiosis with *Frankia*, a nitrogen-fixing bacterium different from *Rhizobium*. It is plausible that the above-reported change in selection pressures characterizes only the legume clade, a result that would lower our argument about a possible impact of symbiosis acquisition on hemoglobin evolution. The two-ratio-L model and the two-ratio-S model were designed to check this hypothesis. Both compare the evolutionary processes of nonsymbiotic genes to that of symbiotic genes but differ with respect to the position of *Casuarina: Casuarina* is grouped with the nonsymbiotic genes in the two-ratio-L model but with legumes in the two-ratio-S model. These two models have the same number of parameters and cannot be compared by a LRT. However, the maximum likelihood value of the two-ratio-S model was higher than the maximum likelihood value of the two-ratio-L model, suggesting that the evolutionary process followed by *Casuarina* symbiotic hemoglobin is closer to that of (symbiotic) legume genes than to that of non-symbiotic genes. This result is in agreement with the hypothesis that the observed change in the selective pressures should be associated with the acquisition of symbiosis.

We performed similar analyses using sequence subsets (results not shown). Within class 2 genes, a model separating symbiotic vs. nonsymbiotic sequences was found optimal (LR = 22.05, $p < 10^{-5}$). Within nonsymbiotic hemoglobins, complex models did not significantly improve the fit compared to the one-ratio model, suggesting that the evolutionary rates of nonsymbiotic genes of class 1 and nonsymbiotic genes of class 2 are similar. Once again, these results provide evidence that the symbiotic/nonsymbiotic function of a gene, not its class, determines its evolutionary process.

Finally, these analyses suggest that the branches immediately following the duplication event might have undergone a specific process characterized by a high ω. The two-ratio model is marginally significantly better from the one-ratio model (LR = 3.79, 1 df, $p = 0.052$). The three-ratio-S model is not significantly different from the two-ratio-S model. However, the ω estimates for this branch (42.24 and 17.8, respectively, for the two-ratio model and the three-ratio-S model) are particularly high. In both models the *dN* for this branch (0.24) is higher than the *dS* (0.0056 and 0.013, respectively), which is indicative of positive selection. Despite the lack of strict significance at the 0.05 level, we tend to interpret these estimates as evidence for the occurrence of an adaptive episode after the duplication event. However, the very low *dS* estimates for this branch suggest that saturation of synonymous substitutions might occur, possibly biasing the adaptive signal upward.

Models of Variable Selective Pressures Among Sites. The above analyses suggest that the acquisition of symbiosis resulted in a twofold increase in ω in class 2 hemoglobin. However, they do not demonstrate that the new selective regime involves molecular adaptation, because the estimated ω is lower

Table 2. Codon models allowing variable selective pressures among sites, class 1 genes

Model	Estimated parameter	ln*L*
One-ratio	$dN/dS = 0.124$	−3,938.367
Neutral	$\omega_0 = 0$ $(p_0 = 0.314)$	−4,080.109
	$\omega_1 = 1$ $(p_1 = 0.686)$	
Selection	$\omega_0 = 0$ $(p_0 = 0.236)$	−3,847.439
	$\omega_1 = 1$ $(p_1 = 0.126)$	
	$\omega_2 = 0.124$ $(p_2 = 0.638)$	(LRT/neutral = 465.34)
"Discrete"	$\omega_0 = 0.035$ $(p_0 = 0.578)$	−3,838.503
	$\omega_1 = 0.228$ $(p_1 = 0.33)$	
	$\omega_2 = 0.971$ $(p_2 = 0.092)$	(LRT/one-ratio = 199.728)
Beta	$p = 0.479$, $q = 2.263$	−3,844.241
Beta & ω	$p = 0.752$, $q = 5.942$	−3,838.378
	$p_0 = 0.916$	
	$\omega_1 = 0.997$ $(p_1 = 0.084)$	(LRT/beta = 11.726)

Note. Models are described in Yang et al. (2000) (and see text).

