Maximum-Likelihood Analysis of Molecular Adaptation in Abalone Sperm Lysin Reveals Variable Selective Pressures Among Lineages and Sites

Ziheng Yang, Willie J. Swanson,† and Victor D. Vacquier‡*

*Galton Laboratory, Department of Biology, University College London, London, England; †Department of Molecular Biology and Genetics, Cornell University; and ‡Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California San Diego, La Jolla, California

Maximum-likelihood models of codon substitution were used to analyze sperm lysin genes of 25 abalone (*Haliotis*) species to identify lineages and amino acid sites under diversifying selection. The models used the nonsynonymous/synonymous rate ratio ($\omega = d_N/d_S$) as an indicator of selective pressure and allowed the ratio to vary among lineages or sites. Likelihood ratio tests suggested significant variation in selective pressure among lineages. The variable selective pressure provided an explanation for the previous observation that the ω ratio is >1 in comparisons of closely related species and <1 in comparisons of distantly related species. Computer simulations demonstrated that saturation of nonsynonymous substitutions and constraint on lysin structure were unlikely to account for the observed pattern. Lineages linking closely related sympatric species appeared to be under diversifying selection, while lineages separating distantly related species from different geographic locations were associated with low evolutionary rates. The selective pressure indicated by the ω ratio was found to vary greatly among amino acid sites in lysin. Sites under potential diversifying selection were identified. Ancestral lysins were inferred to trace the route of evolution at individual sites and to provide lysin sequences for future laboratory studies.

Introduction

Comparison of synonymous (silent, d_S) and nonsynonymous (amino acid–changing, d_N) substitution rates in protein-coding genes provides an important means for understanding molecular evolution. The nonsynonymous/synonymous rate ratio ω ($=d_N/d_S$) measures selective pressure at the protein level. If nonsynonymous mutations are deleterious and are fixed at a lower rate than synonymous mutations, the ω ratio will be <1. If nonsynonymous mutations have no effect on the fitness of the protein and are fixed at the same rate as synonymous mutations, ω should be equal to 1. If nonsynonymous mutations are beneficial and favored by natural selection, ω should be >1. Significantly higher nonsynonymous than synonymous rates are evidence for molecular adaptation. This criterion has been used to identify a number of cases of adaptive molecular evolution (Li 1997). Among the best studied are proteins involved in reproduction (Palumbi 1994; Civetta and Singh 1998; Vacquier et al. 1999). Examples include sperm bindin in sea urchins (Palumbi and Metz 1991; Metz and Palumbi 1996; Biermann 1998), abalone sperm lysin (Vacquier and Lee 1993; Lee, Ota, and Vacquier 1995; Swanson and Vacquier 1995; Metz, Robles-Sikisaka, and Vacquier 1998; Hellberg and Vacquier 1999, 2000), male reproductive proteins in *Drosophila* (Tsaur and Wu 1997; Wyckoff, Wang, and Wu 2000), and other sex-linked genes (Ferris et al. 1997).

Abalones are marine mollusks (order Archeogastropoda) that broadcast and spawn their gametes into seawater where fertilization and embryogenesis occur. Many sympatric abalone species overlap in depth zonation and reproductive seasonality, yet maintain themselves as distinct species. Molecular incompatibilities in the gamete recognition system may be responsible for reproductive barriers between species. Lysin is a ~16 kDa protein that is released by the sperm onto the egg vitelline envelope (VE), an elevated, glycoproteinacious, protective microchamber in which development occurs. Lysin is a nonenzymatic protein that unravels the tightly intertwined glycoprotein fibers of the VE to create a 3-μm-diameter hole through which the sperm passes prior to fusing with the egg cell membrane. The species specificity of lysin's ability to dissolve isolated egg VEs can be quantitatively demonstrated (reviewed in Vacquier et al. 1999). Analysis of mitochondrial DNA sequences permitted dating of the abalone species divergence times and revealed that lysins evolve at surprisingly rapid rates (Metz, Robles-Sikisaka, and Vacquier 1998).

Lysin is the first gamete recognition protein whose crystal structure has been solved (Shaw et al. 1993, 1995; Kresge, Vacquier, and Stout 2000*a*, 2000*b*). Red (*Haliotis rufescens*) and green (*Haliotis fulgens*) abalone sperm lysins differ in 51 out of 134 amino acid positions. Although highly divergent in amino acid sequence, there are several conserved features between these two lysins. For example, the alpha-carbon atom ribbon diagrams of the two lysins are essentially identical and thus superimposable with an average rms deviation of 1.14 Å. Other conserved structural features include one surface of the protein having a hydrophobic patch of 11–16 amino acids (by which lysin dimerizes) and, on the opposite surface, two tracks of basic amino acids which run down the entire length of the lysin. Variations in sequence between species map to the lysin surface, where they could be involved in species-specific recognition of the VE (Kresge, Vacquier, and Stout 2000*a*, 2000*b*). The N- and C-termini lie on the same surface of the lysin. The N-terminal residues 2–12 are always species-unique in sequence, while the C-termi-

Key words: abalone, fertilization, likelihood ratio test, lysin, maximum likelihood, molecular adaptation, molecular evolution, positive selection, reinforcement, sperm-egg recognition.

Address for correspondence and reprints: Ziheng Yang, Galton Laboratory, Department of Biology, University College London, 4 Stephenson Way, London NW1 2HE, United Kingdom. E-mail: z.yang@ucl.ac.uk.

Mol. Biol. Evol. 17(10):1446–1455. 2000

nus is moderately variable, but not species-unique. Recombinant lysins, in which both termini were exchanged between two species, demonstrated that these domains play important roles in species-specific recognition leading to VE dissolution (Lyon and Vacquier 1999).

The only molecule in the VE-binding lysin with high, species-selective affinity is a fibrous glycoprotein of ~1,000 kDa named VERL (vitelline envelope receptor for lysin). VERL is a major structural element of the egg VE. Cloning and sequencing showed that VERL was largely composed of approximately 28 tandemly repeating (intronless) 153–amino acid motifs. Sequencing of VERL repeats from the seven species of Californian abalone showed that VERL repeats were subjected to weak purifying selection and evolved by the process of concerted evolution (Elder and Turner 1995). This is in contrast to lysin, VERL's cognate binding partner, which evolves rapidly in response to strong selection pressure.

Lee, Ota, and Vacquier (1995) determined cDNA sequences for lysin from 20 worldwide abalone species and performed pairwise sequence comparisons to estimate d_S and d_N using the method of Nei and Gojobori (1986; hereinafter referred to as NG). The ω ratio was found to be >1 when closely related species were compared but <1 when divergent species were compared (see fig. 3 in Lee, Ota, and Vacquier 1995). Lee, Ota, and Vacquier (1995) hypothesized that continuous selective pressure may have driven lysin evolution and that the small estimates of the ω ratio in comparisons of divergent species may be due to saturation of nonsynonymous substitutions and functional constraints on lysin structure.

