Molecular evolution of dinoflagellate luciferases, enzymes with three catalytic domains in a single polypeptide

Liyun Liu, Thérèse Wilson, and J. Woodland Hastings*

Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138

Contributed by J. Woodland Hastings, October 15, 2004

Enzymes with multiple catalytic sites are rare, and their evolutionary significance remains to be established. This study of luciferases from seven dinoflagellate species examines the previously undescribed evolution of such proteins. All these enzymes have the same unique structure: three homologous domains, each with catalytic activity, preceded by an N-terminal region of unknown function. Both pairwise comparison and phylogenetic inference indicate that the similarity of the corresponding individual domains between species is greater than that between the three different domains of each polypeptide. Trees constructed from each of the three individual domains are congruent with the tree of the full-length coding sequence. Luciferase and ribosomal DNA trees both indicate that the *Lingulodinium polyedrum* luciferase diverged early from the other six. In all species, the amino acid sequence in the central regions of the three domains is strongly conserved, suggesting it as the catalytic site. Synonymous substitution rates also are greatly reduced in the central regions of two species but not in the other five. This lineage-specific difference in synonymous substitution rates in the central region of the domains correlates inversely with the content of GC_3, which can be accounted for by the biased usage toward C-ending codons at the degenerate sites. RNA modeling of the central region of the *L. polyedrum* luciferase domain suggests a function of the constrained synonymous substitutions in the circadian-controlled protein synthesis.

codon bias | domain duplication | synonymous substitutions

The luciferase (LCF) of the bioluminescent dinoflagellate *Lingulodinium polyedrum* (*Lp*; formerly *Gonyaulax polyedra*) has three homologous tandemly arranged 377-aa-long domains, each of which is catalytically active (1). The LCF of another dinoflagellate, *Pyrocystis lunula* (*Pl*), has a similar structure (2). In both, we found the amino acids in the central region of each of the three domains to be highly conserved intramolecularly at ≈95% identity, compared with ≈65% in the flanking regions. Remarkably, and only in *Lp*, there were fewer synonymous substitutions in the central region.

These results are intriguing from both evolutionary and mechanistic viewpoints. To grasp their significance, we investigated the LCFs of five other bioluminescent dinoflagellates, all of which are photosynthetic: *Alexandrium affine* (*Aa*), *Alexandrium tamarense* (*At*), *Pyrocystis fusiformis* (*Pf*), *Pyrocystis noctiluca* (*Pn*), and *Protoceratium reticulatum* (*Pr*). All were shown to have the same basic configuration, i.e., tandem triplication of catalytic domains following an ≈110-aa-long N-terminal sequence. Fortuitously, it turned out that *Lp* and *Pl* LCFs represent extremes of a spectrum of LCF structures with regard to synonymous substitution rates in the central region of each domain.

Bioluminescence in dinoflagellates has been well characterized only in *Lp*, where the three components involved in light emission are packaged together in unique cellular organelles, termed scintillons, with flashing triggered by a pH change from ≈8 to ≈6. LCF catalyzes the oxidation of the luciferin substrate (LH_2), a linear tetrapyrrole, generating oxyluciferin in an excited state and subsequent light emission. In the scintillon, LH_2 is sequestered by a luciferin binding protein (LBP) at pH 8 but free at pH 6. The activity of LCF itself is also pH-dependent, because of four intramolecularly conserved histidines (3). The bioluminescence is circadian, with daily fluctuations in the abundance of LCF, LBP, LH_2, and the number of scintillons (4–6). Protein synthesis of LCF and LBP are clock-regulated at the translation level.

Although all seven dinoflagellates studied here are believed to have scintillons and emit light rhythmically, only *Lp* and three other species, *Aa*, *At*, and *Pr* (L.L., unpublished data), are known to have LBP. Also, *Lp* is the only species in which the abundance of LCF and the number of scintillons have been demonstrated to undergo a daily cycle. These seven species are able to use the same luciferin and, as shown in the present study, their LCFs all have the four histidine residues important for pH regulation, indicating that bioluminescence in these species is pH-triggered by the same mechanism.

Materials and Methods

Cultures. Cells were grown in F2 medium (38) under a 12 h/12 h light/dark cycle at 19°C. From the Provasoli-Guillard National Center for Culture of Marine Phytoplankton (CCMP; Bigelow Lab, Boothbay Harbor, ME), we obtained *Aa* (CCMP 112), *At* (CCMP 1493), and *Pr* (CCMP 1889). *Pf* was kindly provided by Carrie McDougal (University of California, Santa Barbara). Cells were harvested by vacuum filtration on Whatman paper 51 or by centrifugation at 5,000 × *g* for 10 min.