Table 3. Codon models allowing variable selective pressures among sites, symbiotic genes

Model	Estimated parameter	ln*L*
One-ratio	$dN/dS = 0.29$	−5,443.006
Neutral	$\omega_0 = 0$ $(p_0 = 0.17)$	−5,525.544
	$\omega_1 = 1$ $(p_1 = 0.83)$	
Selection	$\omega_0 = 0$ $(\text{p}_0 = 0.168)$	−5,515.967
	$\omega_1 = 1$ $(p_1 = 0.771)$	
	$\omega_2 = 0.271$ $(p_2 = 0.061)$	(LRT/neutral = 19.154)
"Discrete"	$\omega_0 = 0.08$ $(p_0 = 0.516)$	−5,314.664
	$\omega_1 = 0.455$ $(p_1 = 0.35)$	
	$\omega_2 = 1.042$ $(p_2 = 0.133)$	(LRT/one-ratio = 256.684)
Beta	$p = 0.62$, $q = 1.25$	−5,320.998
Beta & ω	$p = 0.67$, $q = 1.465$	−5,318.512
	$p_0 = 0.984$	
	$\omega_1 = 1.972$ $(p_1 = 0.016)$	(LRT/beta = 4.972)

than one. This does not imply that adaptation is not occurring since the estimated ω in these models is averaged over all sites of the gene, including strongly constrained sites. To try and discriminate amino acid sites potentially undergoing adaptation from amino acid sites under purifying selection, we used models that account for variation of ω among sites (Nielsen and Yang 1998; Yang et al. 2000a). Such models assume a discrete distribution for ω across sites with a fixed number of classes, the proportion of each class and its characteristic ω being parameters (either fixed a priori or estimated from the data). The neutral model assumes two classes of sites (sites under purifying selection, i.e., $\omega = 0$, and neutral sites, i.e., $\omega = 1$), the selection model adds a third class of sites with a free-to-vary ω, and the "discrete" model assumes three site classes, whose relative proportions and ω values are free. Finally, the beta model assumes that ω ratios are distributed among sites according to a beta distribution (10 site classes), and the beta & ω model is an extension of the beta model, having an extra class of sites with an independent, free ω ratio. We applied these models separately for the two groups of class 1 and class 2 symbiotic genes, respectively. Again, models were compared by likelihood ratio tests. Results are presented in Table 2 (class 1 genes) and Table 3 (symbiotic genes). In both groups, the "discrete" model was significantly better than the one-ratio model. This shows that selective pressures are variable among sites. For class 1 genes, the selection model was significantly better than the neutral model, and the beta and ω model was significantly better than the beta model. According to the selection model, 87.4% of sites evolve under purifying selection, whereas 12.6% of sites are neutral. The "discrete" model showed a similar tendency. The free ω parameter is equal to 0.12 when estimated by the selection model and equal to 0.97 and 0.99 when estimated by the "discrete" model and the beta & ω model, respectively. In every case, the freely estimated ω parameter was not higher than one, which does not suggest an effect of positive selection.

For the group of symbiotic genes, the LRT between the selection model and the neutral model, on the one hand, and between the beta & ω model and the beta model, on the other hand, were significant.

	.	....|....|....|....|....|....|	.|....|
Class 1	2	15 25 35	105
Oryza nS AC103891	E	SWAILKKDSANIALRFFLKIFEVAPSASQM	HFEVVKF
Glycine nS GMU47143	E	SWNVMKKNSGELGLKFFLKIFEIAPSAQKL	HFEVTKF
Casuarina nS CGHBIIG	E	SWSAMKPNAGELGLKFFLKIFEIAPSAQKL	HFEVTKF
Solanum nS AY026343	E	SWGSMKKDAGEWGLKFFLKIFEIAPSAKKM	HFEVTKY
Euryale nS AY281295	E	SWGVMKKDAGQLGVKFFAKIFEIAPSAKRM	HFD-VRF
Class 2			
Cichoriun nS CHI7507	E	SWEVMKQDIPALSLYLYAMILEIAPEARGL	QFEVVKE
Brassica nS AY026337	E	SWEILKQDIPKYSLHFFSQILEIAPAAKDM	HFEVVKE
Solanum nS AY026344	D	SWEFMKQDIPQLSLRFFSLILEIAPVAKNM	HFEVVKE
Euryale nS AY281296	E	AWELMKPNVQELGLKMFFLSWADVAGGAGM	HFEVIKV
Casuarina S CAAHEMOGA	E	SWEVLKQNIPAHSLRLFALILEAAPESKYV	HFEVMKG
Lupin lhb LLU50083	D	SFEEFNANIPKNTHRFFTLVLEIAPGAKDL	HFPVVKE
Glycine lhb GMGL03	E	SFEAFKANIPQYSVVFYTSILEKAPAAKDL	QFVVVKE
Vicia lhb VFLB49MR	Q	SWESFKQN-PSYSVLFYTIILEKAPAAKGM	HFVVVKE
Pisum lhb AB009844	D	SWELFKQN-PGYSVLFYNIILKKAPATKGM	HFVVVKE
Med. trun. Lhb MTLB1	D	SYEAFKQNLSGYSVFFYTVILEKAPAAKGL	HFVVVKE