While the higher nonsynonymous than synonymous rates provide unequivocal evidence for adaptive evolution in lysin, the selective pressure remains elusive. Besides bias in the estimation procedure, the inverse relationship between sequence divergence and the ω ratio can also be caused by variable selective pressures (variable ω ratios) among lineages. To distinguish between those factors, we performed a maximum-likelihood (ML) analysis of the lysin gene sequence data. Likelihood models developed recently (Yang 1998; Yang and Nielsen 1998) account for variable ω ratios among branches in the phylogeny and can be used to test adaptive evolution along lineages. Models have also been developed that allow the ω ratio to vary among amino acid sites (Nielsen and Yang 1998; Yang et al. 2000). These models may be used to identify critical amino acids under diversifying selection without knowledge of functionally important domains. The likelihood models are based on more realistic assumptions, as they account for unequal transition and transversion rates and unequal base and codon frequencies. We also infer ancestral lysin proteins to trace evolutionary changes at individual sites.

Materials and Methods

Sequence Data

Sperm lysin cDNA sequences from 25 abalone species are analyzed in this paper; the species are identified in figure 1. Twenty species were used in the study of Lee, Ota, and Vacquier (1995), and five more (*Haliotis pustulata*, *Haliotis varia*, *Haliotis coccinea*, *Haliotis gigantea*, and *Haliotis conicopora*) were included in this paper (Lee and Vacquier 1995). Only codons encoding amino acids in mature lysin were used, with the signal sequence removed. Two codons (amino acids) had alignment gaps in all but one or two species and were removed from all sequences. Amino acids (codons) were numbered according to the lysin of the red abalone (*H. rufescens*; Vacquier and Lee 1993; Lee, Ota, and Vacquier 1995). Amino acids 135 (Gly) and 136 (Lys) were removed, as they are not present in mature lysin (Lyon and Vacquier 1999). The alignment contained 135 codons in each sequence, with one gap (numbered 133a) introduced between amino acids 133 and 134 relative to lysin of the red abalone. The phylogeny of Lee, Ota, and Vacquier (1995) and Lee and Vacquier (1995) was used (fig. 1).

Statistical Analysis

Pairwise sequence comparisons were performed as an exploratory analysis and for comparison with previous results (Lee, Ota, and Vacquier 1995). The NG (Nei and Gojobori 1986) and ML (Goldman and Yang 1994) methods were used to estimate d_S and d_N. The ML method is based on an explicit model of codon substitution which accounts for the transition/transversion rate ratio (κ) and base frequencies at the three codon positions. Parameters κ, ω and sequence divergence t were estimated by ML, while the base frequencies were estimated empirically from the data. For each pairwise comparison, the likelihood ratio test (LRT) could be used to test whether the ω ratio was significantly different from 1. This was done by comparing the log-likelihood values with $\omega = 1$ constrained and without such constraint. If the null hypothesis $\omega = 1$ is correct, twice the log-likelihood difference between the two models asymptotically has a χ^2 distribution with df = 1.

Models of variable ω ratios among lineages were fitted by ML to the alignment of 25 sequences (Yang 1998; Yang and Nielsen 1998). The ''one-ratio'' model assumes the same ω ratio for all branches. The free-ratios model assumes an independent ω ratio for each branch. Comparison of the two models constitutes an LRT of the hypothesis that the ω ratio is identical among lineages. The free-ratios model is parameter-rich and is unlikely to produce accurate estimates for all ω ratios. Nevertheless, it is interesting to estimate the ω ratios without constraints, as knowledge of which lineages are under diversifying selection may provide clues to the selective pressure driving lysin's divergence. For example, if closely related, sympatric species show strong diversifying selection, while distant allopatric species show purifying selection, a selective pressure related to speciation may be implicated. Such a hypothesis can be implemented as a ''two-ratios'' model that assumes different ω ratios for sympatric and allopatric lineages in the phylogeny.

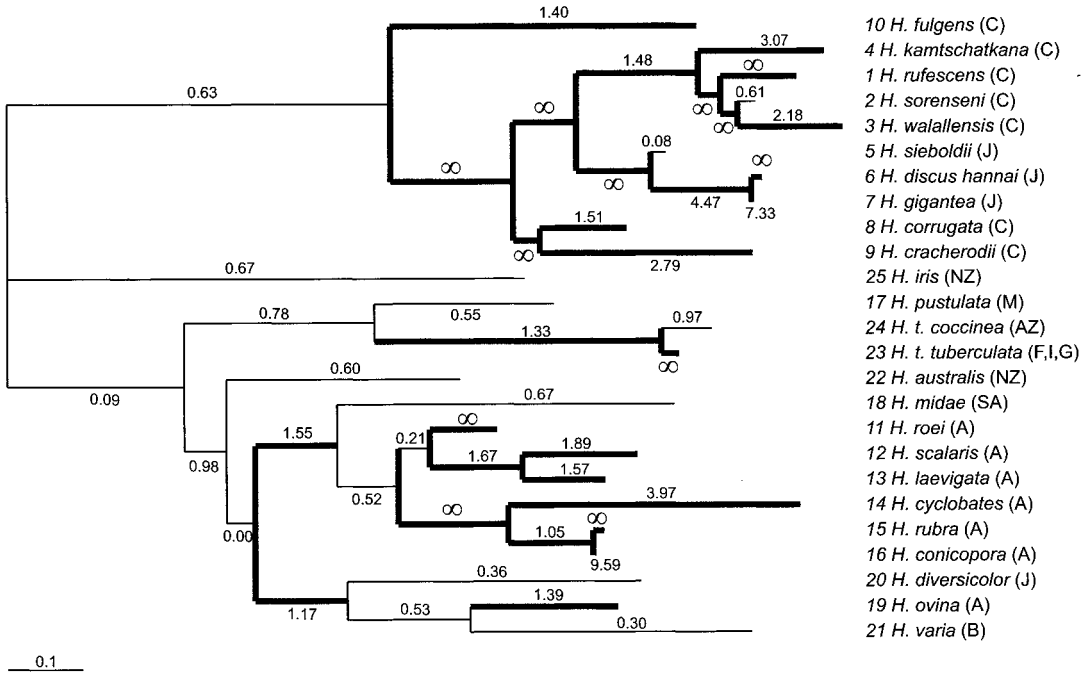

FIG. 1.—Phylogeny for the sperm lysin genes from 25 abalone (genus *Haliotis*) species. Letters in parentheses indicate the collecting sites: Australia (A), Azores (AZ), Borneo (B), California (C), France (F), Greece (G), Italy (I), Japan (J), Madagascar (M), and New Zealand (NZ). Analysis in this paper used the unrooted topology only. Branches are drawn in proportion to their lengths, defined as the expected number of nucleotide substitutions per codon. Maximum-likelihood estimates of branch lengths were obtained under the ''free-ratios'' model, which assumes an independent ω ratio (d_N/d_S) for each branch in the tree. Estimates of the ω ratios under that model are shown along branches, and branches for which the estimated ω ratios are >1 are drawn in thick lines.