Molecular Cloning. DNA and RNA were isolated by using a phenol-based method (7). Partial LCF sequences were PCR-amplified from genomic DNA with two pairs of degenerate primers. The first pair, 5′-TGYAARGGNTTYGAY-TAYGG-3′ and 5′-GCNGAYTCCATYTCCCARAA-3′, corresponds to amino acid sequences CKGFDYG and FWEMESG, respectively, which are located at the beginning and the end of the most conserved central region of the three intramolecular domains of the *Lp* LCF. The second pair, 5′-GASTAT-GAGAASCGNGTNCRRCATGA-3′ and 5′-TGRCCRCC-NGAGACNGTGATCCA-3′, is derived from the amino acid sequences DYENGVRDD and WITVSGGQ. The former is present in the N-terminal region of *Lp* LCF and LBP, and the

Abbreviations: *At*, *Alexandrium tamarense*; *Aa*, *Alexandrium affine*; *Lp*, *Lingulodinium polyedrum*; *Pf*, *Pyrocystis fusiformis*; *Pl*, *Pyrocystis lunula*; *Pn*, *Pyrocystis noctiluca*; *Pr*, *Protoceratium reticulatum*; *Ns*, *Noctiluca scintillans*; GST, glutathione *S*-transferase; LCF, luciferase; LBP, luciferin binding protein; LH_2, luciferin; ENC, effective numbers of codons; CBI, codon bias index; GC_3, GC content at the third position of codons; K_s, synonymous substitutions per site; K_a, nonsynonymous substitutions per site.

Data deposition: The LCF sequences for the five previously undescribed species reported in this article have been deposited in the GenBank database [accession nos. AY766382 (*Aa*), AY766382 (*At*), AY766384 (*Pf*), AY766385 (*Pn*), and AY766386 (*Pr*)].

*To whom correspondence should be addressed. E-mail: hastings@fas.harvard.edu.

latter is present in the conserved central region of the three domains in LCF. Based on the partial sequences, specific primers were synthesized to clone 5′ and 3′ cDNA ends by PCR with a rapid amplification of cDNA ends (RACE) kit (Invitrogen) (for *Pf*), screening libraries (for *At* and *Pn*), or inverse PCR with *Apa*I-digested circular DNA (for *Aa* and *Pr*) (8).

Tree-Building Methods. DNA sequence alignment was performed with the program CLUSTALW (39). Selection and evaluation of DNA substitution models, rate parameters, and base frequencies were carried out with MODETEST (9) and MAPPS (10). The distant trees of the full-length ORFs and ribosomal RNA were constructed with MEGA (Version 2.1) (11). The maximum-likelihood tree of the individual domains and the parsimony tree of the N-terminal sequences were generated with PHYLIP 3.6B (12) and PAUP (Version 4b) (13), respectively. For all trees, bootstrap values were estimated from 1,000 replicates, and tree files were processed with the program TREEVIEW (14).

Calculation of Synonymous and Nonsynonymous Substitutions, Effective Numbers of Codons (ENC), Codon Bias Index (CBI), and GC_3. The cumulative numbers of synonymous and nonsynonymous differences were obtained with SNAP (www.hiv.lanl.gov/content/hiv-db/SNAP/WEBSNAP/SNAP.html). The rates of synonymous and nonsynonymous substitutions were estimated with DNASP 4.05 (15) and K-ESTIMATOR 6.0 (16). ENC, CBI, and GC_3 all were calculated with DNASP. The software for computing the base usage at the third position of codons was developed by Xianghui Liu (Harvard University).

Results

All Dinoflagellate LCFs Have Three Catalytic Domains in a Single Polypeptide. We sequenced LCF genes from five bioluminescent species, *Aa*, *At*, *Pn*, *Pf*, and *Pr*, and compared them with the sequences of *Lp* and *Pl*. The seven LCFs are similar at both genome and protein levels (see Fig. 5, which is published as supporting information on the PNAS web site). All exist as tandem repeats separated by an intergenic region of ≈0.3 to ≈3.0 kb, depending on the species (Fig. 1). Although the noncoding regions share little similarity (except for those of the two closely related *Alexandrium* species), the coding sequences are highly conserved. The predicted proteins all possess the same organization, in which a single polypeptide consists of four parts: an N-terminal region (≈100 residues) of unknown function followed by three contiguous and homologous catalytic domains.

Lp LCF does not resemble any other sequences in the GenBank database, except for its N-terminal region, where similar sequences occur in comparably located regions of both LBP and glutathione *S*-transferase (GST) (17), with sequence identities of 46% and 31%, respectively, compared with *Lp* LCF. The sequence identities of the seven LCF genes in this region average 60%, ranging from 47% between *At* and *Pn* to 77% between the two *Alexandrium* species.

In the rest of the LCF sequence, which embraces the three repeated domains, the identities are similar across species, averaging ≈84%. When the individual domains are considered separately, the values for sequence identities are also ≈84%. However, the identities between different domains of any given species are lower (≈73%).

The extent of intramolecular amino acid sequence conservation is not uniform within different regions of each of the three individual domains. According to the degree of sequence conservation, each individual catalytic domain can be subdivided into three major parts: the well conserved central part of ≈150 residues (>95% amino acid similarity) and the less conserved N- and C-terminal parts of 101–125 residues each (≈75% amino acid similarity) (Fig. 1). The more conserved central region is believed to form the catalytic structural fold and the luciferin-binding site, as suggested also by the crystal structure, deletion mapping, and point mutagenesis (W. Schultz, L.L., and J.W.H., unpublished data).