Fig. 2. Amino acid alignment. Sites 12 and 21, sites assigned to a category with ω > 1 with codon model. Sites 13 and 109, covarion sites ($p = 0.001$) with the model *class*. Site 2, covarion site (p = 0.001) with the model *function*. *n*S nonsymbiotic genes; S, symbiotic genes; lhb, leghemoglobins.

The freely estimated ω parameter was higher than one in both cases, suggesting that some sites of symbiotic hemoglobins are undergoing positive selection. This, however, seems to concern only a small proportion of sites (6% under the selection model, < 2% under the beta & ω model). Sites were assigned to ω categories according to their empirical bayesian posterior probabilities (Yang et al. 2000b). Two sites were unambiguously assigned to a class with $\omega > 1$, both under the selection and beta + ω model: sites 12 and 21 (Fig. 2).

Covarion Test

The covarion model (Galtier 2001; Pupko and Galtier 2002) aims at representing the notion that a site evolves slowly when it is functionally important but changes rapidly when neutral. Thus, a functional innovation can be detected from the phylogenetic history of a protein by seeking a lineage in which several amino acid positions show a significant increase, or decrease, in the evolutionary rate. For a given amino acid site, highly different rates in the two subgroups potentially reveal a change in functional constraints. Several "cutting points" (defining the two subtrees between which site-specific rate variation is sought) were tried, with the aim of detecting changes in functional constraints (i.e., in the evolutionary rate) associated with the duplication event (class 1 vs. class 2) or the acquisition of the symbiotic function (nonsymbiotic vs. symbiotic). For each amino acid site, a likelihood ratio test was performed according to the method developed by Pupko and Galtier (2002). Two models are contrasted: the null model assumes a constant evolutionary rate across the tree for the considered site, while the alternative hypothesis allows one specific rates for each of the two subtrees. To cope with the problem of multiple testing (one test per site), the significance of the number of sites showing a covarion-like pattern was assessed as in Pupko and Galtier (2002): under the hypothesis of no covarion, and assuming independent sites, this number should follow a Binomial distribution B(n, p) where n is the number of sites analysed and p the type I error of the LRT. The JTT model of amino acid substitution was used (Jones et al. 1992). A discretized gamma distribution of rates across sites (four categories) was assumed.

Table 4. Covarion test

	Group 1	Group 2	*P*	Covarion sites	Binomial test
Class	Class 1	Class 2	0.05	**15** (E2, E3, Q4, W12, N21, I22, S38, T59, M62, M67, C69, A73, R76, A78, F1 10)	**
			0.01	**1** (P127)	
			0.005	**2** (K56,V83)	
			0.001	**2** (A13, K109)	**
Class_nS	nS class 1	nS class 2	0.05	**7** (A13, S36, S47, P50, M62, G101, L113)	
			0.01	**1** (K117)	
			0.005	**0**	
			0.001	**0**	
Function	nS	S	0.05	**7** (W12, A23, F26, M62, K109, P121, W133)	
			0.01	**3** (A13, A73, V108)	
			0.005	**2** (F110, S134)	
			0.001	**1** (E2)	**
Function_c2	class 2 nS	class 2 S	0.05	**13** (Q39, LSI,' K53, C69, D85, T87, G101, E106, V108, F1 10, K117, P121, W125)	**
			0.01	**3** (E2,W12,F26)	
			0.005	**0**	

Note: Number and nature of detected covarion sites, for various splits of the data set. nS, nonsymbiotic genes; S, symbiotic genes; Cr, number of covarion positions detected by the program. In parentheses are indicated the nature of covarion sites, using *Oryza* sequence (accession number AC103891) as reference. p = type I error.**The number of detected covarion sites is significant at the 99% level. Positions of covarion sites are given using *Oryza sativa* hemoglobin 1 sequence as reference (accession number AC103891).