Models of variable ω ratios among sites were used to test for the presence of sites under diversifying selection (with $\omega > 1$) and to identify them (Nielsen and Yang 1998). Note that in this paper, a site refers to an amino acid or codon rather than a nucleotide. We use the following five models for the ω distribution (table 1), implemented in the CODEML program of the PAML package (Yang 1999; Yang et al. 2000). Model M1 (neutral) assumes two classes of sites in the protein: the conserved sites at which $\omega = 0$ and the neutral sites at which $\omega = 1$. Model M2 (selection) adds a third class of sites with ω as a free parameter, thus allowing for sites with $\omega > 1$. Model M3 (discrete) uses a general discrete distribution with three site classes, with the proportions (p_0, p_1, and p_2) and the ω ratios (ω_0, ω_1, and ω_2) estimated from the data. Model M7 (beta) uses a beta distribution $B(p, q)$, which, depending on parameters p and q, can take various shapes (such as L, J, U, and inverted U shapes) in the interval (0, 1). Model M8 (beta and ω) adds an extra class of sites to the beta (M7) model, with the proportion and the ω ratio estimated from the data, thus allowing for sites with $\omega > 1$. From these models, we construct three LRTs (table 2), which compare M0 (one ratio) with M3 (discrete), M1 (neutral) with M2 (selection), and M7 (beta) with M8 (beta & ω), respectively. When the alternative models (M2, M3, and M8) suggest the presence of sites with $\omega > 1$, all three tests can be considered tests of positive selection

Table 1
Parameter Estimates and Log-Likelihood Values Under Models of Variable ω Ratios Among Sites

Model	p	Parameters	ℓ	d_N/d_S	Positively Selected Sites
M0: one ratio ...	1	$\omega = 0.929$	−4,682.42	= ω	None
M1: neutral	1	$p_0 = 0.212$, $\omega_0 = 0$ $p_1 = 0.788$, $\omega_1 = 1$	−4,577.12	0.788	Not allowed
M2: selection ...	3	$p_0 = 0.204$, $\omega_0 = 0$ $p_1 = 0.515$, $\omega_1 = 1$ $p_2 = 0.281$, $\omega_2 = 3.816$	−4,481.17	1.587	**4**, **7**, **9**, **10**, **12**, **14**, **32**, **33**, **36**, *41*, **44**, **64**, **67**, **70**, **74**, **83**, **86**, **87**, **113**, *115*, **120**, *126*, *127*, **132**, *133a*
M3: discrete	5	$p_0 = 0.329$, $\omega_0 = 0.085$ $p_1 = 0.402$, $\omega_1 = 0.911$ $p_2 = 0.269$, $\omega_2 = 3.065$	−4,464.35	1.218	**4**, **7**, **9**, **10**, **12**, *14*, **32**, **33**, **36**, **41**, *44*, *64*, **67**, **70**, **74**, **83**, *86*, **87**, **113**, **120**, *126*, *127*, **132**, *133a*
M7: beta	2	$p = 0.283$, $q = 0.211$	−4,524.72	0.572	Not allowed
M8: beta & ω ...	4	$p_0 = 0.731$, $p = 0.310$, $q = 0.291$ $p_1 = 0.269$, $\omega = 2.955$	−4,464.66	1.173	**4**, **7**, **9**, **10**, **12**, *14*, **32**, **33**, **36**, *41*, **44**, *64*, *67*, **70**, **74**, **83**, *86*, **87**, **113**, **120**, *126*, **132**, *133a*

NOTE.—p = number of parameters in the ω distribution; d_N/d_S = average over sites. Sites inferred under selection at the 99% level are listed in bold, and those at the 95% level are in italic.

Table 2
Likelihood Ratio Statistics ($2\Delta\ell$)

Comparison	$2\Delta\ell$	df	$\chi^2_{1\%}$
M0 (one ratio) vs. M3 (discrete) . . .	436.14	4	13.28
M1 (neutral) vs. M2 (selection) . . .	191.90	2	9.21
M7 (beta) vs. M8 (beta & ω)	120.12	2	9.21

(Nielsen and Yang 1998; Yang et al. 2000). However, the comparison of M0 with M3 may also be considered a test of variable ω values among sites. After ML estimates of parameters are obtained, the Bayes theorem is used to calculate the posterior probabilities of site classes for each site (Nielsen and Yang 1998). If the ω ratios for some site classes are >1, sites with high posterior probabilities for those classes are likely to be under diversifying selection.

Lysin sequences for extinct ancestral nodes in the phylogeny were inferred using the empirical Bayes approach (Yang, Kumar, and Nei 1995) to trace amino acid changes at individual sites. The analysis was performed using the codon (nucleotide) sequences. Posterior probabilities for synonymous codons (codons that encode the same amino acid) at each ancestral node were summed to calculate the probability for that amino acid. All statistical analyses were performed using the CODEML program in the PAML package (Yang 1999).

Computer Simulation

Computer simulation was performed to examine the effect of saturation of nonsynonymous substitutions (e.g., due to functional constraints on lysin) on the analysis of variable selective pressures among lineages. Codon sequences were simulated using the EVOLVER program in the PAML package. The program generates a codon sequence for the root of the tree and ''evolves'' the sequence along branches in the phylogeny using specified branch lengths and substitution parameters. The ω ratio was assumed to be either constant across sites or variable according to a discrete distribution. Each simulated data set was analyzed in the same way as the original data either by pairwise comparison or by ML joint analysis. Values of parameters used in the simulation are described later.

Results

Pairwise Comparison

The ML method (Goldman and Yang 1994) was used to estimate d_S and d_N, with the transition/transversion rate ratio κ fixed at 1.6 to reduce sampling errors. This was the ML estimate from joint analysis of all sequences. The estimates are plotted in figure 2*A*. Estimates obtained using the NG method (Nei and Gojobori 1986) are not shown but were very similar to the ML estimates, especially for small distances (see also fig. 3 in Lee, Ota, and Vacquier 1995). At large distances ($d_S > 0.5$, $d_N > 0.2$), NG gave slightly lower estimates of d_S and d_N than did ML. Both the transition/transversion bias and the base frequency biases were weak, and the two methods produced very similar estimates.

An intriguing pattern in figure 2*A* is that $d_N > d_S$ at low sequence divergences ($d_S < 0.2$), and $d_N < d_S$ at high divergences ($d_S > 0.4$), as found by Lee, Ota, and Vacquier (1995). In figure 2*B*, ω is plotted against the sequence divergence *t*, where the inverse relationship between the two is strikingly clear. Sequence divergence *t* is defined as the expected number of nucleotide substitutions per codon and is about $3 \times (0.243d_S + 0.757d_N)$, where 0.243 and 0.757 are the proportions of synonymous and nonsynonymous sites in the lysin genes (Goldman and Yang 1994). Lee, Ota, and Vacquier (1995) hypothesized that lysin evolution may be continuous and attributed the smaller estimates of ω in distant comparisons to saturation of nonsynonymous substitutions due to constraints on lysin structure. In later sections, we examine possible variation of selective pressure along lineages as an alternative explanation.