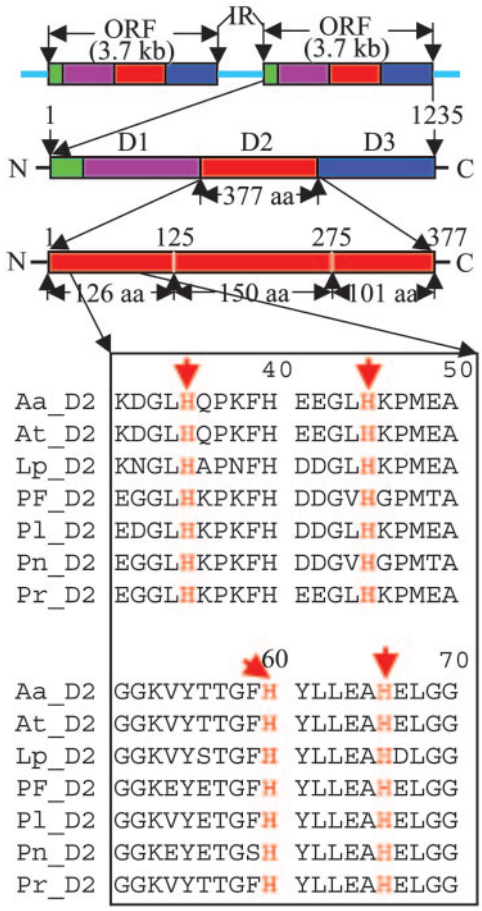

Fig. 1. Bar representation of the organization (*Upper*) and partial sequence alignments (*Lower*) of dinoflagellate LCFs. (*Upper Top*) Two copies of LCF genes arranged in tandem, separated by an intergenic region. (*Upper Middle*) The structure of *Lp* LCF, showing the N-terminal region followed by three repeated domains, D1, D2, and D3. (*Upper Bottom*) A diagram of the second domain of *Lp* subdivided into three regions: the N-terminal, the central, and the C-terminal. The central region (126–275) is more conserved than the flanking ones and is likely the catalytic core. (*Lower*) Alignment of sequences 30–70 of the N-terminal second domains of seven LCFs showing the four conserved histidines at positions 35, 45, 60, and 66.

As noted above, the regulation of LCF activity in *Lp* by pH has been attributed to four histidines in the N-terminal region of each of the domains (3). They are found in all three domains of all seven LCF genes, as shown for D2 (Fig. 1).

Phylogenetic Analysis Suggests an Early Origin of *L. polyedrum* LCF. Given the unique structure of dinoflagellate LCF, we constructed phylogenetic gene trees from three sets of DNA sequences: the full-length coding region, the individual domains, and the N-terminal regions. The tree from the full-length ORFs resolves the LCF genes of seven species into three clades (Fig. 2*A*). The first comprises the genes from three *Pyrocystis* species, in which the gene of *Pn* is closer to *Pf* than to *Pl*, and the second comprises two *Alexandrium* species and *Pr*. *Lp* itself constitutes the third clade, being approximately equidistant from the other two.

The three intramolecularly conserved domains were treated as separate units to construct the so-called domain tree, showing that the individual domains of the LCFs from different species

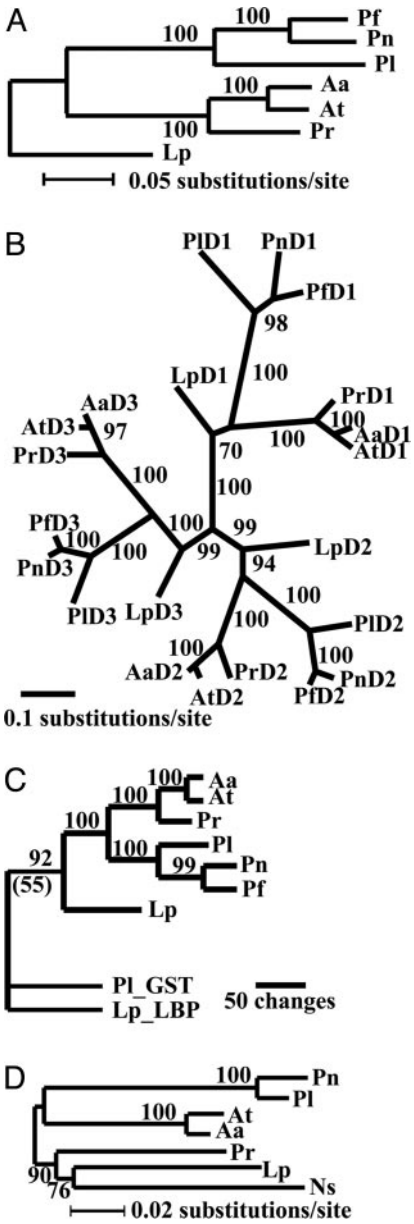

Fig. 2. Molecular phylogeny of dinoflagellate LCFs based on their nucleotide sequences. Numerals are the percentage of the bootstrap values. (*A*) Distance tree of the full-length coding sequences constructed by using the neighbor-joining method, separating seven LCF genes into three clades. The first includes LCFs from two *Alexandrium* species and *Pr*; the second comprises three *Pyrocystis* LCFs; and the third is made up of *Lp* LCF as an orphan. (*B*) Maximum likelihood tree (natural logarithmic likelihood = −11505.72613) of the catalytic domains. The corresponding domains of different LCFs group together, and the domains of *Lp* separate early from those of the other LCFs. (*C*) Parsimony tree (tree length = 694; consistency index = 0.7334 and retention index = 0.5727) of the N-terminal domains of the seven LCFs and of *Lp*_LBP and *Pl*_GST. The tree was generated using PAUP 4.0B. The bootstrap value in support of *Lp* LCF, *Lp*_LBP, and *Pl*_GST as a clade was lower (in parentheses) when the data were analyzed with maximum likelihood method. (*D*) Small subunit ribosomal RNA tree for the seven species studied here along with *Noctiluca scintillans* (*Ns*).