Results are presented in Table 4. The model class compares class 1 and class 2 genes. It found 2 significant ($p < 0.001$) covarion positions (sites 13 and 109, which are variable in class 1 but very conserved in class 2 genes), and 15 with a p value of 0.05. The numbers of covarion positions detected with a p value of 0.005 and 0.01 (two and one, respectively) were not significant. The model class-nS, which compares groups of nonsymbiotic genes of class 1 and nonsymbiotic genes of class 2, found no significant results. The model function compares the two subgroups of nonsymbiotic and symbiotic genes, respectively. It detected 13 covarion positions, with only 1 significant ($p < 0.001$). This corresponds to the second site, which is conserved in nonsymbiotic sequences but variable in symbiotic sequences. The model function_c2, which compares nonsymbiotic and symbiotic genes of class 2, detected 16 covarion positions, but with a low level of significance. A visual inspection of detected sites revealed various patterns of rate increase/decrease. For example, site 73 is conserved in class 1 genes (Val), different but conserved in class 2 nonsymbiotic genes (Ile), and very variable for class 2 symbiotic genes. Site 62 was variable for class 1 genes (Met, Thr, Val, Lys, Leu), and conserved in class 2 genes (Val in nonsymbiotic hemoglobins, but Glu for most of symbiotic sequences). It should be noted that no site was detected as significantly slower (i.e., constrained) in symbiotic than in nonsymbiotic genes. These results are best interpreted in the light of structural data. The quaternary structure plays an important role in the function of many hemoglobins by facilitating allostery and cooperativity for regulation of ligand binding. The dimer interface region is composed of the amino acid residues at the beginning of the CD helix region (residues 32 to 38) and the G helix (residues 104 to 113) in rice (Hargrove et al. 2000; Goodman and Hargrove 2001). These amino acid residues are relatively conserved among nonsymbiotic hemoglobins (Arredondo-Peter et al. 1998: Hargrove et al. 2000), suggesting that the quaternary structures of other members of this group could be similar. The interaction between the two subunits is mediated by a hydrogen bond between the side chain of Glu106 of one subunit and Ser36 on the other. These two polar interactions surround a hydrophobic core consisting of the side chains of Val33, Vall07, and Phel10 of each subunit. The protein coded by the nonsymbiotic class 2 gene of *Arabidopsis*, in contrast, is predicted not to achieve dimerization because of an Ala, not Ser, residue at position 36 (Goodman and Hargrove 2001). The soybean leghemoglobin is also monomeric. Presumably, its dimerization is prevented by the presence of hydrophobic residues at positions 36 and 106. Some of the residues composing the dimer interface region were detected as covarion sites. Site 36 is very conserved (serine) in class 1, variable in class 2 nonsymbiotic genes (hydrophilic Gly and Glu, hydrophobic Ala and Val), and relatively conserved in class 2 symbiotic genes (Ala). The site 106 is hydrophilic (Glu) in nonsymbiotic genes but hydrophobic (mostly Val) in symbiotic ones.

The covarion analysis, therefore, is in agreement with, and somewhat corroborates, inferences that could be made from crystallography data about structural evolution in class 1 vs. class 2 hemoglobin. It did not allow, however, detection of any episode of adaptation associated with the acquisition of symbiosis.

Discussion

The evolutionary history of the plant hemoglobin gene family is unusual in that it involves two mechanisms known for generating molecular adaptation: gene duplication and between-species interaction. In this study, we applied elaborate models of molecular evolution with the goal of determining the nature of the selective forces underlying the evolution of these sequences. Ohta (1988) proposed that after gene duplication, natural selection should favor fixation of mutations that lead to the adaptation of one (or both) copy to a new function. Once the new function is established, positive selection stops, and purifying selection acts to maintain it. For protein coding genes, this means that nonsynonymous substitutions should be accelerated following the duplication (Li et al. 1985; Lynch and Conery 2000) and then slow down due to an increased effect of purifying selection.

Plant hemoglobin genes apparently match this prediction in that the two branches immediately following the duplication event show a *dN* value higher than the *dS* value. Then the nonsymbiotic hemoglobin evolved under purifying selection in both class 1 and class 2. In *Casuarina* and in legumes, class 2 hemoglobin was recruited for a symbiotic function, and this resulted in a significant increase of the *dN*/*dS* ratio.

