Syst. Biol. 49(1):87–100, 2000

Phylogenetic Relationship Among Horseshoe Crab Species: Effect of Substitution Models on Phylogenetic Analyses

XUHUA XIA
Molecular Laboratory, Department of Ecology & Biodiversity, University of Hong Kong, Pokfulam Road, Hong Kong; E-mail: xxia@hkusua.hku.hk

Abstract.—The horseshoe crabs, known as living fossils, have maintained their morphology almost unchanged for the past 150 million years. The little morphological differentiation among horseshoe crab lineages has resulted in substantial controversy concerning the phylogenetic relationship among the extant species of horseshoe crabs, especially among the three species in the Indo-Pacific region. Previous studies suggest that the three species constitute a phylogenetically unresolvable trichotomy, the result of a cladogenetic process leading to the formation of all three Indo-Pacific species in a short geological time. Data from two mitochondrial genes (for 16S ribosomal rRNA and cytochrome oxidase subunit I) and one nuclear gene (for coagulogen) in the four species of horseshoe crabs and outgroup species were used in a phylogenetic analysis with various substitution models. All three genes yield the same tree topology, with *Tachypleus-gigas* and *Carcinoscorpius-rotundicauda* grouped together as a monophyletic taxon. This topology is significantly better than all the alternatives when evaluated with the RELL (resampling estimated log-likelihood) method. [*Carcinoscorpius rotundicauda; Limulus polyphemus;* molecular phylogenetics; RELL; substitution model; *Tachypleus gigas*; *T. tridentatus*.]

There are four extant species of horseshoe crabs: *Limulus polyphemus*, distributed along the eastern coast of North America and the Gulf of Mexico, and *Tachypleus gigas*, *T. tridentatus,* and *Carcinoscorpius rotundicauda*, distributed in the Indo-Pacific region. These species of horseshoe crabs, in spite of their wide geographic distribution, exhibit few morphological differences, and are also very similar to a fossil specimen (*Mesolimulus walchi*) found in the Jurassic lithographic shale of Germany (Sekiguchi and Sugita, 1980). This implies that horseshoe crabs, now reputed as living fossils, have maintained their morphology for the past 150 million years or more (Fisher, 1984).

The morphological similarity among the horseshoe crabs has given rise to much difficulty in elucidating phylogenetic relationships among the horseshoe crab lineages. Many non-morphological characters have consequently been amassed and have resulted in at least a partial resolution of the horseshoe crab phylogeny. It is generally agreed that the Atlantic species, *L. polyphemus*, is a sister taxon to the three Indo-Pacific species. This is supported from serological evidence (Shuster, 1962), phylogenetic analyses of amino acid sequences of coagulogen and the fibrinopeptide-like peptide C (Shishikura et al., 1982; Srimal et al., 1985; Sugita and Shishikura, 1995), immunological comparisons of hemocyanins (Sugita, 1988), two-dimensional electrophoresis of general proteins (Miyazaki et al., 1987), interspecific hybridization experiments (Sekiguchi and Sugita, 1980), cladistic appraisals of morphological characters (Fisher, 1984), and mtDNA genes (Avise et al., 1994).

What remains unresolved is the phylogenetic relationship among the three Indo-Pacific species. *T. gigas* and *T. tridentatus* were grouped together on the basis of morphological traits (Fisher, 1984), but *C. rotundicauda* and *T. tridentatus* appear to be more closely related on the basis of amino acid sequence divergence of a fibrinopeptide-like protein (Shishikura et al., 1982) and coagulogen (Srimal et al., 1985), and on interspecific hybridization studies (Sekiguchi and Sugita, 1980). In a phylogenetic study of the four species of horseshoe crabs employing two-dimensional electrophoresis of skeletal and cardiac muscles, the similarity index is the greatest between *T. gigas* and *C. rotundicauda* for cardiac muscles, but between *T. tridentatus* and *T. gigas* for skeletal muscles (Miyazaki et al., 1987). Phylogenetic analyses based on two partial mitochondrial genes, 16S ribosomal RNA (rRNA) and cytochrome oxidase subunit I (*COI*), also yield conflicting topologies

(Avise et al., 1994). For example, the most-parsimonious (MP) tree from the 16S rRNA gene and one of the two MP trees from *COI* grouped *T. gigas* with *T. tridentatus*, but alternative topologies grouping *T. gigas* and *C. rotundicauda* together or grouping *T. tridentatus* and *C. rotundicauda* together are only two and four steps longer. In addition, the other MP tree from *COI* grouped *T. gigas* with *C. rotundicauda* as a monophyletic taxon. Such a confusing array of conflicting results has led to the conclusion that the three species constitute a phylogenetically unresolvable trichotomy, resulting from a cladogenetic process in which all three Indo-Pacific species formed within a short geological time (Avise et al., 1994).

The molecular phylogenetic study of horseshoe crabs by Avise et al. (1994) has several shortcomings when judged by today's standards of data analysis. First, the phylogenetic methods (parsimony and UPGMA) used in the paper are known to generate biased estimates of phylogenetic relationships (Kuhner and Felsenstein, 1994; Li, 1997; Takezaki and Nei, 1994). Second, there was no study of the substitution patterns of nucleotide and codon sites of the two mitochondrial genes (16S rRNA and *COI*). Without such a study, an investigator choosing phylogenetic methods is prone to errors. Third, the phylogenetic methods used in previous studies prevent researchers from evaluating quantitatively the relative statistical support of various phylogenetic hypotheses. Without such an evaluation, very little can be said about reliability of the best tree or the resolution power of the data set (Kishino and Hasegawa, 1989; Kishino et al., 1990; Penny and Hendy, 1986). Fourth, the third codon positions of the *COI* gene were discarded in the MP analysis, presumably because of substitution saturation. However, no effort is spent on checking the extent of substitution saturation. As I will show later, the third codon positions of the *COI* gene still retain useful phylogenetic information and contribute to a better phylogenetic resolution. Fifth, there was no outgroup species for the *COI*-based tree. Sixth, no nuclear gene was included in the study. All these suggest the necessity of a more up-to-date phylogenetic analysis of an expanded data set. In this paper I will show that phylogenetic relationships among the horseshoe crab species can be sufficiently resolved with available mitochondrial and nuclear genes.

Materials and Methods

Nucleotide sequences of two partial mitochondrial genes (16S rRNA and *COI*) and amino acid sequences of a nuclear gene (encoding coagulogen) were used in this study. For mitochondrial genes, a tick, *Ixodes hexagonus* (Black and Roehrdanz, 1998) and a brine shrimp, *Artemia franciscana,* were used as outgroups. Two complete 16S rRNA sequences from the brine shrimp have been reported (Black and Roehrdanz, 1998; Palmero et al., 1988), differing in two transitions and three indels, but these differences do not alter their phylogenetic relationship with the horseshoe crabs and the tick because four of the five changes are unique in the brine shrimp sequence. Only one 16S sequence (Palmero et al., 1988) is used in this study. For the coagulogen gene, only the amino acid sequences from the four species of horseshoe crabs were analyzed, because I could not find an outgroup species in GenBank. The species and genes with their GenBank Locus names are listed in Table 1.