The hypothesis $\omega = 1$ can be tested using an LRT comparing the null model H_0 with $\omega = 1$ fixed and the alternative model with ω estimated from the data. Suppose the log-likelihood values under the two models are ℓ_0 and ℓ_1, respectively. Then, $2\Delta\ell = 2(\ell_1 - \ell_0)$ can be

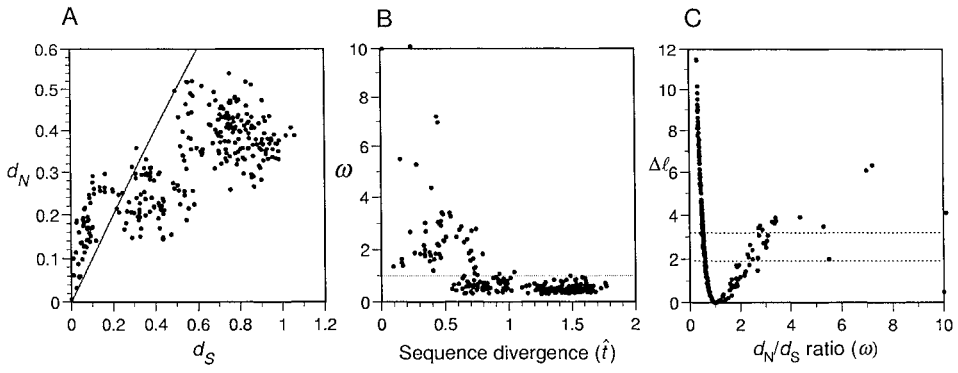

FIG. 2.—Pairwise comparison of sperm lysin genes from 25 abalone species. *A*, Estimates of d_S and d_N using ML (Goldman and Yang 1994). *B*, The ω ratio (d_N/d_S) plotted against the estimated sequence divergence *t*, defined as the expected number of nucleotide substitutions per codon. *C*, The LRT statistic $\Delta\ell = \ell_1 - \ell_0$, where ℓ_0 and ℓ_1 are the likelihood values with and without the constraint $\omega = 1$, respectively. $2\Delta\ell$ can be compared with χ^2 with df = 1 to test the null hypothesis $\omega = 1$. The 95% and 99% critical values, $\frac{1}{2}\chi^2 = 1.92$ and 3.32, are indicated on the graph. In the comparisons between *Haliotis discus hannai* and *Haliotis gigantea* and between *Haliotis rubra* and *Haliotis conicopora*, no synonymous difference was found, and ω is set to 10 in the plots.

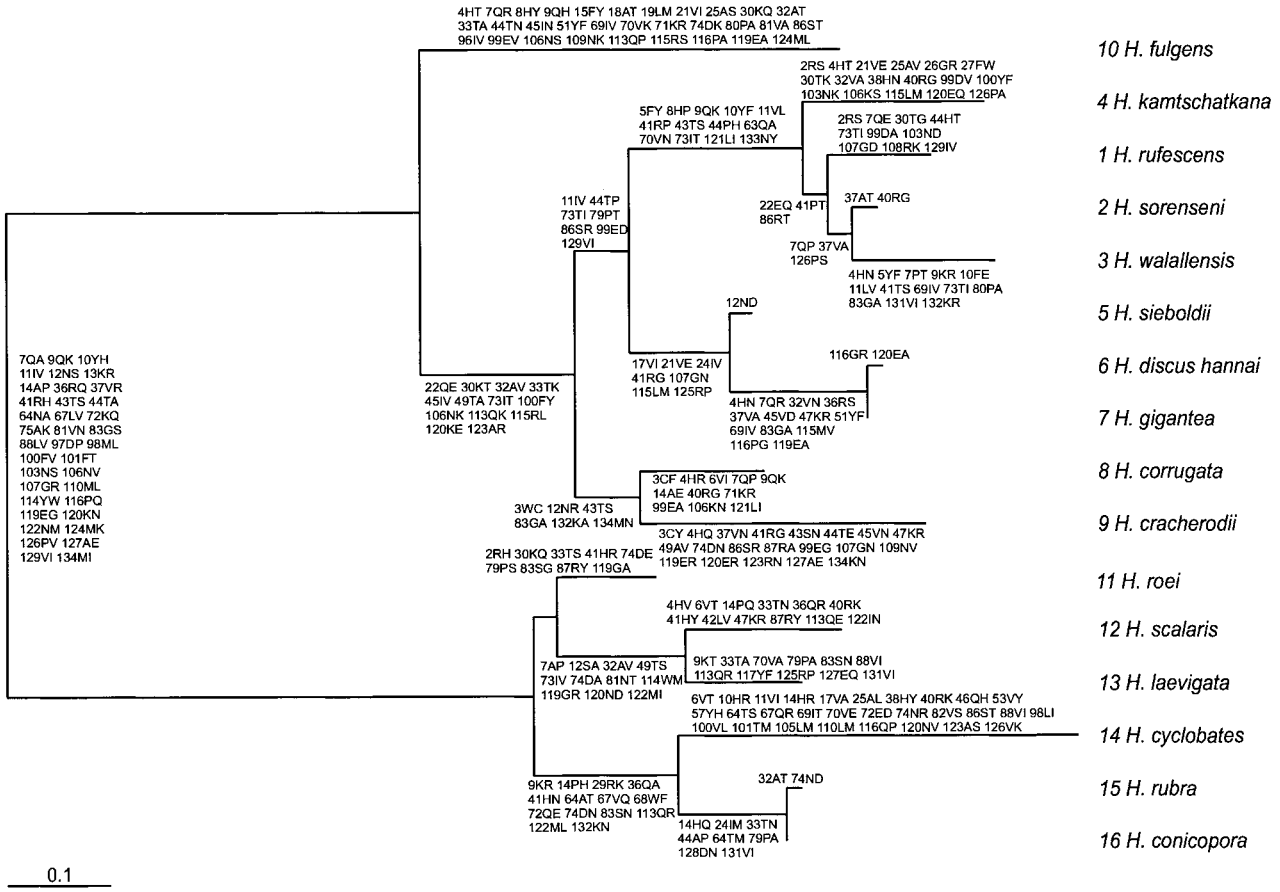

FIG. 3.—Phylogeny for a subset of 16 abalone species, with inferred amino acid changes shown along branches. Branch lengths were estimated, and ancestral sequences were reconstructed under a codon substitution model that assumed an ω ratio for branches within the North Pacific (species 1–10) and Australian (species 11–16) clades and another ω ratio for the branch connecting the two clades.

compared with a χ^2 distribution with df = 1 to test whether ω is significantly different from 1. Use of the χ^2 approximation is reliable when the sample size is ''large'' or when the sequence is long. As the lysin gene has only about 130 codons, we performed computer simulations to check the reliability of the χ^2 approximation. We examined two pairwise comparisons, with the estimated ω being >1 and <1, respectively. The first comparison was between *H. rufescens* and *H. sorenseni.* The estimates under H_1 were $\hat{t} = 0.149$ and $\hat{\omega} = 5.572$, with $\ell_1 = -634.04$, while under H_0, $\hat{t} = 0.149$ and $\ell_1 = -636.09$. Thus, $2\Delta\ell = 2(\ell_1 - \ell_0) = 4.10$, with a P value of 0.04 from the χ^2 distribution. We simulated 200 replicate data sets using parameter estimates from H_0, and for each data set, we calculated the log-likelihood values under H_0 and H_1. The test statistics among replicates are used to construct a histogram. The observed value is at the fifth percentile of the simulated distribution. This P value is close to that from the χ^2 distribution. A second case is that of the comparison between *H. rufescens* and *H. roei.* Estimates of parameters under H_1 were $\hat{t} = 1.411$ and $\hat{\omega} = 0.482$, with $\ell_1 = -910.15$, while under H_0, $\hat{t} = 1.348$ and $\ell_1 = -914.01$. Thus, $2\Delta\ell = 7.73$, and $P = 0.0054$ from the χ^2 distribution. The P value estimated from 1,000 replicates simulated under H_0 was 0.010. In both comparisons, the P values from the χ^2 distribution are close to those from the simulated distributions. While the limitations of pairwise comparison should be borne in mind, the χ^2 approximation to the LRT appears usable even for pairwise comparisons and small genes such as lysin. Use of the χ^2 approximation for data of multiple sequences, which contain more information, is expected to be more reliable.