group together more closely than different domains of the same LCF (Fig. 2*B*). The relationship of the LCF from different species is reflected in each group; for example, the three *Pyrocystis* LCFs always are grouped together whether they are in the group formed by D1, D2, or D3. The three groups converge

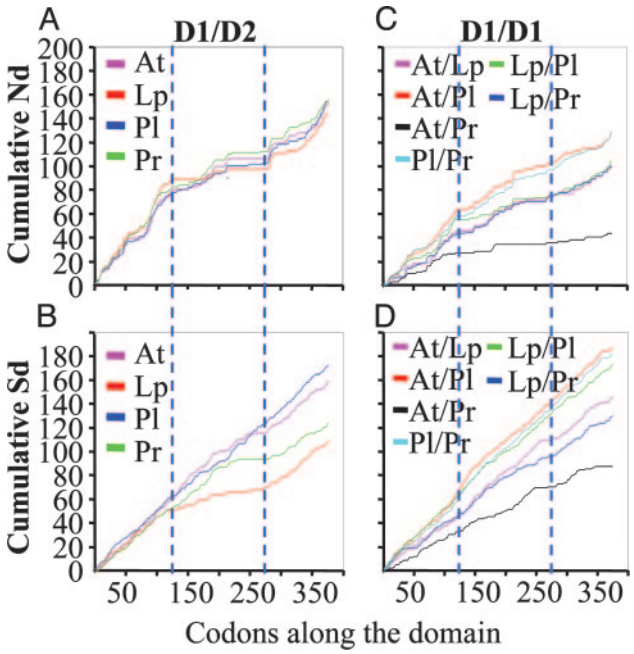

Fig. 3. Cumulative numbers of synonymous and nonsynonymous substitutions along the domains; numbers were counted with the SNAP program. For intramolecular comparisons, only D1 and D2 are shown; similar results were obtained for D1/D3 and D2/D3. For intermolecular comparisons, only the first domain pairs are shown; the other two gave the same result. (*A*) The changes in the cumulative numbers of the nonsynonymous substitution rates between the domains are similar along the domains for all LCFs, with pronounced reductions in the central regions. (*B*) The cumulative numbers of synonymous substitutions for intramolecular domains, especially for the first domain as compared with one of the other two, are lower in the central region for *Lp* and *Pr* LCFs but not for others. (*C* and *D*) Similar comparisons also were carried out pairwise for corresponding domains of different LCFs. Both the cumulative numbers of nonsynonymous substitutions and those of synonymous substitutions increase at about the same rate along the entire domain; the absolute numbers are smaller for the more closely related LCFs, such as *At* vs. *Pr*.

around *Lp* LCF, suggesting that it diverged early from the other six.

The tree based on the N-terminal sequences of *Lp* LBP, *Pl* GST, and all of the LCFs places the first two together (Fig. 2*C*). The N-terminal sequence of *Lp* LCF is firmly placed with those of LCFs of other species by PAUP analysis but only weakly by distant and likelihood analysis at bootstrap values of ≈55%.

A tree of the small subunit ribosomal RNA sequences, including the sequence of *Noctiluca scintillans* (*Ns*) as an outgroup, was consistent with the LCFs (Fig. 2*D*). It has been established that *Ns*, a heterotrophic dinoflagellate, represents the basal branch in dinoflagellates by using molecular sequences as well as morphological characters (18).

Synonymous Substitution Rates Within the Intramolecular Domains Vary Greatly Among Different LCFs. Cumulative numbers of both nonsynonymous and synonymous substitutions between pairs of domains in a given LCF were counted and plotted against codon number. By comparing nonsynonymous substitutions between the intramolecular domains pairwise for the different LCFs, we saw very striking regions of conservation in all cases, especially in the central parts of all of the domains, from about codon 126 to 275 (Fig. 3*A*). This finding reasonably may be attributed to the fact that the active site is located in this region. But in all cases, there is also a less extensive island of conservation centered around codon 65, possibly related to the portion of the molecule associated with regulation by pH.