Sequence alignment is straightforward for the *COI* gene and the coagulogen gene, but more complicated for the 16S gene. I used CLUSTALW (Thompson et al., 1994) to do preliminary alignment and then used the published secondary structure for termite 16S rRNA (Kambhampati et al., 1996) for further adjustment. Two segments are virtually impossible to align and are not included in the phylogenetic analysis. The aligned sequences of both genes can be found at http://web.hku.hk/~xxia/research/data/data.htm and at the *Systematic Biology* web site (www.utexas.edu/ftp/depts/systbiol/).

The tree to be resolved is (Outgroup1, Outgroup2 (*L. polyphemus*, *T. gigas*, *T. tridentatus*, C. *rotundicauda*)), with 15 possible topologies for the four species of horseshoe crabs. I have previously outlined evidence showing that the Atlantic species, *L. polyphemus*, is a sister taxon to the three Indo-Pacific species. Including *L. polyphe-*

TABLE 1. Species and DNA sequences (LOCUS name in GenBank) used for the phylogenetic analysis. There are two *COI* sequences from two different individuals for *Limulus polyphemus* and *Carcinoscorpius rotundicauda*.

OTUs	16S rRNA	*COI*	Coagulogen
Limulus polyphemus	LPU09397	LPU09391, LPU09392	758141
Tachypleus tridentatus	TTU09393	TTU09387	356167
Tachypleus gigas	TGU09394	TGU09388	352176
Carcinoscorpius rotundicauda	CRU09396	CRU09389, CRU09390	354133
Ixodes hexagonus	AF081828	AF081828	
Artemia franciscana	MIASRR16	MTAFDNA	

mus in the ingroup in this study, rather than using it to root the three Indo-Pacific species, is for a special purpose. The mitochondrial genes, which typically have a rapid evolutionary rate, are often plagued by substitution saturation. I included the *L. polyphemus* in the ingroup just to see if the two mitochondrial genes can still identify the root given the long time of divergence between the outgroup and the ingroup.

Maximum likelihood values for the 15 possible topologies were obtained by using the PAML package (Yang, 1997a) with the BASEML program for the 16S rRNA gene, and with the CODEML program for the *COI* gene. The nucleotide-based substitution models, such as the F84 model in the DNAML program in the PHYLIP package (Felsenstein, 1993), the HKY85 model (Hasegawa et al., 1985), the TN93 model (Tamura and Nei, 1993), and the general time-reversible model (REV; Yang, 1994a) were used with the BASEML program. A recently developed codon-based model (Yang et al., 1998) was used for the codon-based phylogenetic reconstruction. For the analysis involving the amino acid sequences of coagulogen, the empirical substitution matrix (Jones et al., 1992) was used. The amino acid-based model is designated as the JTT-F model.

The relative support for each of the 15 phylogenetic hypotheses was evaluated by using the RELL method (Kishino and Hasegawa, 1989; Kishino et al., 1990), which has been implemented in the RELL program in the PAML package (Yang, 1997a). The difference in maximum likelihood between the best tree and other trees, the standard error of these differences, and the estimated bootstrap probabilities for the 15 topologies were used to evaluate alternative phylogenetic hypotheses.

The validity of phylogenetic analyses, especially those involving the dating of speciation events, depends much on our understanding of substitution patterns (Lockhart et al., 1996; Yang, 1996a). To understand substitution patterns, we need to know how each site of nucleotide or amino acid sequences has changed during evolution. The empirical pattern of substitution is often obtained by doing all possible pairwise comparisons. Such comparisons, however, are not statistically independent and could introduce biases (Felsenstein, 1992; Nee et al., 1996; Xia et al., 1996). For example, if one species has recently experienced a large number of $A \rightarrow G$ transitions and few other substitutions, then all pairwise comparisons between this species and the other species will each contribute one data point with a large $A \leftrightarrow G$ transition bias.

One way to avoid such a problem of nonindependence is to reconstruct ancestral states of DNA sequences and estimate the number of substitutions between neighboring nodes on the phylogenetic tree (Gojobori et al., 1982; Tamura and Nei, 1993; Xia, 1998a; Xia et al., 1996; Xia and Li, 1998). Reconstruction of ancestral states using the maximum likelihood method (Yang et al., 1995) was performed by the BASEML program in the PAML package (Yang, 1997a).

As will be shown later, the two mitochondrial genes show substantial rate heterogeneity among sites, and substitution models with gamma-distributed rates were used to accommodate such rate heterogeneity. The results based on these models were compared with those from substitution models, assuming equal rates among sites.

This comparison is important because models with gamma-distributed rates, although providing better fit to observed substitution patterns (Yang, 1993, 1994b; Yang and Kumar, 1996), do not necessarily produce better phylogenies than those models that assume no rate heterogeneity among sites (Yang, 1997b).

Results and Discussion

Phylogenetic Analysis of the 16S Gene

The validity of phylogenetic analyses depends much on our understanding of the underlying substitution patterns that govern the evolution of nucleotide or amino acid sequences. A substitution model typically has two categories of parameters for describing a substitution pattern (Kumar et al., 1993; Li, 1997; Yang, 1997a): the rate ratio parameters (often symbolized by α, β, δ, γ, etc.), which specify the relative frequencies of different kinds of substitutions, and the frequency parameters (i.e., π_A, π_C, π_G, and π_T), which are nucleotide frequencies. Intuitively, we expect more realistic substitution models to offer better phylogenetic resolutions.

I examined the substitution pattern of the six species by reconstructing ancestral states based on the following partially specified tree: (*A. franciscana*(*I. hexagonus*(*L. polyphemus, T. tridentatus, T. gigas, C. rotundicauda*))). Pairwise comparisons were made between neighboring nodes along the branches, and two features are apparent. First, transitions occur more frequently than tranversions, which is easy to accommodate by using either the F84 or the HKY85 model. Second, the transition/transversion ratio is not constant along different branches. For example, the branch leading to *A. franciscana* experienced 26 transitions and 4 transversions, whereas the branch leading to *I. hexagonus* registered 50 transitions and 49 transversions. For this reason, I used the HKY85 model implemented in the BASEML program that allows estimation of one transition/transversion ratio (κ) for each branch (i.e., nhomo = 2 in the control file for BASEML).