Only in symbiotic hemoglobin genes did we detect evidence for adaptation when models allowing variable ω across sites were used. The posterior assignment of sites to categories, however, detected only two positively selected positions in symbiotic hemoglobins. Site 12 is a Trp in most species, and because this amino acid is encoded by a single codon, any substitution of Trp must be nonsynonymous. When a (necessarily nonsynonymous) substitution occurred at site 12, Trp was replaced by Tyr or Phe, two aromatic, voluminous amino acids as well. Such changes are not expected to dramatically change the structure and function of the protein. This site is thus likely to have evolved under purifying, not positive, selection. That single-codon amino acids induce false-positive detection might be a general feature of methods based on the Goldman–Yang codon model, highlighting the need for a visual inspection of sites detected as positively selected. Site 21, in contrast, is highly variable and shows amino acids of different chemical properties. It is a candidate for having evolved under recurrent positive selection in symbiotic hemoglobins. More research about its role in the structure and function of hemoglobin is needed in order to know whether, and how, these variations relate to adaptation. The covarion approach did not provide additional evidence for symbiosis-associated positive selection in class 2 hemoglobin. Overall, these results are in agreement with the hypothesis of a Red Queen-like evolution of hemoglobin in symbiotic plants. They do not, however, firmly demonstrate that this hypothesis is true.

It should be noticed that many symbiotic species carry more than one class 2 hemoglobin gene (Fig. 1), possibly as a consequence of recent genome duplications (Shoemaker et al. 1996; Young et al. 2003). In several instances, functional differences between the proteins encoded by these genes were reported. In lupine, the leghemoglobin II gene is expressed exclusively in nodules, as it is the general case for leghemoglobin, but leghemoglobin I is also expressed in nonnodule tissues, suggesting a potential additional function (Strozycki et al. 2000). In pea (*Vicia faba*), two isohemoglobins of class 2 (Lb I and Lb IV) were described. Lb IV shows a higher O_2-binding activity than Lb I (Uheda and Syono 1982a) and is thereby more effective in supporting nitrogen fixation and oxygen consumption of isolated bacteroids (Uheda and Syono 1982b). The expression of Lb IV is evenly distributed in the central tissue of efficient pea nodules, whereas that of LbI is restricted to a smaller region, from infection zone II to the distal part of nitrogen fixation zone III (Uheda and Syono 1982a). The Lb IV/Lb I concentration ratio changes during nodule development: Lb IV is synthesized mainly in older nodules that actively fix nitrogen (Uheda and Syono 1982a). This heterogeneity in function and pattern of expression would contribute to a more efficient nitrogen fixation (Kawashima et al. 2001). Similar observations were made in *Glycine max* (Appleby 1962; Fuchsman et al. 1976; Fuchsman and Appleby 1979; Verma et al. 1979). This multiplicity of symbiotic genes might possibly explain why the adaptive signal we extracted from coding sequences is relatively low. Once several hemoglobins with various biochemical properties are present in the genome of a symbiotic plant, an adaptation involving the tuning of the amount of oxygen in nodules, if required, might occur most frequently through a change of the expression pattern, not the primary sequence, of available hemoglobin genes. Such an adaptive event would not be detectable from coding sequence analysis. Subfunctionalization (i.e., the split of an ancestral function into several subfunctions after duplication [Force et al. 1999]) is seen here as allowing recurrent adaptive responses without the need for recurrent sequence changes. Given the importance of gene duplications for adaptation in eukaryotic genomes (e.g., see Wagner 2001), this might well be a general process in this kingdom.

One illustration might be given by the evolutionary pattern of mammalian mitochondrial proteins.

Among the many proteins located in the mitochondrion, some are encoded by the mitochondrial genome, and other by the nuclear genome. The mitochondrial genome is of prokaryotic origin and, apparently, constrained to a small size in animals; it shows virtually no gene duplication. Nuclear genes encoding proteins targeted to the mitochondrion, in contrast, are often duplicated and expressed in a tissue-specific way. Schmidt et al. (2001) analyzed the evolutionary process of amino acid sites involved in a chemical interaction between a mitochondrion-encoded and a nucleus-encoded subunit of the cytochrome oxidase complex in primates. They found that in nucleus-encoded subunits, sites interacting with a mitochondrion-encoded subunit evolve more slowly, on average, than noninteracting sites, suggesting an increased purifying selection at interacting sites. In mitochondrion-encoded subunits, in contrast, sites interacting with a nucleus-encoded subunit were found to evolve faster, on average, than noninteracting sites. This can be interpreted in adaptive terms: beneficial changes in the mitochondrion-encoded component must involve primary sequence changes, whereas the nucleus component can adapt by modifying the relative proportion of distinct paralogues in various tissues.