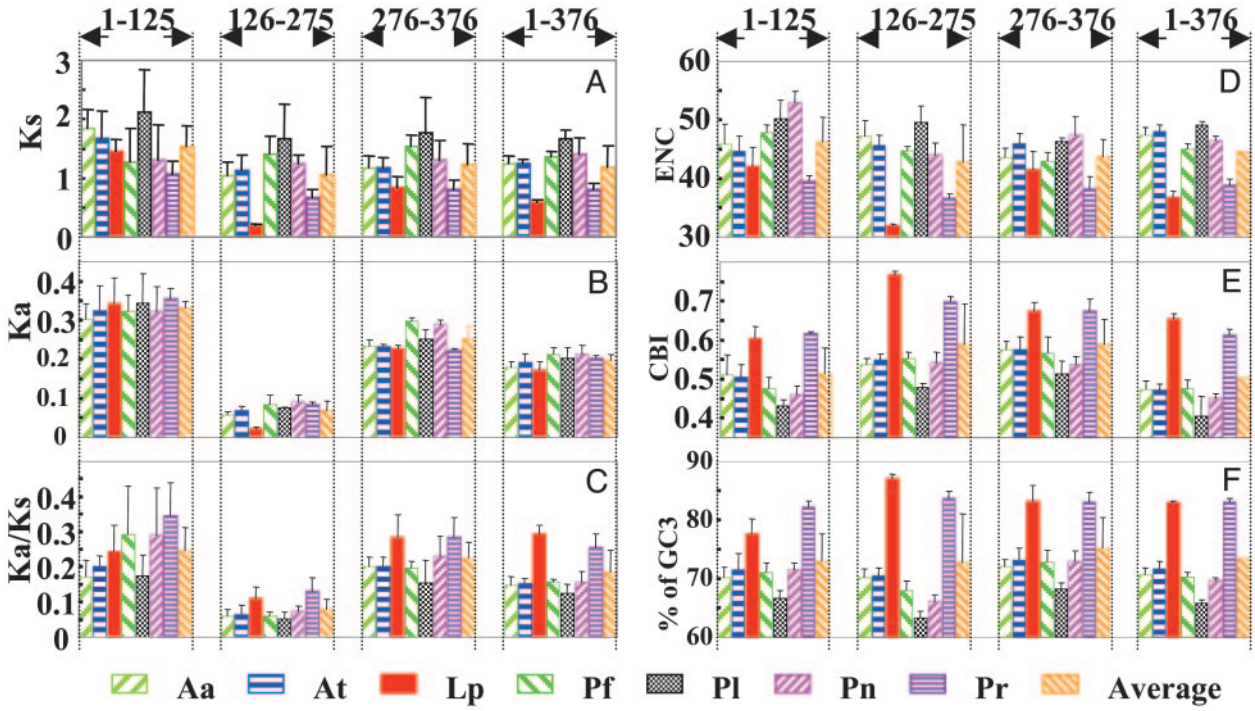

Fig. 4. Relationship between synonymous substitution rates and codon bias. The three intramolecular domains (D1, D2, and D3) are compared pairwise for entire ORFs (1–376) or as three regions, the N-terminal (codon 1–125), the central (codon 126–275), and the C-terminal (codon 276–376). For each LCF, the values from three pairwise comparisons were averaged, and their standard deviations were computed. Regional analysis of synonymous (K_s) and nonsynonymous (K_a) substitution rates and their ratios are shown in *A*, *B*, and *C*, respectively. Note that the scales on the *y*-axis are different for K_s and K_a. The codon bias is measured in ENC and CBI (*D* and *E*). GC_3 contents for each region are indicated in *F*. All values were calculated with the DNASP package. The values in each category are all arranged from left to right for the LCFs from *Aa*, *At*, *Lp*, *Pf*, *Pl*, *Pn*, *Pr*, and the averages.

The frequencies of synonymous substitutions in the central subdomain of *Lp* LCF, and to a lesser degree those of *Pr* LCF, are also considerably lower than in the flanking regions, whereas in the other LCFs, the rates do not differ very much over the length of the domains (Fig. 3*B*).

The corresponding domains of the LCFs were compared similarly for different pairs of species. Changes in synonymous substitution rates are more or less constant over the full domain lengths, with the cumulative numbers (thus the slope of the line) being lower in more closely related LCFs, such as *At* and *Pr* (Fig. 3*D*). Similar patterns are true for the nonsynonymous substitutions, but in the region of the central subdomain, the number is slightly less in some pairs than those in the flanking regions, especially for *At* vs. *Pr* (Fig. 3*C*).

The heterogeneity described above for synonymous and nonsynonymous substitutions in different regions of the intramolecular conserved domains can be more accurately expressed in terms of the values of K_s (synonymous substitutions per site), K_a (nonsynonymous substitutions per site), and the ratio of K_a to K_s. In these cases, multiple hits were corrected using Jukes–Cantor's one-parameter (19) or Kimura's two-parameter (20) methods. The K_s values vary with lineage, with *Lp* lowest, *Pr* medium, and *Pl* highest, whether the entire domain or the subdomains are considered (Fig. 4*A*). Most strikingly, the K_s value of the central region of the domain in *Lp* LCF is >8-fold lower than in *Pl* LCF, but only 2-fold lower or less if the N- or C-terminal regions are compared. Values of K_a and K_a/K_s for all LCFs are lower in the central region (codon 126–275) (Fig. 4 *B* and *C*). The K_a/K_s value suggests that purifying selection acts more strongly on that region.

Codon Bias and Differences in Synonymous Substitution Rates. Differences in synonymous substitution rates can result from a biased codon usage. One measure for such a bias is the ENC; if all codons are indifferently used, ENC is 61, whereas if one codon is used exclusively for each amino acid, ENC is 20 (21). ENC values for the central regions of the domains in *Lp* and *Pr* LCFs (codons 126 to 275) are ≈32 and 37, respectively, but they are 45–50 for other LCFs (Fig. 4*D*). Another measure for biased codon usage, CBI, quantifies the extent to which a particular set of codons is selectively used (22). *Lp* and *Pr* LCFs have much greater values of CBI (Fig. 4*E*). Together, ENC and CBI establish that some codons are favored over others in the central region of catalytic domains of *Lp* and *Pr* LCFs. Our previous analysis of codon usage in *Lp* LCF did not include codons that are not the same in all three domains (1), and so it did not reveal the true codon bias between the central and flanking regions. Although we had observed correctly that the same sets of favored codons are used for both regions, the conclusion that the biased codon usage is not the factor responsible for the differential synonymous substitution rate was an artifact of our computation choice.