The LRT statistics are plotted in figure 2*C* as a function of the estimated ω ratio. For 12 comparisons, ω is >1 at the 1% significance level (fig. 2*C*). In 13 additional pairs, ω is >1 at the 5% level. It is noteworthy that a large estimate of the ω ratio is not necessarily strong evidence for adaptive evolution, as an estimate based on very few changes is unreliable. In two comparisons, that between *H. discus hannai* and *H. gigantea* and that between *H. rubra* and *H. conicopora,* two nonsynonymous differences and no synonymous differences were found. The estimate of ω in both comparisons was ∞ ($d_N = 0.0065$, $d_S = 0$). However, this ratio is not significantly >1; the test statistic $2\Delta\ell = 1.0$ is rather small, with $P = 0.3$. Because there were only two changes, we are unable to rule out chance effects.

For distant comparisons, the estimated ω ratios are <1. In 107 comparisons, ω is significantly <1 at the 1% level, while in 49 additional comparisons, ω is signifi-

cantly <1 at the 5% level. For the remaining 119 comparisons, the null hypothesis $\omega = 1$ cannot be rejected. While many more pairs show $\omega < 1$ than $\omega > 1$, the pairwise comparisons are not independent and the patterns are not easy to interpret.

Variable Selective Pressures and Positive Selection Along Lineages

The one-ratio model assumes the same ω ratio for all lineages (fig. 1) and involves a total of 58 parameters: 47 branch lengths, 9 parameters for the nucleotide frequencies at the three codon positions, the transition/transversion rate ratio κ, and the d_N/d_S ratio ω. The log-likelihood value under this model was $\ell_0 = -4{,}682.42$, with parameter estimates $\kappa = 1.56$ and $\omega = 0.929$. This ω ratio was an average over all sites and lineages. While larger than estimates from most other genes (see, e.g., Li 1997; Yang and Nielsen 1998), this ratio was <1. The free-ratios model, which assumes an independent ω ratio for each branch, was then applied to the same data. The tree in figure 1 has 47 branches, so 46 additional ω parameters are involved in this model. Estimates of the ω ratios are shown along branches in figure 1. The likelihood value under this model was $\ell_1 = -4{,}627.27$. Comparison of $2\Delta\ell = 2(\ell_1 - \ell_0) = 2 \times 55.15 = 110.30$ with the χ^2 distribution (df = 46) suggests rejection of the one-ratio model, with $P = 0.3 \times 10^{-6}$. The ω ratios are extremely variable among lineages.

Estimates of ω ratios in figure 1 suggest that the ω ratios for recent lineages linking closely related or sympatric species are most often >1, while those for branches separating distantly related species are all <1. For example, in the North Pacific clade (species 1–10 from California and Japan), all except two very short branches have estimates of $\omega > 1$. A similar pattern is seen among closely related species in the Australia clade (species 11–16). One explanation is that lysin may be under pressure to evolve to establish reproductive isolation during sympatric speciation, while such pressure is absent in allopatric lineages or after reproductive barriers are well established. One difficulty with testing such a hypothesis is that the models used here only describe the average pattern along each lineage. If a lineage undergoes a short episode of positive selection but is under purifying selection most of the time, the average ω ratio for the lineage may not be >1. Another difficulty with testing the hypothesis is that we do not know whether a node represents a sympatric or an allopatric speciation event. Here, we used a phylogeny (fig. 3) for a subset of species from the North Pacific (species 1–10) and Australia (species 11–16) clades to fit a two-ratios model, assuming a ratio ω_0 for the branch connecting the two clades and another ratio ω_1 for branches within the two clades. The one-ratio model for this subset data gives $\ell_0 = -2{,}719.60$, with the estimate $\hat{\omega} = 1.450$. The two-ratios model gives $\ell_1 = -2{,}698.46$, with $\hat{\omega}_0 = 0.272$ between clades and $\hat{\omega}_1 = 2.396$ within clades. The LRT suggested that the two ω ratios were significantly different; $2\Delta\ell = 2(\ell_1 - \ell_0) = 2 \times 21.14 = 42.28$, and $P < 0.8 \times 10^{-10}$ with df = 1. Furthermore, an LRT comparing models with $\omega_1 = 1$ constrained and without such constraint suggested that ω_1 was significantly >1; $2\Delta\ell = 2 \times (-2{,}698.46 - (-2{,}712.97)) = 29.02$ is much greater than $\chi^2_{1\%} = 6.63$ with df = 1.

However, the two branches separating the Japanese (*H. sieboldii, H. discus hannai,* and *H. gigantea*) and Californian species probably represent allopatric rather than sympatric speciation. When those two branches were assigned the ratio ω_0, the same conclusions as above were reached. The two-ratios model gave $\ell_1 = -2{,}706.10$ with $\hat{\omega}_0 = 0.477$ and $\hat{\omega}_1 = 2.326$, and the two-ratios model with $\omega_1 = 1$ constrained gave $\ell_0 = -2{,}718.76$. Thus, ω_1 was significantly $>\omega_0$ ($2\Delta\ell = 27.00$) and also significantly >1 ($2\Delta\ell = 25.32$). Nevertheless, the large estimates of ω for the two lineages under the free-ratios model, which were averages over all sites, did not appear to be due to chance effects and were in contradiction to the hypothesis.

Variable Selective Pressures Among Sites and Identification of Amino Acids Under Diversifying Selection

Table 1 lists parameter estimates and log-likelihood values under models of variable ω ratios among sites. Model M0 (one ratio) assumes the same ratio for all sites and fits the data far more poorly than any of the other models, which account for variable ω ratios across sites. For example, M3 (discrete) involves four more parameters than M0 (one ratio), and the LRT statistic $2\Delta\ell = 436.14$ is much greater than the critical value $\chi^2_{1\%} = 13.28$ with df = 4 (table 2). The results suggest extreme variation in selective pressure among amino acid sites.