The results (Table 2) support topology 4, in which *L. polyphemus* is the sister group of the three Indo-Pacific species and *T. gigas* and *C. rotundicauda* form a monophyletic group (Fig. 1 and Table 2). The RELL sup-

Table 2. Likelihood statistics for the 16S rRNA gene. The OTUs are represented by two letters that are the initials of the genus and the species names (e.g., Lp stands for *Limulus polyphemus*). l_i = likelihood value for topology i; l_{max} = the largest likelihood value, being equal to −1348.632; SE = standard error for the difference $l_i - l_{max}$; P(RELL) = estimated bootstrap probability expressed as percentage values. Results based on HKY85 model with separate transition/transversion ratios for different branches. The topologies are rooted by *Ixodes hexagonus* and *Artemia franciscana*.

Topology	l_i	$l_i - l_{max}$	SE	P(RELL)
1. ((LpTt)(CrTg))	−1399.38	−10.91	5.69	0.74
2. (Cr((LpTt)Tg))	−1399.85	−11.38	7.15	4.64
3. (Tg((LpTt)Cr))	−1402.09	−13.62	6.86	0.28
4. (Lp((CrTg)Tt))	−1388.47	0.00	0.00	86.24
5. (Tt((CrTg)Lp))	−1401.02	−12.56	5.14	0.00
6. (Lp((CrTt)Tg))	−1399.47	−11.00	7.63	5.76
7. ((CrTt)(LpTg))	−1410.98	−22.52	8.61	0.00
8. (Tg(Lp(CrTt)))	−1410.55	−22.08	8.90	0.02
9. (Cr((LpTg)Tt))	−1408.10	−19.63	9.17	0.34
10. (Tt((LpTg)Cr))	−1410.73	−22.26	8.42	0.00
11. (Lp(Cr(TgTt)))	−1401.54	−13.08	7.34	1.50
12. (Cr(Lp(TgTt)))	−1408.41	−19.95	8.86	0.24
13. ((TgTt)(LpCr))	−1409.13	−20.66	8.89	0.10
14. (Tg((LpCr)Tt))	−1410.93	−22.46	9.05	0.12
15. (Tt((LpCr)Tg))	−1411.60	−23.13	8.40	0.02

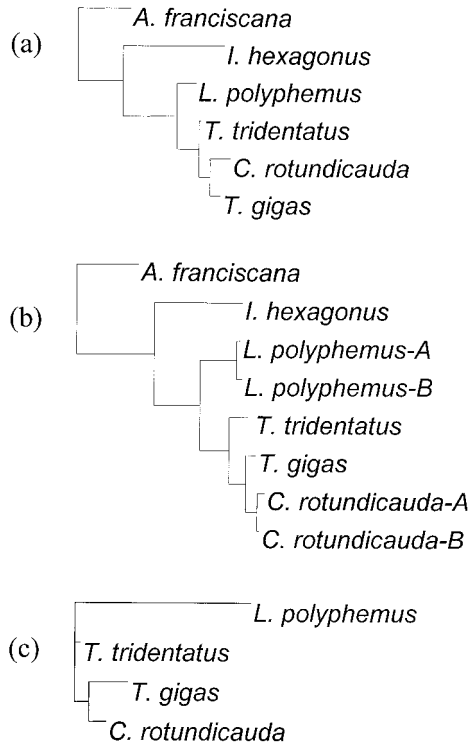

FIGURE 1. The topology with the largest likelihood values from the 15 possible topologies. (a) 16S rRNA; (b) *COI*; (c) coagulogen.

port for topology 4 (86.24%) is larger than that from the HKY85 model that assumes a single κ, for which the support is 80.52%. The log-likelihood of the former model is larger than the latter by 11.414, which is significant ($P < 0.01$), based on a likelihood ratio test with 9 df.

Because nucleotide frequencies differ substantially in the 16S gene among the six species (e.g., the frequency of T is 0.361 for *I. hexagonus* and 0.265 for *A. franciscana*), a more appropriate model should relax the assumption of stationarity. When I used such a non-homogeneous model, e.g., by specifying nhomo = 3 in the BASEML program, a larger likelihood value (−1348.632) and a stronger RELL support (91.06%) were obtained for topology 4 from the 16S gene.

Phylogenetic Analysis of the COI *and Coagulogen Gene*

The *COI* gene was analyzed by using a recently developed codon-based model (Yang et al., 1998). Codon-based models have several advantages over nucleotide-based models for protein-coding DNA sequences. Nucleotide-based models are inherently awkward in describing substitution patterns in protein-coding genes because the substitution rate at nucleotide sites of a protein-coding gene depends not only on whether the substitution is a transition or transversion and whether the site is located in a functionally important segment or not, but also on codon-specific properties, such as the codon position at which the site is, whether the substitution is synonymous or nonsynonymous, and how similar the two amino acids are to each other when the substitution is nonsynonymous. (Xia, 1998a; Yang, 1996b). In short, the nucleotide-based substitution model cannot handle the complexity of substitutions involving codons.

Several codon-based models have recently been proposed (Goldman and Yang, 1994; Muse and Gaut, 1994; Yang et al., 1998). The one by Muse and Gaut (1994) is restrictive in two ways. First, it does not have separate rate parameters for transitions and transversions. Second, it assumes that nonsynonymous substitutions occur equally likely. For example, two very different nonsynonymous substitutions, one involving AAT ↔ GAT (resulting in Asn ↔ Asp, with Grantham's [1974] distance = 23 between the two amino acids) and the other involving TGT ↔ CGT (resulting in Cys ↔ Arg, with Grantham's distance = 180), were assumed by the model to have the same substitution rate.

This assumption is known to be false. Amino acid substitutions occur more frequently between similar amino acids than between dissimilar ones (Clarke, 1970; Epstein, 1967; Grantham, 1974; Kimura, 1983: 152; Miyata et al., 1979; Sneath, 1966; Xia and Li, 1998; Zuckerkandl and Pauling, 1965). Similar amino acid replacements do occur more frequently in the *COI* gene in the eight OTUs in our study (Fig. 2). The genetic code of the invertebrate mitochondrion has 62 sense codons and 202 possible nonsynonymous codon pairs in which one codon can mutate into the other through a single nucleotide substitution (e.g., CTT ↔ ATT). If all nonsynonymous substitutions occur with equal frequency, then the amino acid pairs involved in nonsynonymous substitutions will have Grantham's distance ranging from 5 (Leu ↔ Ile) to 215 (Cys ↔