This discussion, although mostly based on the hemoglobin example, leaves us a bit pessimistic about the power of between-species sequence comparison per se for understanding molecular adaptation in eukaryotes. Current methods appear to unambiguously detect positive selection only when the signal is very strong. In the case of plant hemoglobins, this analysis was useful in proposing hypothesis to be tested by alternative approaches, including population genetics, structural biology, and expression data analysis. A more exploratory perspective would be the characterization and sequence analysis of hemoglobin genes from the few nonsymbiotic legume species (De Faria et al. 1989), and from actinorhizal (non legume) symbiotic species other than *Casuarina*, in which symbiotic exchanges are not very dependent on oxygen concentration.

References

Appleby C (1962) The oxygen equilibrium of leghemoglobin. Biochim Biophys Acta 60:226–235

Appleby C (1992) The origin and functions of haemoglobin in plants. Sci Prog 76:365–398

Arredondo-Peter R, Hargrove MS, Moran JF, Sarath G, Klucas RV (1998) Plant hemoglobins. Plant Physiol 118:1121–1125

Bielawski J, Yang Z (2003) Maximum likelihood methods for detecting adaptive evolution after gene duplication. J Struct Funct Genomics 3:201–212

Biermann CH (1998) The molecular evolution of sperm bindin in six species of sea urchins (Echinodea: Strongylocentrotidae). Mol Biol Evol 15:1761–1771

Civetta A, Singh RS (1998) Sex-related genes, directional selection, and speciation. Mol Biol Evol 15:901–909

Cueto M, Hernandez-Perera O, Martin R, Bentura M, Rodrigo J, Lamas S, Golvano M (1996) Presence of nitric oxide synthase activity in roots and nodules of *Lupinus albus*. FEBS Lett 398:159–164

Davies M, Puppo A (1992) Direct detection of radicals in intact soybean nodules: presence of nitric oxide-leghemoglobin complexes. Biochem J 281:197–201

De Faria SM, Lewis GP, Sprent JI, Sutherland JM (1989) Occurrence of nodulation in the Leguminosae. New Phytol 111:607–619

Force A, Lynch M, Bryan Picket F, Amores A, Yan Y, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545

Fuchsman W, Appleby C (1979) Separation and determination of the relative concentrations of the homogeneous components of soybean leghemoglobin by isoelectric focusing. Biochim Biophys Acta 579:314–324

Fuchsman W, Barton C, Stein M, Thompson J, Willet RM (1976) Leghemoglobin: different roles for different components?. Biochem Biophys Res Commun 68:387–392

Galtier N (2001) Maximum likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol 18:866–873

Golding GB, Dean AM (1998) The structural basis of molecular adaptation. Mol Biol Evol 15:355–369

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736

Goodman MD, Hargrove MS (2001) Quaternary structure of rice nonsymbiotic hemoglobin. J Biol Chem 276:6834–6839

Grossman LI, Schmidt TR, Wildman DE, Goodman M (2001) Molecular evolution of aerobic energy metabolism in primates. Mol Phylogenet Evol 18:26–36

Guldner E, Desmarais E, Galtier N, Godelle B (2004) Molecular evolution of plant hemoglobin: two hemoglobin genes in Nymphaeaceae *Euryale ferox*. J Evol Biol 17:48–54

Hargrove MS, Brucker EA, Stec B, Sarath G, Arredondo-Peter R, Klucas RV, Olson JS, Phillips GN Jr (2000) Crystal structure of a nonsymbiotic plant hemoglobin. Struct Fold Des 8:1005–1014

Huelsenbeck J, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755

Hughes AL, Nei M (1988) Nucleotide substitution at major histocompatibility complex loci reveals overdominant selection. Nature 335:167–170

Hughes AL, Nei M (1989) Nucleotide substitution at major histocompatibility complex class II loci: Evidence for overdominant selection. Proc Natl Acad Sci USA 86:958–962

Hunt RA, Watts RA, Trevaskis B, Llewellyn DJ, Burnell J, Dennis ES, Peacock WJ (2001) Expression and evolution of functionally distinct haemoglobin genes in plants. Plant Mol Biol 47:677–692