All three models that allow for the presence of positively selected sites, i.e., M2 (selection), M3 (discrete), and M8 (beta & ω), do suggest the presence of such sites (table 1). Allowing for the presence of positively selected sites (with $\omega > 1$) improves the fit of the models significantly. For example, the neutral model (M1) does not allow for sites with $\omega > 1$. The selection model (M2) adds an additional site class, with the ω ratio estimated to be 3.8. The log-likelihood improvement was huge, as $2\Delta\ell = 191.90$ should be compared with $\chi^2_{1\%} = 9.21$ with df = 2 (table 2). Comparison between M7 (beta) and M8 (beta & ω) produced similar results (table 2). M7 (beta) fits the data better than M1 (neutral), as it allows for sites at which $0 < \omega < 1$ (table 1).

Posterior probabilities for site classes calculated under M3 (discrete) are plotted in figure 4. ML estimation suggests that the three site classes are in proportions $p_0 = 0.329$, $p_1 = 0.402$, and $p_2 = 0.269$, with the ratios $\omega_0 = 0.085$, $\omega_1 = 0.911$, and $\omega_2 = 3.065$ (table 1). Those proportions are the prior probabilities that any site belongs to each of the three classes. The data (codon configurations in different species) at a site alter the prior probabilities dramatically, such that the posterior probabilities may be very different from the prior probabilities. For example, the posterior probabilities for site 1 are 0.944, 0.056, and 0.000, and this site

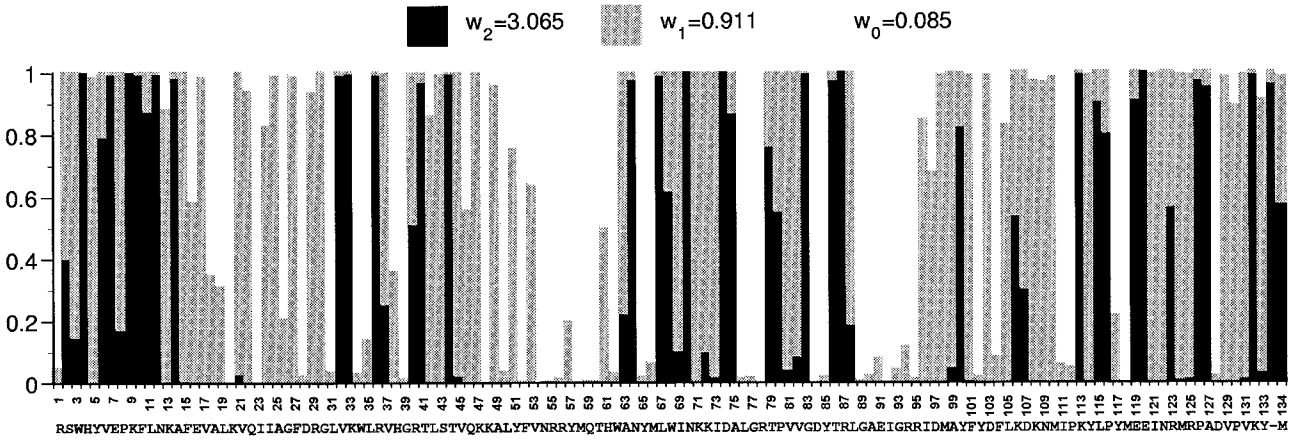

FIG. 4.—Posterior probabilities for site classes with different selective pressures (measured by the ω ratio) for amino acid sites along the sequence. The lysin sequence of the red abalone (*Haliotis rufescens*) is shown below the *x*-axis. Model M3 (discrete) is applied to the data of figure 1.

is very likely to be under strong purifying selection. The probabilities for site 4 are 0.000, 0.000, and 1.000, and this site is almost certainly under diversifying selection. The analysis was also performed under models M2 (selection) and M8 (beta & ω), and the results (not presented) were highly similar. For example, the probabilities that site 127 belongs to the class of positively selected sites (with the ω ratio being 3.82 under M2, 3.07 under M3, and 2.96 under M8; table 1) were 0.979, 0.957, and 0.950 under the three models, respectively. For site 116, the corresponding probabilities were 0.872, 0.806, and 0.802. Table 1 lists sites inferred to be under positive selection under different models at the 95% cutting point.

Sites inferred to be under diversifying or purifying selection under M3 (discrete) (table 1) are mapped onto the alpha-carbon ribbon diagram of red abalone sperm lysin in figure 5 (Shaw et al. 1993, 1995; Kresge, Vacquier, and Stout 2000*a*, 2000*b*). The 23 sites under positive selection are scattered over the entire primary sequence (fig. 4), whereas in the crystal structure (fig. 5), they cluster at the top and bottom of the molecule. Both the N- and the C-termini lie close together, and about half the sites in the N-terminus are inferred to be under positive selection. Other sites under positive selection, such as sites 14, 64, 67, 70, 74, 83, 86, and 87, cluster at the top of the protein. The majority of conserved sites are found in the internal portions of the alpha helices, where they are involved in interhelical interactions which determine the conserved fold common to all lysins.

Computer Simulation to Examine the Effect of Saturation and Functional Constraints

The pairwise comparison (fig. 2) and likelihood analysis of variable selective pressures among lineages (fig. 1) assumed that the selective pressure was the same over all amino acid sites. This assumption was unrealistic, both from biological considerations and from the analysis discussed above (tables 1 and 2). This unrealistic assumption is expected to cause underestimation of nonsynonymous rates, with greater biases at larger distances. We performed computer simulations to examine this effect, as models that account for variation of ω both among sites and among lineages have not yet been implemented.

We distinguished between two factors. The first was saturation, that is, inadequacy of the estimation procedure to correct for multiple nonsynonymous substitutions at the same site at high sequence divergences. When the substitution model is correct, the ML method does not appear to involve such a bias. Previous simulations (Yang and Nielsen 2000) suggest that the ML method tends to overestimate rather than underestimate ω in short sequences when $\omega > 1$. We generated a data set under the one-ratio model with $\kappa = 1.6$ and $\omega = 3$ and with branch lengths three times as large as those estimated from the lysin data. The sequences were more divergent and the ω ratio was much higher than in lysin. For this data set, NG underestimated the ω ratio in almost all pairwise comparisons, with the average $\omega = 1.8$. ML produced estimates around the correct value of 3 (with the average $\omega = 3.0$) although with large sampling errors at high divergences.

The second factor was variation of selective pressures or ω ratios among sites. To examine whether such a variation caused the LRT to suggest incorrectly significant differences among lineages, we simulated data sets under model M3 (discrete) using parameter estimates obtained from the lysin data under the same model (table 1). Each simulated replicate was then analyzed using two models: the one-ratio model assuming one ω for all branches, and the free-ratios model assuming one ω ratio for each branch. Both models assume no variation among sites. The statistic $2\Delta\ell$ was then calculated. As the free-ratios model is computation-intensive, only 10 replicates were simulated. In none of the 10 replicates did the LRT reject the null model of one ω ratio for all lineages. The statistic $2\Delta\ell$ ranged from 32.5 to 60.2 among replicates. Note that if the LRT is strictly