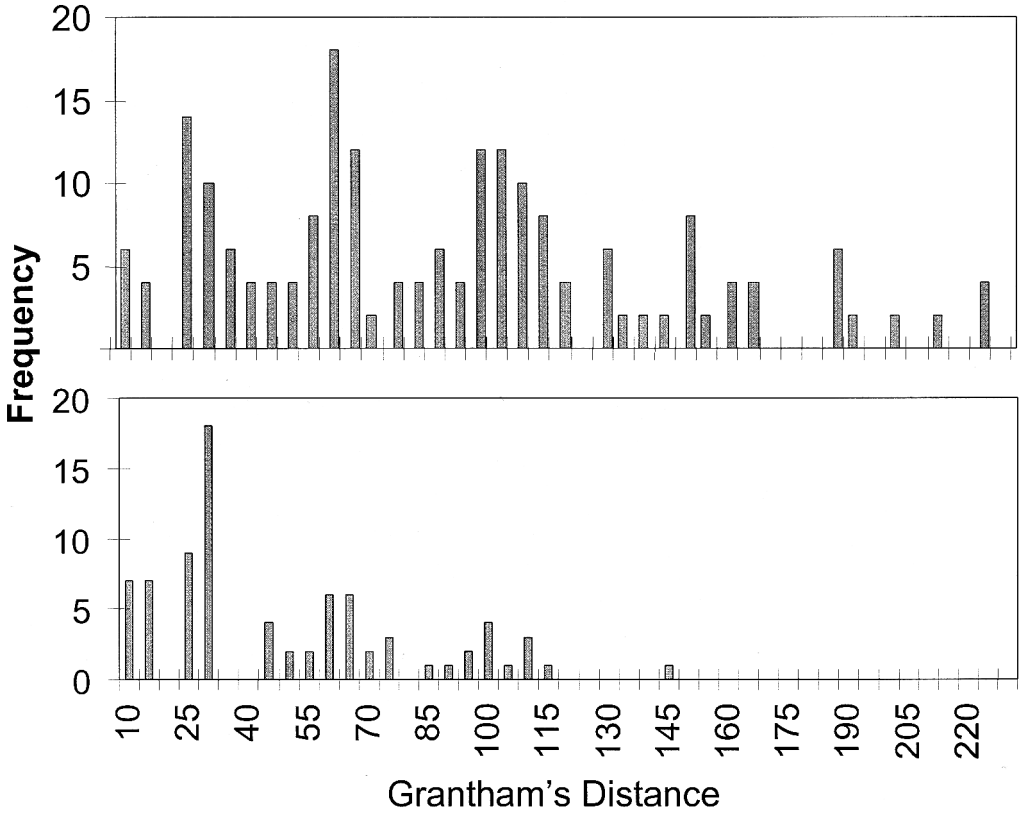

FIGURE 2. Frequency distribution of nonsynonymous substitutions along the range of Grantham's distances for the *COI* gene. Top panel: The expected frequency distribution when all nonsynonymous substitutions occur with equal frequency. Bottom panel: The observed distribution. Nonsynonymous substitutions involving very dissimilar amino acids (with a large Grantham's distance) are rare or absent.

Trp), as shown in the top panel of Figure 2. The observed nonsynonymous substitutions involved amino acids pairs with only small to moderate Grantham's distances (Fig. 2, bottom panel). The bottom distribution is significantly more right-skewed than the top, indicating a strong purifying selection that has eliminated amino acid replacements involving large Grantham's distances.

The model by Goldman and Yang (1994) accommodates potentially different substitution rates among different amino acid replacements by using dissimilarity measures between amino acids. This model has subsequently been improved (Yang et al., 1998), and it is this improved model, referred to as the YRH98 model, that is used in this study for analyzing the *COI* gene. The YRH98 model uses dissimilarity indices between amino acids to accommodate the rate heterogeneity among different kinds of amino acid substitutions.

There are currently two amino acid dissimilarity indices in use, Grantham's distance (Grantham, 1974) and Miyata's distance (Miyata et al., 1979). The latter has been shown to fit the observed codon substitution data better than Grantham's distance for mammalian mitochondrial DNA (Yang et al., 1998), but this is not true for the *COI* gene from my group of OTUs. In this study, the YRH98 model is used in conjunction with Grantham's distance.

The best-supported topology is the same as that from the 16S gene (Table 3 and Fig. 1). The likelihood values for the 15 topologies derived from the *COI* gene are highly correlated with those derived from the 16S gene (Fig. 3), indicating that the two genes contain similar phylogenetic information.

TABLE 3. Likelihood statistics for the *COI* gene. See Table 2 for abbreviations. *L. polyphemus* and *C. rotundicauda* are each represented by two sequences: i.e., Lp should have been written as (Lp1, Lp2), and Cr as (Cr1, Cr2). Based on the geometric + G74 model (Yang et al., 1998). The topologies are rooted by *Ixodes hexagonus* and *Artemia franciscana*.

Topology	l_i	$l_i - l_{max}$	SE	P(RELL)
1. ((LpTt)(CrTg))	−2556.24	−16.33	8.49	0.10
2. (Cr((LpTt)Tg))	−2566.63	−26.73	11.16	0.02
3. (Tg((LpTt)Cr))	−2566.63	−26.73	11.16	0.12
4. (Lp((CrTg)Tt))	−2539.90	0.00	0.00	90.06
5. (Tt((CrTg)Lp))	−2554.95	−15.05	8.98	3.36
6. (Lp((CrTt)Tg))	−2550.26	−10.35	6.95	2.10
7. ((CrTt)(LpTg))	−2571.33	−31.43	11.87	0.00
8. (Tg(Lp(CrTt)))	−2571.33	−31.43	11.87	0.00
9. (Cr((LpTg)Tt))	−2571.82	−31.91	11.78	0.00
10. (Tt((LpTg)Cr))	−2565.72	−25.82	11.55	0.08
11. (Lp(Cr(TgTt)))	−2550.09	−10.19	7.07	4.00
12. (Cr(Lp(TgTt)))	−2571.44	−31.54	11.93	0.00
13. ((TgTt)(LpCr))	−2571.44	−31.54	11.93	0.00
14. (Tg((LpCr)Tt))	−2571.82	−31.91	11.78	0.00
15. (Tt((LpCr)Tg))	−2565.72	−25.81	11.55	0.16

One shortcoming of the existing codon-based models is that they disregard the factors constraining synonymous codon usage in protein genes. Synonymous codon usage and consequently the synonymous substitution rate may be affected by the tRNA availability (Berg and Kurland, 1997; Gouy and Gautier, 1982; Ikemura, 1992; Sorensen et al., 1989; Xia, 1998b), the level of gene expression (Bennetzen and Hall, 1982; Berg and Martelius, 1995; Sharp et al., 1988; Sharp and Devine, 1989), ribonucleotide availability (Xia, 1996), and the conservativeness of the protein segment (Akashi, 1994). These studies suggest that the pattern of synonymous substitutions in protein genes may differ between species, between genes, and even between different segments of the same gene. All these would contribute to the rate heterogeneity among

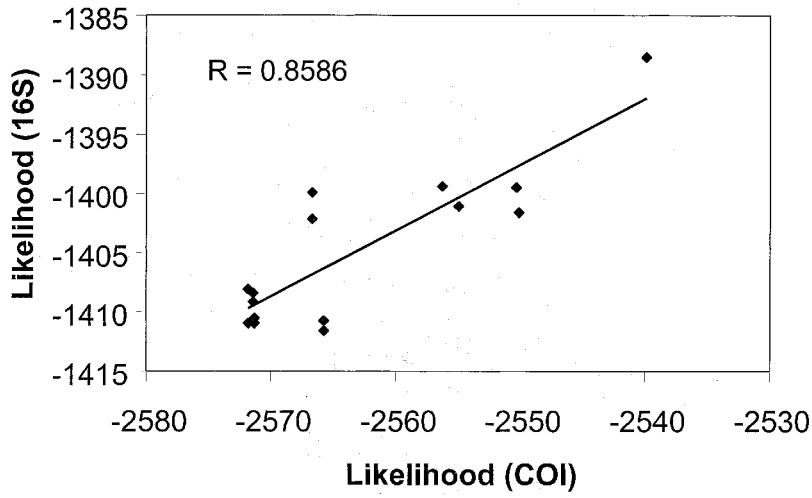

FIGURE 3. The two mitochondrial genes (*COI* and 16S) contain similar phylogenetic information, as shown by the high correlation in likelihood values for the 15 possible topologies between the two genes.