The percentages of GC_3 in the central region of *Lp* and *Pr* LCFs are 87% and 84%, respectively, whereas they are <72% in the central region of the other LCFs (Fig. 4*F*). We tabulate the frequency of each individual base in the central region, revealing that the percentages of cytosine-ending (but not guanine) bases vary among the different LCFs. The percentages of C-ending codons negatively correlate with the synonymous substitution rates in the central region of different LCFs (Fig. 4*A*), with the highest cytosine value of nearly 58% for *Lp* LCF, the medium value of 49% for *Pr* LCF, and the lowest value of 36% for *Pl* (Table 1). Close examination of the usage of individual codons shows that degenerate codons coding for glycine, serine, and proline are predominantly those ending in cytosine bases in *Lp* and *Pr* LCFs, but not others. For example, among the four

Table 1. Base composition of the third position of the triplet codons in the central regions

Species	G	A	T	C
Aa	30.2	9.3	18.9	41.6
At	30.7	8.7	19.1	41.6
Lp	30.2	2.7	9.6	57.6
Pf	30.7	8.4	22.0	38.9
Pl	29.3	9.3	25.3	36.0
Pn	30.0	8.0	24.0	38.0
Pr	34.4	3.8	11.6	50.2

codons coding for serine (TCA, TCG, TCC, and TCT), TCC accounts for ≈83%, 65%, 41%, and 35% in the central region for *Lp*, *Pr*, *Pf*, and *Pl* LCFs, respectively.

Discussion

Origin and Evolution of Dinoflagellate LCFs. The comparison of LCF genes from seven dinoflagellates reveals that they are highly conserved and share a common origin. This finding is in agreement with previous results showing cross-reactions between LCF and luciferin from different species by the criteria of biochemical assays, immunoreactivity, and DNA hybridizations (23, 24).

Our results can be summarized as follows. (*i*) All seven LCFs have three tandem, conserved domains. (*ii*) In each of these three domains, a central region of ≈150 aa is especially conserved. (*iii*) In *Lp* and *Pr*, the number of synonymous substitutions in these central subdomains is also low, nearly 10 times less than in the border regions. (*iv*) The degree of intramolecular homology between the three domains of a given LCF is lower than between corresponding domains of different LCFs. (*v*) Phylogenetic analysis, based on either LCF or ribosomal DNA sequences, indicates that *Lp* LCF diverged early from the other six.

One scenario that might explain our results assumes that gene triplication, creating the three active LCF domains, occurred only once, in a species ancestral to all. The entire LCF gene subsequently was retained in descendents where luminescence provided a selective advantage under strong selective pressure for preserving catalytic activity, thereby restricting the frequency of nonsynonymous substitutions in the central part of each domain (Fig. 4*B*).

However, a second selective pressure also must have originated at one point, possibly after *Lp* and *Pr* diverged, resulting in a strikingly low rate of synonymous substitutions, as is now found in the central subdomains of *Lp* LCF and, to a lesser extent, in *Pr*. Whatever this pressure might have been, it apparently did not arise or later disappeared during the course of evolution of the other species.

Functional implications of this constraint can be inferred from theoretical calculations of RNA secondary structures. As shown in Table 2, the stabilities of the RNA secondary structures of both the central regions and the full domains of the *Lp* and *Pr* LCFs are greater than those of the others by ≈10 and ≈30 kcal/mol, respectively. Stability could be important for a regulatory mechanism, such as binding of a protein, antisense RNAs, small RNAs, or other small molecules, possibly for circadian control, which is known to be mediated by translational control in *Lp* (24). Conversely, in *Pl*, where the rate of synonymous substitutions is the highest of all seven species and uniform along the domains, the abundance of LCF does not change along the daily cycle.

Pros and Cons of Multiple Catalytic Domains in a Single Peptide. What benefits might result from having three homologous domains in a single polypeptide? Based on the fact that the proteins are packaged at a high concentration in the small scintillon organelles, we previously proposed that up to a 3-fold greater enzyme activity could be obtained without an increase in the colloidal osmotic pressure of the scintillon (1). Fluorescence microscopy shows that scintillon-like structures are present in all seven species (L.L., unpublished data). Another possibility is that, even though individual domains exhibit a pH-dependent activity, the pH-regulatory mechanism might be more effective with the three-domain structure.

Several other enzymes are known to contain from two to four internal repeats, with some evident or inferred functional advantages. Among them are several kinds of kinases (25–29) and polysaccharide-degrading enzymes such as chitinases (30–32), xylanases (33, 34), and endoglucanases (35). Homology between the domains varies from ≈30% to 99%; activities of the individual domains have not been demonstrated in all cases, and pH and temperature optima, as well as substrate preference, have been reported to differ among the domains. Hydrolases are the third class of multidomain enzymes. Rat intestinal lactase-phlorizin hydrolase is of special interest because none of its four tandem duplicated domains (with average similarities of ≈50%) is catalytically active alone, whereas a construct bearing domains III and IV is active. Domains I and II function as an intramolecular chaperone to promote the maturation of the enzyme-containing domains III and IV (36). Enzymes with more than four internal domain duplications have not been reported.