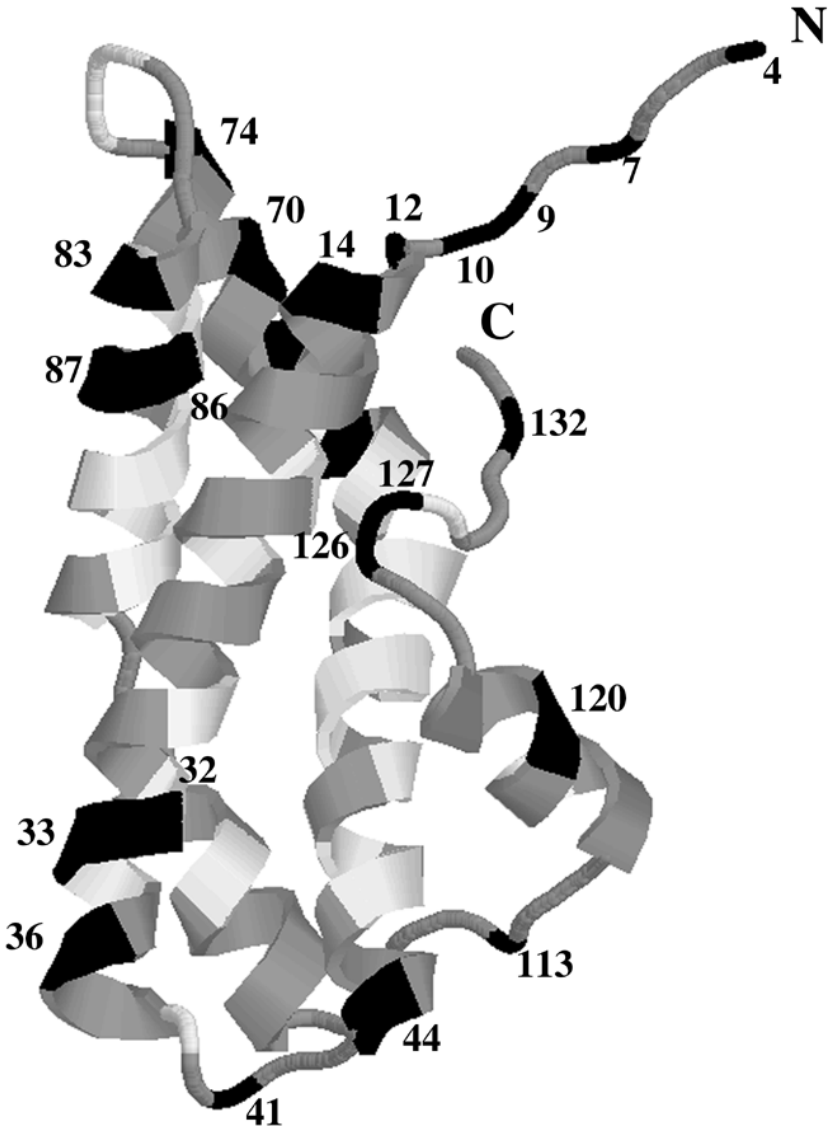

FIG. 5.—Lysin crystal structure from the red abalone *Haliotis rufescens* (Shaw et al. 1993). The structure starts from amino acid 4 (His). Sites under diversifying selection are in black and labeled, except for sites 64 and 67, which are on the face opposite that shown in the figure. Sites under purifying selection are indicated in white. The discrete model (M3) is used (see table 1).

correct, there will be a 5% chance of rejecting the null hypothesis at the 5% level.

In another set of simulations, 12 replicates were simulated under model M3 (discrete) with three classes of sites in proportions 0.2, 0.3, and 0.5 with ω ratios of 0.1, 1, and 10. The branch lengths used were three times those estimated from the lysin data. The $2\Delta\ell$ statistics ranged from 40.4 to 78.9 among replicates. The null model of one ω ratio for all lineages was incorrectly rejected in four replicates at the 1% level ($\chi^2 = 71.20$) and in two more replicates at the 5% level ($\chi^2 = 62.83$). The extreme nonsynonymous rate variation among sites and the high sequence divergence had a substantial effect on the LRT, causing far more false positives than indicated by the χ^2 distribution. Nevertheless, the observed statistic, 110.30, was well outside the range from the simulation performed under even such extreme parameter values. It is thus safe to conclude that the ω ratios are variable among lineages, even though the LRT ignores variable selective pressures at sites.

Tracing Evolutionary Changes at Amino Acid Sites by Reconstructing Ancestral Lysins

Codons and amino acids at ancestral nodes were inferred using the empirical Bayes method (Yang, Kumar, and Nei 1995) for the subtree of species 1–16 (fig. 3). The codon-based model assumed two ω ratios, one for lineages within the North Pacific and Australian clades, and another for the branch between the two clades. Ancestral amino acids were then used to map changes onto branches in the phylogeny (fig. 3). The ancestral lysin proteins may be useful for future laboratory studies.

As mentioned earlier, in two pairwise comparisons, that between *H. discus hannai* and *H. gigantea* and that

between *H. rubra* and *H. conicopora*, two nonsynonymous differences and no synonymous differences were found. These two species pairs deserve special attention. Ancestral reconstruction suggests that the ancestor of *H. discus hannai* and *H. gigantea* had GGA (Gly) at site 116 and GAA (Glu) at site 120, both with probability 1.00, and that the differences between the two species are due to a GGA (Gly) → AGA (Arg) change at site 116 and a GAA (Glu) → GCA (Ala) change at site 120 along the *H. discus hannai* lineage. The ancestor of *H. rubra* and *H. conicopora* had GCC (Ala) at site 32 with probability 1.00 and AAC (Asn) at site 74 with probability 0.97, so the differences between the two species are most likely due to a GCC (Ala) → ACC (Thr) change at site 32 and an AAC (Asn) → GAC (Asp) change at site 74 along the *H. rubra* lineage. Note that previous analysis suggests that sites 32, 74, and 120 ($P > 0.99$) and site 116 ($P \approx 0.8$) are under diversifying selection (table 1 and fig. 4). While the individual pairwise comparisons discussed earlier cannot rule out chance effects as the cause for the observed differences in the two species pairs, the combined evidence makes it most likely that those differences occurred not by chance, but by selective pressure. It would be interesting to examine the efficiency of the egg-sperm interaction within and between species for these species pairs.

Discussion

ML analysis of the lysin gene sequences reveals significant variation in selective pressure among sites. Many amino acid sites are clearly under strong purifying selection, with the nonsynonymous rate close to 0, while some other sites are under diversifying selection, with the nonsynonymous rate elevated to more than three times the synonymous rate. The analysis also suggests that the selective pressure indicated by the ω ratio is variable among lineages. Purifying selection clearly operates on conserved sites in all lineages, and variable selective pressures among lineages appear to be due to the presence or absence of diversifying selective pressure.

While the data provide support for variable selective pressures among lineages, indicating episodic evolution in lysin, it is unclear what may have caused this variation. Estimates of the ω ratios under the free-ratios model suggest that recent lineages of closely related sympatric species tend to be under diversifying selection, while old lineages separating distantly related species tend to lack such pressure. One hypothesis is that lysin evolution is driven by the selective pressure to establish cross-species reproductive barriers, possibly through reinforcement (Dobzhansky 1940). However, this explanation is contradicted by the high estimates of the ω ratio for the two branches separating the California and Japanese species, which most likely represent an allopatric speciation event.