TABLE 4. Likelihood statistics for the nuclear coagulogen gene for four horseshoe crab species. Derived from the amino acid-based substitution model with JTT-F empirical substitution matrix.

Topology	l_i	$l_i - l_{max}$	SE	P(RELL)
((Tt,Lp)Cr,Tg)	−898.31	0	0	76.08
((Tt,Cr)Lp,Tg)	−900.76	−2.45	2.59	4.32
((Tt,Tg)Lp,Cr)	−900.56	−2.25	2.84	19.6

synonymous substitutions. Existing codon-based models all use a single rate parameter for synonymous substitutions.

The coagulogen gene also favors the grouping of *T. gigas* and *C. rotundicauda* (Fig. 1 and Table 3), although none of the three phylogenetic hypotheses can be rejected statistically. Note that there is no outgroup to root the four horseshoe crab species. However, the root should be somewhere along the branch leading to *L. polyphemus* because otherwise we would have to assume an extraordinarily fast evolutionary rate in the *L. polyphemus* lineage.

The JTT-F model (Yang et al., 1998) was used in the maximum likelihood reconstruction. The main reason for using the empirical JTT-F model is that it has fewer parameters to estimate than a mechanical model, which makes it attractive given the limited length of coagulogen sequences. One problem with using the empirical JTT-F model is that the data are insufficient for deriving an amino acid substitution matrix exclusively from protein-coding genes of invertebrate mitochondrial DNA; consequently, the substitution matrix is based on a large-scale compilation of heterogeneous genes that may not necessarily reflect the substitution patterns of the *COI* gene in this study. Alternatively, one could use a mechanical substitution model that does not use an empirical substitution matrix, but instead uses an amino acid dissimilarity matrix (e.g., Grantham's distance or Miyata's distance) to accommodate rate heterogeneity among different kinds of amino acid substitutions. However, amino acid dissimilarity matrices are not really derived entirely independently of the empirical data. For example, if a matrix derived entirely on the basis of certain amino acid properties is then found to fit the empirical substitution data poorly, then the matrix will simply be discarded. In short, it is the empirical data, few of which are from arthropod mitochondria, that decide whether a distance matrix makes sense or not.

A maximum parsimony reconstruction revealed 83 amino acid changes for topologies 1 and 2 in Table 4, 84 changes for topology 3. Thus, although the coagulogen tree is consistent with the topology derived from mitochondrial genes, it adds very little to resolving the phylogenetic relationship among the three Indo-Pacific species.

Combined Analysis of the COI *and the 16S Gene*

An alternative way of evaluating the relative support of the 15 possible topologies is to apply significance tests to the differences in likelihood between a hypothesized topology and the topology with the largest maximum likelihood value (Kishino and Hasegawa, 1989). The two sets of results, one from the 16S rRNA gene (Table 2) and the other from the *COI* gene (Table 3), can be combined for the significance test. Let $L_{16S.i}$ and $L_{COI.i}$ be the likelihood values obtained from the 16S rRNA gene and the *COI* gene, respectively, for topology i (i = 1, 2, . . ., 15), and $SE_{16S.i}$ and $SE_{COI.i}$ be the corresponding standard errors. We now compute

$$L_{sum.i} = L_{16S.i} + L_{COI.i} \quad (1)$$

$$L_{max} = \max(L_{sum.i}) \quad (2)$$

$$L_{diff.i} = L_{sum.i} - L_{max} \quad (3)$$

$$SE_{L_{diff.i}} = \sqrt{SE^2_{16S.i} + SE^2_{COI.i}} \quad (4)$$

$$z_i = \left| \frac{L_{diff.i}}{SE_{L_{diff.i}}} \right| \quad (5)$$

where z_i can be used to perform a standard z-test. If $z_i > 1.96$, then the topology i is rejected at the 0.05 significance level. This test has been used in previous studies (Kishino

TABLE 5. Combining analysis of results from separate analyses of the *COI* and the 16S sequences for testing the hypothesis of $l_{diff.i} = 0$. See equations (1–5) for abbreviations. The results are derived from data in Tables 2 and 3. The topologies are rooted by *Ixodes hexagonus* and *Artemia franciscana*.

Topology	$l_{sum.i}$	$l_{diff.i}$	SE	z_i	$P(z_i = 0)$
1. ((LpTt)(CrTg))	−3955.61	−27.24	10.27	2.654	0.0080
2. (Cr((LpTt)Tg))	−3966.48	−38.11	13.16	2.896	0.0038
3. (Tg((LpTt)Cr)	−3968.72	−40.34	13.07	3.087	0.0020
4. (Lp((CrTg)Tt))	−3928.37	0.00	0.00		
5. (Tt((CrTg)Lp))	−3955.97	−27.60	10.51	2.627	0.0086
6. (Lp((CrTt)Tg))	−3949.73	−21.36	9.32	2.291	0.0220
7. ((CrTt)(LpTg))	−3982.32	−53.95	14.76	3.655	0.0003
8. (Tg(Lp(CrTt)))	−3981.89	−53.51	14.54	3.680	0.0002
9. (Cr((LpTg)Tt))	−3979.92	−51.54	14.75	3.495	0.0005
10. (Tt((LpTg)Cr))	−3976.45	−48.08	14.14	3.399	0.0007
11. (Lp(Cr(TgTt)))	−3951.63	−23.26	9.47	2.457	0.0140
12. (Cr(Lp(TgTt)))	−3979.86	−51.48	14.93	3.449	0.0006
13. ((TgTt)(LpCr))	−3980.57	−52.20	14.79	3.530	0.0004
14. (Tg((LpCr)Tt))	−3982.75	−54.37	14.46	3.759	0.0002
15. (Tt((LpCr)Tg))	−3977.32	−48.95	13.91	3.518	0.0004

and Hasegawa, 1989), and I wrote it out explicitly just to show that it is really appropriate only for a single comparison, not for multiple comparisons. For this reason, the interpretation of the test result is heuristic.