Table 2. Predicted free energy levels for the RNA secondary structures folded from either the central region of the domains or the full domain

Location	*Aa*	*At*	*Pr*	*Lp*	*Pf*	*Pl*	*Pn*
Central region							
D1	−170.2 ± 0.6	−168.9 ± 0.6	−180.9 ± 0.4	−180.8 ± 0.4	−156.6 ± 0.5	−168.6 ± 0.2	−164.2 ± 0.3
D2	−177.4 ± 0.6	−171.0 ± 1.4	−184.1 ± 0.7	−172.8 ± 0.8	−168.3 ± 0.2	−164.9 ± 0.5	−167.5 ± 0.6
D3	−166.2 ± 0.8	−175.6 ± 1.5	−184.1 ± 0.7	−182.6 ± 0.7	−186.2 ± 0.6	−163.7 ± 0.5	−177.0 ± 0.3
Avg	−171.1 ± 5.7	−171.8 ± 3.4	−181.4 ± 2.4	−178.8 ± 5.3	−170.4 ± 14.9	−165.7 ± 2.6	−169.6 ± 6.7
Full domain							
D1	−450.2 ± 2.2	−421.5 ± 0.5	−460.4 ± 0.6	−478.1 ± 0.5	−417.2 ± 1.7	−436.9 ± 1.4	−447.0 ± 0.3
D2	−432.2 ± 0.9	−431.4 ± 0.4	−491.6 ± 0.6	−460.3 ± 0.5	−424.8 ± 0.4	−406.3 ± 0.9	−422.0 ± 0.5
D3	−429.7 ± 1.7	−435.4 ± 0.5	−463.3 ± 0.3	−450.2 ± 0.9	−444.9 ± 0.7	−407.7 ± 0.5	−448.9 ± 1.7
Avg	−437.4 ± 11.2	−429.4 ± 7.2	−471.8 ± 17.2	−462.9 ± 14.1	429.0 ± 14.3	−417.0 ± 17.3	−439.3 ± 15.0

The mRNA sequences of the central regions (450 bases) or the full sequences of the individual domains from seven LCFs were imported into MFOLD (Version 4.0) for predicting the secondary structures (http://bioweb.pasteur.fr/seqanal/interfaces/mfold-simple.html). The five structures with the lowest free energy were used to obtain the average values for each sequence. Avg, averaged value of the three domains. The window sizes for folding the central regions and the full domains are 7 and 15, respectively. The predicted structures were viewed and printed with XRNA (Version 1.1.10; http://rna.ucsc.edu/rnacenter/xrna/xrna.html).

The above examples serve to illustrate that the architectures, intramolecular interactions, and functional consequences of the repeated domains are diverse, and, in some cases, interactions between internal domains are advantageous, as in the lactase-phlorizin hydrolase just discussed. If so, why are there so few enzymes with the multidomain configuration as compared with gene duplications, which are widespread? Several features favoring gene duplication can be cited. First, it allows not only the duplication and divergence of the coding sequences but also of the regulatory sequences, so that differential gene expression can be achieved both temporally and spatially, giving more flexibility, whereas with domain duplication, gene expression is controlled by the same promoter for all domains. Second, for gene duplication, the number of copies can be very large. For example, the rice GST family consists of 59 putative genes and 2 pseudogenes (37). Conversely, increasing domain duplications ultimately will increase the mRNA length to a point where transcription and translation efficiencies could be compromised. Finally, in the case of domain duplication, any frameshift mutation in the most upstream domain will result in activity loss for that and all downstream domains.

The origin of domain duplications for some of the enzymes discussed above is clearly traceable. The mammalian hexokinases and lactase-phlorizin hydrolases have multiple catalytic domains, whereas their microbial counterparts have only one. Because all seven dinoflagellate LCFs carry the three domains, the duplication events must have occurred before the diversification of these species.

To conclude, our study of dinoflagellate LCFs describes the previously unexamined evolution of intramolecularly duplicated domains, each with catalytic activity. We showed that the amino acid sequences of the intramolecular repeats are more divergent within the gene of a single species than between the corresponding domains of those of different species. We also found the intramolecular synonymous substitution rates to vary among the different species and even in some species (*Lp* and *Pr*) along the domains. In *Lp*, we propose that the dramatically reduced synonymous substitutions in the central regions may relate to a specific secondary RNA structure functioning in the rhythmic control of LCF synthesis.

We are grateful to Drs. Sarah Mathews, Daniel Hartl, and Liming Li for advice and critical readings of the manuscript. This work was supported in part by National Science Foundation Grant MCB-0343407 and Office of Naval Research Grant N00014-03-1-0173.