Another factor that may cause variable ω ratios among lineages is population size fluctuation. The importance of selection relative to random drift increases with the population size. Thus, a slightly deleterious mutation will have a greater chance of getting fixed, and a slightly advantageous mutation will have a reduced chance of fixation in a smaller population (Ohta 1973). The majority of lineages with estimated $\omega < 1$ involve potential dispersal (founding) events, which might result in small population sizes. When the population is small, random fixations of mutations may be as important in lysin evolution as positive selection. Small population sizes may also lead to low animal density and low sperm concentration, and thus reduced sperm competition, a factor that may be driving lysin evolution. While differences in population size may create variable ω ratios among lineages, they do not seem likely to change the direction of selection or to produce both lineages with $\omega > 1$ and $\omega < 1$.

An additional hypothesis incorporates the evolution of VERL repeats by concerted evolution (Elder and Turner 1995). It has been hypothesized that lysin evolves to match the VERL repeats, which are being homogenized by concerted evolution (Swanson and Vacquier 1998). Positive selection may favor lysins which match the new VERL repeat or lysins which tend to conserve lysin's three dimensional structure while diversifying its amino acid side chains. Thus, VERL changes first, with the change propagated by concerted evolution, and lysin adapts to match its cognate VERL repeat while conserving its overall three-dimensional structure. The rate and extent of homogenization of repeating motifs may vary between lineages and species (Modi 1993), and there is evidence that the homogenization of VERL repeats differs between species (unpublished data). This variation in VERL homogenization rates may lead to variation in the selective pressure on lysin among lineages. For example, once lysin-VERL interaction has been optimized, purifying selection may act on lysin until a new VERL repeat type appears and is propagated by concerted evolution. Instead of being the driving force in achieving reproductive isolation (the mismatch of lysin and VERL), the species specificity of abalone fertilization may evolve as a by-product of the way in which these two cognate gamete recognition proteins evolve.

Acknowledgments

We thank Charles Aquadro, James Mallet, Nicole Kresge, and Andrew Pomiankowski for discussions. This study was supported by BBSRC grant 31/G10434 to Z.Y. and NIH grant HD12986 to V.D.V. W.J.S. is supported by an NSF/Alfred P. Sloan fellowship in Molecular Evolution.

LITERATURE CITED

BIERMANN, C. H. 1998. The molecular evolution of sperm bindin in six species of sea urchins (*Echinodea*: *Strongylocentrotidae*). Mol. Biol. Evol. **15**:1761–1771.

CIVETTA, A., and R. S. SINGH. 1998. Sex-related genes, directional selection, and speciation. Mol. Biol. Evol. **15**:901–909.

DOBZHANSKY, T. 1940. Speciation as a stage in evolutionary divergence. Am. Nat. **74**:302–321.

ELDER, J. F. J., and B. J. TURNER. 1995. Concerted evolution of repetitive DNA sequences in Eukaryotes. Q. Rev. Biol. **70**:297–320.

FERRIS, P. J., C. PAVLOVIC, S. FABRY, and U. W. GOODENOUGH. 1997. Rapid evolution of sex-related genes in *Chlamydomonas*. Proc. Natl. Acad. Sci. USA **94**:8634–8639.

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

HELLBERG, M. E., and V. D. VACQUIER. 1999. Rapid evolution of fertilization selectivity and lysin cDNA sequences in tuguline gastropods. Mol. Biol. Evol. **16**:839–848.

———. 2000. Positive selection and propeptide repeats promote rapid interspecific divergence of a gastropod sperm protein. Mol. Biol. Evol. **17**:458–466.

KRESGE, N., V. D. VACQUIER, and C. D. STOUT. 2000*a*. 1.35 and 2.07 A resolution structures of the red abalone sperm lysin monomer and dimer reveal features involved in receptor binding. Acta Crystallogr. D **56**:34–41.

———. 2000*b*. The high resolution crystal structure of green abalone sperm lysin: implications for species-specific binding of the egg receptor. J. Mol. Biol. **296**:1225–1234.

LEE, Y.-H., T. OTA, and V. D. VACQUIER. 1995. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. Mol. Biol. Evol. **12**:231–238.

LEE, Y.-H., and V. D. VACQUIER. 1995. Evolution and systematics in *Haliotidae* (*Mollusca*: *Gastropoda*): inferences from DNA sequences of sperm lysin. Mar. Biol. **124**:267–278.

LI, W.-H. 1997. Molecular evolution. Sinauer, Sunderland, Mass.

LYON, J. D., and V. D. VACQUIER. 1999. Interspecies chimeric sperm lysins identify regions mediating species-specific recognition of the abalone egg vitelline envelope. Dev. Biol. **214**:151–159.

METZ, E. C., and S. R. PALUMBI. 1996. Positive selection and sequence arrangements generate extensive polymorphism in the gamete recognition protein bindin. Mol. Biol. Evol. **13**: 397–406.

METZ, E. C., R. ROBLES-SIKISAKA, and V. D. VACQUIER. 1998. Nonsynonymous substitution in abalone sperm fertilization genes exceeds substitution in introns and mitochondrial DNA. Proc. Natl. Acad. Sci. USA **95**:10676–10681.

MODI, W. S. 1993. Heterogeneity in the concerted evolution process of a tandem satellite array in meadow mice (*Microtus*). J. Mol. Evol. **37**:48–56.

NEI, M., and T. GOJOBORI. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.

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

OHTA, T. 1973. Slightly deleterious mutant substitutions in evolution. Nature **246**:96–98.

PALUMBI, S. R. 1994. Genetic divergence, reproductive isolation and marine speciation. Annu. Rev. Ecol. Syst. **25**:547–572.

PALUMBI, S. R., and E. C. METZ. 1991. Strong reproductive isolation between closely related tropical sea urchins (genus *Echinometra*). Mol. Biol. Evol. **8**:227–239.

SHAW, A., P. A. FORTES, C. D. STOUT, and V. D. VACQUIER. 1995. Crystal structure and subunit dynamics of the abalone sperm lysin dimer: egg envelopes dissociate dimers, the monomer is the active species. J. Cell Biol. **130**:1117–1125.

SHAW, A., D. E. MCREE, V. D. VACQUIER, and C. D. STOUT. 1993. The crystal structure of lysin, a fertilization protein. Science **262**:1864–1867.

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

———. 1998. Concerted evolution in an egg receptor for a rapidly evolving abalone sperm protein. Science **281**:710–712.

TSAUR, S. C., and C.-I. WU. 1997. Positive selection and the molecular evolution of a gene of male reproduction, *Acp26Aa* of *Drosophila.* Mol. Biol. Evol. **14**:544–549.

VACQUIER, V. D., and Y.-H. LEE. 1993. Abalone sperm lysin: unusual mode of evolution of a gamete recognition protein. Zygote **1**:181–196.

VACQUIER, V. D., W. J. SWANSON, E. C. METZ, and C. D. STOUT. 1999. Acrosomal proteins of abalone spermatozoa. Adv. Dev. Biochem. **5**:49–81.

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

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

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

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

———. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. **17**:32–43.

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

MANOLO GOUY, reviewing editor

Accepted May 29, 2000