The test above is similar to the test implemented in the DNAPARS program in PHYLIP. In the latter, the differences in the number of character changes (steps) between the MP tree and alternative trees, and the SE of such differences, are calculated. A difference is declared significant at the 0.05 level if it is >1.96 SE.

Topology 4 is the best of the 15 possible topologies (Tables 2 and 3). Of the remaining 14 possible topologies, all can be rejected at the 0.05 significance level (Table 5). The same conclusion can be made if we use alternative substitution models, such as HKY85, TN93, or REV models rather than the HKY85 model for the 16S gene. This consistent and overwhelming support for just 1 of the 15 possible topologies is quite remarkable because we typically encounter alternative topologies that receive similar amount of support, leaving us undecided and depressed.

A similar test based on maximum parsimony statistics produce only slightly different results. For the 16S gene, the number of steps for topology 4 (Fig. 1) is 402 steps, which is the same as that for the alternative topology that groups *T. gigas* and *T. tridentatus* as a monophyletic taxon. However, the former is more strongly supported by the *COI* gene, with 394 steps, than the latter, with 407 steps. A combined analysis based on maximum parsimony statistics rather than maximum likelihood statistics can reject all alternative topologies at the 0.05 level except two: topologies 6, with $P = 0.0581$, and topology 11, with $P = 0.0790$. These two topologies differ from topology 4 in the relative position of the three Indo-Pacific species, with topology 6 grouping *C. rotundicauda* and *T. tridentatus* together, and topology 11 grouping *T. tridentatus* and *T. gigas* together.

Phylogenetic Information in the Third Codon Positions

Avise et al. (1994) did not use a codon-based model, and excluded the third codon position of the *COI* gene from phylogenetic analysis. This might be wise when the third codon positions have experienced substitution saturation. In such cases, the phylogenetic tree reconstructed from only the third codon positions will often be rather different from that reconstructed from the first and second codon positions.

Phylogenetic reconstruction using only the first and second codon positions of the *COI* gene and the F84 model supports the *COI* topology in Figure 1, for which the RELL support is 80.32%. This topology is also supported when I apply various phylogenetic methods to the third codon positions only. The distance methods (both neighbor-joining and Fitch-Margoliash), using the proportion of different nucleotides between the two sequences as the distance, produced the same *COI* topology as shown in Figure 1. The simple proportion is used because correction for multiple hits is unreliable for very diverged taxa. In addition, the maximum parsimony method also generates the same *COI* topology based on the third codon positions.

When the maximum likelihood method with the F84 model is applied to the third codon positions of the *COI* gene, the *COI* topology in Figure 1 is obtained when the transition/transversion ratio (κ) is set to 4.91, with the maximum likelihood value of −1132.34. However, when κ is set to 5.19, a tree with a slightly larger maximum likelihood value (−1132.19) is obtained which groups *T. tridentatus* and *T. gigas* as a monophyletic group. Given that the partial *COI* gene has only 194 codons and consequently only 194 third codon positions, such a minor difference can be attributed to sampling error. In short, the phylogenetic tree from the third codon positions of the *COI* gene is concordant with that from the first and second codon positions, and the wisdom of discarding the third codon positions without checking the presence of phylogenetic information (Avise et al., 1994) is therefore questionable.

Rate Heterogeneity over Sites and Gamma Models

One of the most perplexing problems in phylogenetics is the rate heterogeneity among sites. All popular phylogenetic methods appear to fail when different sites of the gene differ in substitution rate (Gaut and Lewis, 1995; Huelsenbeck, 1995; Kuhner and Felsenstein, 1994; Lockhart et al., 1996; Tateno et al., 1994; Yang, 1995, 1997b).

There are two sources of rate heterogeneity over sites along molecular sequences. The first, caused by differential selection pressure exerted on different sites, is well exemplified by different codon positions in a protein-coding gene (Xia, 1998a; Yang, 1996b). The distribution of substitutions over nucleotide sites for the *COI* gene is significantly different from the expectation based on the Poisson distribution (Table 6). This rate heterogeneity among codon sites can be accommodated with a codon-based model. For example, if we use a nonoverlapping window of three neighboring sites (a codon) as a sampling unit, then the distribution of substitutions follows closely the Poisson distribution ($X^2 = 6.82$, df = 5, $P = 0.2345$).

The second source of rate heterogeneity is caused by differential selection pressure exerted on different gene segments. Different parts of a gene differ in functional importance (Irwin et al., 1991; Xia, 1998a), and the functionally important sites tend to be more conservative than those functionally unimportant sites. Such rate heterogeneity among gene segments would lead to both rate heterogeneity among sites and autocorrelation in substitution rates between neighboring sites. For example, the distribution of substitutions over sites for the 16S sequences follows the Poisson distribution quite closely, based on a goodness-of-fit test ($X^2 = 2.68$, df = 3, $P = 0.4436$). However, if a nonoverlapping window of two or more neighboring sites is used as a sampling unit, then the distribution deviates signifi-

TABLE 6. Fitting a Poisson distribution to observed nucleotide substitutions for the *COI* gene. The sequences are 576 nucleotides long after deleting unresolved codons. x, – number of substitutions at a nucleotide site; N_{obs}, number of sites with x substitutions (e.g., there are 328 nucleotide sites that have experienced no substitution); N_{exp}, expected number of sites experiencing x substitutions. The Poisson parameter $\lambda = 0.68$, $P = 0.0000$, df = 3.

x	N_{Obs}	N_{exp}	X^2
0	328	295.74	3.52
1	144	200.21	15.78
2	83	67.77	3.42
3	24	15.29	4.96
≧4	3	2.98	0.00
Sum	582	582	27.68

TABLE 7. Fitting a Poisson distribution to nucleotide substitutions for the 16S gene, with a nonoverlapping window of three neighboring sites. The sequence length is 363 after pairwise deletion of gaps. Symbols are the same as in Table 6. The Poisson parameter $\lambda = 1.99$, $P = 0.0002$, df = 5. The distribution follows closely a negative binomial distribution, with $X^2 = 2.3$, df = 5 (the last group was broken into two so that there are seven groups), $P = 0.8069$. The negative binomial coefficient $k = 2.8321$.

x	N_{Obs}	N_{exp}	X^2
0	27	16.51	6.66
1	33	32.89	0.00
2	22	32.75	3.53
3	15	21.74	2.09
4	10	10.83	.06
5	8	4.31	3.15
≧6	6	1.97	8.26
Sum	121	121	23.76

cantly from the Poisson distribution and follows closely the negative binomial distribution (Table 7).

It is not surprising that substitution models with gamma-distributed rates fit the data much better than those that assume equal rates among sites (Table 8). However, it is surprising that the former performed relatively poorly compared with the latter in recovering the "true" tree (i.e., the topologies in Fig. 1), the RELL support for the "true" tree being much smaller for the former than for the latter (Table 8). Thus, the incorporation of gamma-distributed rates in the substitution models does not seem to help in phylogenetic analysis in these two invertebrate mitochondrial genes.