Jacobsen-Lyon K, Jensen EO, Jorgensen J-E, Marcker KA, Peacock WJ, Dennis ES (1995) Symbiotic and nonsymbiotic hemoglobin genes of *Casuarina glauca*. Plant Cell 7:213–223

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comp Appl Biosci 8:275–282

Jun HK, Sarath G, Moran JF, Becana M, Klucas RV, Wagner FW (1994) Characteristics of modified leghemoglobins isolated from soybean (*Glycine max* Merr) root nodules. Plant Physiol 104:1231–1236

Kawashima K, Suganuma N, Tamaoki M, Kouchi H (2001) Two types of pea leghemoglobin genes showing different O_2-binding affinities and distinct patterns of spatial expression in nodules. Plant Physiol 125:641–651

Li W, Luo C, Wu C (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150–174

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

Miller R, McRae D, Al-Jobore A, Berndt W (1988) Respiration supported nitrogenase activity of isolated rhizobium meliloti bacteroids. J Cell Biochem 38:35–49

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936

Ohta T (1988) Further simulation studies on evolution by gene duplication. Evolution 42:375–386

Pupko T, Galtier N (2002) A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc R Soc Lond B Biol Sci 269:1313–1316

Savolainen V, Chase MW, Hoot SB, Morton CM, Soltis DE, Bayer C, Fay MF, de Bruijn AY, Sullivan S, Qiu YL (2000) Phylogenetics of flowering plants based on combined analysis of plastid atpB and rbcL gene sequences. Syst Biol 49:306–362

Schmidt TR, Wu W, Goodman M, Grossman LI (2001) Evolution of mitochondria- and nuclear-encoded subunit interaction in cytochrome c oxidase. Mol Biol Evol 18:563–569

Shoemaker RC, Polzin K, Labate J, Specht J, Brummer EC, Olson T, Young N, Concibido V, Wilcox J, Tamulonis JP, Kochert G, Boerma HR (1996) Genome duplication in soybean (*Glycine* subgenus *soja*). Genetics 144:329–338

Strozycki PM, Karlowski WM, Dessaux Y, Petit A, Legocki AB (2000) Lupine leghemoglobin I: expression in transgenic Lotus and tobacco tissues. Mol Gen Genet 263:173–182

Swanson WJ, Vacquier VD (2002) The rapid evolution of reproductive proteins. Nat Rev Genet 3:137–144

Trevaskis B, Watts A, Andersson CR, Llewellyn DJ, Hargrove MS, Olson JS, Dennis ES, Peacock WJ (1997) Two hemoglobin genes in Arabidopsis thaliana: The evolutionary origins of leghemoglobins. Proc Natl Acad Sci USA 94:12230–12234

Uheda E, Syono K (1982a) Physiological role of leghaemoglobin heterogeneity in pea root nodule development. Plant Cell Physiol 23:75–84

Uheda E, Syono K (1982b) Effects of leghaemoglobin components on nitrogen fixation and oxygen consumption. Plant Cell Physiol 23:85–90

Verma DPS, Ball S, Guérin C, Wanamaker L (1979) Leghemoglobin biosynthesis in soybean root nodules: characterization of the nascent and released peptides and the relative rate of synthesis of the major leghemoglobins. Biochemistry 18:476–483

Wagner A (2001) Birth and death of duplicated genes in completely sequenced eukaryotes. Trends Genet 17:237–239

Wagner FW, Sarath G (1987) Biochemical changes in stressed and senescent soybean root nodules. In: Thompson ED, Nothnagel NA, Huffaker RC (eds) Plant senescence: Its biochemistry and physiology. American Society of Plant Physiologists, Rockville, MD, pp 190–197

Wyckoff GJ, Wang W, Wu CI (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. Comp Appl Biosci 13:555–556

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysosyme evolution. Mol Biol Evol 15:568–573

Yang Z, Nielsen R, Goldman N, Krabbe Pedersen A-M (2000a) Codon-substitution models for heteregeneous selection pressure at amino acid sites. Genetics 155:431–449

Yang Z, Swanson WJ, Vacquier VD (2000b) Maximum likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. Mol Biol Evol 17:1446–1455

Young ND, Mudge J, Ellis THN (2003) Legume genomes: more than peas in a pod. Curr Opin Plant Biol 6:199–204

Zhang J (2003) Evolution by gene duplication: an update. Tends Ecol Evol 18:292–298