1. Li, L., Hong, R. & Hastings, J. W. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 8954–8958.
2. Okamoto, O. K., Liu, L., Robertson, D. L. & Hastings, J. W. (2001) *Biochemistry* **40,** 15862–15868.
3. Li, L., Liu, L., Hong, R., Robertson, R. L. & Hastings, J. W. (2001) *Biochemistry* **40,** 1844–1849.
4. Johnson, C. H., Roeber, J. & Hastings, J. W. (1984) *Science* **223,** 1428–1430.
5. Morse, D., Milos, P. M., Roux, E. & Hastings, J. W. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 172–176.
6. Fritz, L., Morse, D. & Hastings, J. W. (1990) *J. Cell Sci.* **95,** 321–328.
7. Bugos, R. C., Chiang, V. L., Zhang, X. H., Campbell, E. R., Podila, G. K. & Campbell, W. H. (1995) *BioTechniques* **19,** 734–737.
8. Digeon, J. F., Guiderdoni, E., Alary, R., Michaux-Ferriere, N., Joudrier, P. & Gautier, M. F. (1999) *Plant Mol. Biol.* **39,** 1101–1112.
9. Posada, D. & Crandall, K. A. (1998) *Bioinformatics* **14,** 817–818.
10. Bollback, J. (2002) *Mol. Biol. Evol.* **19,** 1171–1180.
11. Kumar, S., Tamura, K. & Nei, M. (1994) *Comp. Appl. Biosci.* **10,** 189–191.
12. Felsenstein, J. (1993) *Inferring Phylogenies* (Sinauer, Sunderland, MA).
13. Swofford, D. L. (2003) PAUP* *Phylogenetic Analysis Using Parsimony (*and other methods)* Version 4.0b (Sinauer, Sunderland, MA).
14. Page, R. D. M. (1996) *Comp. Appl. Biosci.* **12,** 357–358.
15. Rozas, J., Sanchez-DelBarrio, J. C., Messeguer, X. & Rozas, R. (2003) *Bioinformatics* **19,** 2496–2497.
16. Comeron, J. M. (1999) *Bioinformatics* **15,** 763–764.
17. Okamoto, O. K. & Hastings, J. W. (2003) *J. Phycol.* **39,** 519–526.
18. Edvardsen, B., Shalchian-Tabrizi, K., Jakobsen, K. S., Medlin, L. K., Dahl, E., Brubak, S. & Paasche, E. (2003) *J. Phycol.* **39,** 395–408.
19. Jukes, T. H. & Cantor, C. R. (1969) in *Mammalian Protein Metabolism*, ed. Munro, H. N. (Academic, New York), pp. 21–123.
20. Kimura, M. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 454–458.
21. Wright, F. (1990) *Gene* **87,** 23–29.
22. Sharp, P. M. & Li, W. H. (1987) *Nucleic Acids Res.* **15,** 1281–1295.
23. Schmitter, R. E., Njus, D., Sulzman, F. M., Gooch, V. D. & Hastings, J. W. (1976) *J. Cell. Physiol.* **87,** 123–134.
24. Knaust, R., Urbig, T., Li, L., Taylor, W. & Hastings, J. W. (1998) *J. Phycol.* **34,** 167–172.
25. Tsai, H. J. & Wilson, J. E. (1996) *Arch. Biochem. Biophys.* **329,** 17–23.
26. Tsai, H. J. (1999) *Arch. Biochem. Biophys.* **369,** 149–156.
27. Poorman, R. A., Randolph, A., Kemp, R. G. & Heinrikson, R. L. (1984) *Nature* **309,** 467–469.
28. Wothe, D. D., Charbonneau, H. & Shapiro, B. M. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 5203–5207.
29. Maile, T., Kwoczynski, S., Katzenberger, R. J., Wassarman, D. A. & Sauer, F. (2004) *Science* **304,** 1010–1014.
30. Howard, M. B., Ekborg, N. A., Taylor, L. E., Weiner, R. M. & Hutcheson, S. W. (2004) *J. Bacteriol.* **186,** 1297–1303.
31. Tanaka, T., Fujiwara, S., Nishikori, S., Fukui, T., Takagi, M. & Imanaka, T. (1999) *Appl. Environ. Microbiol.* **65,** 5338–5344.
32. Zhu, Q. S., Deng, Y. P., Vanka, P., Brown, S. J., Muthukrishnan, S. & Kramer, K. J. (2004) *Bioinformatics* **20,** 161–169.
33. Marrone, L., McAllister, K. A. & Clarke, A. J. (2000) *Protein Eng.* **13,** 593–601.
34. Gilbert, H. J., Hazlewood, G. P., Laurie, J. I., Orpin, C. G. & Xue, G. P. (1992) *Mol. Microbiol.* **6,** 2065–2072.
35. Eberhardt, R. Y., Gilbert, H. J. & Hazlewood, G. P. (2000) *Microbiology* **146,** 1999–2008.
36. Ouwendijk, J., Peters, W. J. M., van de Vorstenbosch, R. A., Ginsel, L. A., Naim, H. Y. & Fransen, J. A. M. (1998) *J. Biol. Chem.* **273,** 6650–6655.
37. Soranzo, N., Sari Gorla, M., Mizzi, L., De Toma, G. & Frova, C. (2004) *Mol. Genet. Genomics* **271,** 511–521.
38. Guillard, R. & Ryther, J. (1962) *Can J. Microbiol.* **8,** 229–239.
39. Higgins, D., Thompson, J. & Gibson, T. (1994) *Nucleic Acids Res.* **22,** 4673–4680.