The only case in which the gamma model did better than the non-gamma model involves amino acid sequences of the coagulogen gene (Table 8). However, there is no strong indication of rate heterogeneity among sites for this gene because the distribution of amino acid substitutions follows the Poisson distribution quite closely ($X^2 = 4.4$, df = 2, $P = 0.111$).

Our result suggests that a gamma model that fits the observed substitution pattern better than a non-gamma model does not necessarily mean that the former will provide better phylogenetic resolution than the latter. This seemingly paradoxical result is in fact not difficult to understand. For a simple illustration, suppose the following four sequences are evolving according to the JC69 model (Jukes and Cantor, 1969). Each sequence consists of one slow-evolving segment (first four nucleotides) and one fast-evolving segment (last four nucleotides):

Seq1: ACGTACGT

Seq2: ACGTCGTA

Seq3: ACGTGTAC

Seq4: ACGTTACG

A JC69 + gamma model would provide a significantly better fit to the data than its non-gamma counterpart, but the better fit has nothing to do with phylogenetic resolution, although it is expected to improve branch length estimates. If the first four sites of one sequence become fast-evolving, while the first four sites of the other three

TABLE 8. Comparison of RELL support for the "true" tree (Fig. 1) between models with gamma-distributed rates over sites (Gamma) and models assuming equal rates over sites (No Gamma).

		RELL Support for Tree 4		*Likelihood value for Tree 4*	
Gene	Model	No Gamma	Gamma	No Gamma	Gamma
COI	geo+G74	90.06	84.56	−2539.91	−2234.41
16S	JC69	78.38	73.68	−1467.53	−1456.35
	F84	80.52	49.32	−1399.88	−1378.16
	HKY85	80.44	48.80	−1398.44	−1377.02
	TN93	80.38	47.76	−1397.79	−1376.14
	REV	72.14	40.14	−1387.32	−1365.67
Coagulogen	JTT-F	70.3	76.08	−906.87	−898.31

remain slow-evolving, then this situation violates the existing gamma models, which one would then expect to yield misleading phylogenetic results.

Phylogeny and the Horseshoe Crab Evolution

Most horseshoe crab fossils have been found away from the sea, and it is generally believed that the ancestors of the horseshoe crab inhabited brackish or freshwater environments (Chatterji, 1994:12–13). Three extant horseshoe crab species (*L. polyphemus*, *T. tridentatus*, and *T. gigas*) are highly marine, but *C. rotundicauda* was reported to occur in the river Hooghly as far as 90 km away from the sea (Chatterji, 1994:26). It is not known whether the endurance of freshwater environment in *C. rotundicauda* is ancestral or secondarily derived. The present analysis, showing that *C. rotundicauda* represents a most recently derived lineage, suggests that endurance is a derived character. These results also suggest that the common ancestor of modern horseshoe crabs was a marine inhabitant, which implies that the ancient forms of horseshoe crabs represented by fossils found in ancient freshwater environments are failed evolutionary offshoots.

C. rotundicauda and *T. gigas* are similar in size, and both much smaller than the other two horseshoe crab species. Whereas the former is found in brackish and occasionally in freshwater environments, and breeds in muddy shores, the latter is a purely marine form and breeds on sandy beaches (Chatterji, 1994:25–35). This difference in habitat preference might have contributed to obscuring the phylogenetic relationship between the two species, which are grouped as a monophyletic taxon in my phylogenetic analysis of molecular sequences.

In summary, a combination of an increased amount of data and a judicious choice of phylogenetic methods have resulted in a better-resolved phylogenetic relationships among the extant species of horseshoe crabs. The phylogenetic information in the two mitochondrial genes is sufficient to reject 14 of 15 possible topologies for the four extant species of horseshoe crabs. Both mitochondrial and nuclear genes support the topology that groups *Carcinoscorpius rotundicauda* and *Tachypleus gigas* together as a monophyletic taxon. My results from comparisons among substitution models suggest that the current methods of handling the problem of rate heterogeneity over sites do not help in resolving conflicting phylogenies and should be used with caution.

ACKNOWLEDGMENT

This project was supported by a RGC grant from Hong Kong Government (HKU 7259/97M) and CRCG grants from University of Hong Kong (335/023/0022 and 10202258/27662/25400/323/01) to X.-X. I thank J. C. Avise and two anonymous reviewers for their comments, which resulted in extensive reanalysis and revision. C. Simon helped me with the alignment of the 16S gene by incorporating information on secondary structures. I am also grateful to B. Morton for his review.

REFERENCES

AKASHI, H. 1994. Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. Genetics 136:927–935.

AVISE, J. C., W. S. NELSON, AND H. SUGITA. 1994. A speciational history of 'living fossils': Molecular evolutionary patterns in horseshoe crabs. Evolution 48:1986–2001.

BENNETZEN, J. L., AND B. D. HALL. 1982. Codon selection in yeast. J. Biol. Chem. 257:3026–3031.

BERG, O. G., AND C. G. KURLAND. 1997. Growth rate-optimized tRNA abundance and codon usage. J. Mol. Biol. 270:544–550.

BERG, O. G., AND M. MARTELIUS. 1995. Synonymous substitution-rate constants in *Escherichia coli* and *Salmonella typhimurium* and their relationship to gene expression and selection pressure. J. Mol. Evol. 41: 449–456.

BLACK, W. C. I., AND R. L. ROEHRDANZ. 1998. Mitochondrial gene order is not conserved in arthropods: Prostriate and metastriate tick mitochondrial genomes. Mol. Biol. Evol. 15:1772–1785.

CHATTERJI, A. 1994. The horseshoe crab—a living fossil. Project Swarajya Publication, Cuttack, Orissa, India.

CLARKE, B. 1970. Selective constraints on amino-acid substitutions during the evolution of proteins. Nature 228:159–160.

EPSTEIN, C. J. 1967. Non-randomness of amino-acid changes in the evolution of homologous proteins. Nature 215:355–359.

FELSENSTEIN, J. 1992. Estimating effective population size from samples of sequences: Inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. 59:139–147.

FELSENSTEIN, J. 1993. PHYLIP: Phylogeny inference package, version 3.5. Department of Genetics, Univ. Washington, Seattle.

FISHER, D. C. 1984. The Xiphosurida: Archetypes of bradytely? Pages 196–213 *in* Living fossils. (N. Eldredge, and S. M. Stanley, eds.). Springer, New York.

GAUT, B. S., AND P. O. LEWIS. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol. Biol. Evol. 12:152–162.

GOJOBORI, T., W. H. LI, AND D. GRAUR. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. 18:360–369.

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

GOUY, M., AND C. GAUTIER. 1982. Codon usage in bacteria: Correlation with gene expressivity. Nucleic Acids Res. 10: 7055–7064.

GRANTHAM, R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862–864.

HASEGAWA, M., H. KISHINO, AND T. A. YANO. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22: 160–174.

HUELSENBECK, J. P. 1995. The performance of phylogenetic methods in simulation. Syst. Biol. 44:17–48.

IKEMURA, T. 1992. Correlation between codon usage and tRNA content in microorganisms. Pages 87–111 *in* Transfer RNA in protein synthesis (D. L. Hatfield, B. Lee, and J. Pirtle, eds.). CRC Press, Boca Raton, Florida.

IRWIN, D. M., T. D. KOCHER, AND A. C. WILSON. 1991. Evolution of the cytochrome *b* gene of mammals. J. Mol. Evol. 32:128–144.

JONES, D. T., W. R. TAYLOR, AND J. M. THORNTON. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8: 275–282.

JUKES, T. H., AND C. R. CANTOR. 1969. Evolution of protein molecules. Pages 21–123 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

KAMBHAMPATI, S., K. M. KJER, AND B. L. THORNE. 1996. Phylogenetic relationship among termite families based on DNA sequence of mitochondrial 16S ribosomal RNA gene. Insect Mol. Biol. 5:229–238.

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambridge Univ. Press, Cambridge.

KISHINO, H., AND M. HASEGAWA. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29:170–179.

KISHINO, H., T. MIYATA, AND M. HASEGAWA. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. 31:151–160.

KUHNER, M. K., AND J. FELSENSTEIN. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11:459–468.

KUMAR, S., K. TAMURA, AND M. NEI. 1993. MEGA: Molecular evolutionary genetics analysis, version 1.0. The Pennsylvania State Univ., University Park.

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

LOCKHART, P. J., A. W. LARKUM, M. STEEL, P. J. WADDELL, AND D. PENNY. 1996. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. Proc. Natl. Acad. Sci. USA 93:1930–1934.

MIYATA, T., S. MIYAZAWA, AND T. YASUNAGA. 1979. Two types of amino acid substitution in protein evolution. J. Mol. Evol. 12:219–236.

MIYAZAKI, J., K. SEKIGUCHI, AND T. HIRABAYASHI. 1987. Application of an improved method of two-dimensional electrophoresis to the systematic study of horseshoe crabs. Biol. Bull. 172:212–224.

MUSE, S. V., AND B. S. GAUT. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. 11:715–724.

NEE, S., E. C. HOLMES, A. RAMBAUT, AND P. H. HARVEY. 1996. Inferring population history from molecular phylogenies. Pages 66–80 *in* New uses for new phylogenies (P. H. Harvey, A. J. L. Brown, J. Maynard Smith, and S. Nee, eds.). Oxford Univ. Press, Oxford.

PALMERO, I., J. RENART, AND L. SASTRE. 1988. Isolation of cDNA clones coding for mitochondrial 16S ribosomal RNA from the crustacean *Artemia*. Gene 68:239–248.

PENNY, D., AND M. HENDY. 1986. Estimating the reliability of evolutionary trees. Mol. Biol. Evol. 3:403–417.

SEKIGUCHI, K., AND H. SUGITA. 1980. Systematics and hybridization in the four living species of horseshoe crabs. Evolution 34:712–718.

SHARP, P. M., E. COWE, D. G. HIGGINS, D. C. SHIELDS, K. H. WOLFE, AND F. WRITE. 1988. Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens;* a review of the considerable within-species diversity. Nucleic Acids Res. 16:8207–8211.

SHARP, P. M., AND K. M. DEVINE. 1989. Codon usage and gene expression level in *Dictyostelium discoideum:* Highly expressed genes do "prefer" optimal codons. Nucleic Acids Res. 17:5029–5038.

SHISHIKURA, F., S. NAKAMURA, K. TAKAHASHI, AND K. SEKIGUCHI. 1982. Horseshoe crab phylogeny based on amino acid sequences of the fibrino-peptide-like peptide C. J. Ex. Zool. 223:89–91.

SHUSTER, C. N., JR. 1962. Serological correspondence among horseshoe "crabs" (Limulidae). Zoologica 47:1–8.

SNEATH, P. H. A. 1966. Relations between chemical structure and biological activity. J. Theor. Biol. 12:157–195.

SORENSEN, M. A., C. G. KURLAND, AND S. PEDERSEN. 1989. Codon usage determines translation rate in *Escherichia coli.* J. Mol. Biol. 207:365–377.

SRIMAL, S., T. MIYATA, S. KAWABATA, T. MIYATA, AND S. IWANAGA. 1985. The complete amino acid sequence of coagulogen isolated from southeast Asian horseshoe crab, *Carcinoscorpius rotundicauda.* J. Biochem. 98:305–318.

SUGITA, H. 1988. Immunological comparisons of hemocyanins and their phylogenetic implications. Pages 315–334 *in* Biology of horseshoe crabs (K. Sekiguchi, ed.). Science House, Tokyo.

SUGITA, H., AND F. SHISHIKURA. 1995. A case of orthologous sequences of hemocyanin subunits for an evolutionary study of horseshoe crabs: Amino acid sequence comparison of immunologically identical subunits of *Carcinoscorpius rotundicauda* and *Tachypleus tridentatus*. Zool. Sci. Tokyo 12:661–667.

TAKEZAKI, N., AND M. NEI. 1994. Inconsistency of the maximum parsimony method when the rate of nucleotide substitution is constant. J. Mol. Evol. 39:210–218.

TAMURA, K., AND M. NEI. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10:512–526.

TATENO, Y., N. TAKEZAKI, AND M. NEI. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varied with site. Mol. Biol. Evol. 11:261–277.

THOMPSON, J. D., D. G. HIGGINS, AND T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

XIA, X. 1996. Maximizing transcription efficiency causes codon usage bias. Genetics 144:1309–1320.

XIA, X. 1998a. The rate heterogeneity of nonsynonymous substitutions in mammalian mitochondrial genes. Mol. Biol. Evol. 15:336–344.

XIA, X. 1998b. How optimized is the translational machinery in *E. coli, S. typhimurium,* and *S. cerevisiae*? Genetics 149:37–44.

XIA, X., M. S. HAFNER, AND P. D. SUDMAN. 1996 On transition bias in mitochondrial genes of pocket gophers. J. Mol. Evol. 43:32–40.

XIA, X., AND W.-H. LI. 1998. What amino acid properties affect protein evolution? J. Mol. Evol. 47:557–564.

YANG, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.

YANG, Z. 1994a. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39:105–111.

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

YANG, Z. 1995. Evaluation of several methods for estimating phylogenetic trees when substitution rates differ over nucleotide sites. J. Mol. Evol. 40:689–697.

YANG, Z. 1996a. Among-site rate variation and its impact on phylogenetic analysis. Trends Ecol. Evol. 11:367–372.

YANG, Z. 1996b. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42:587–596.

YANG, Z. 1997a. PAML: A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555–556.

YANG, Z. 1997b. How often do wrong models produce better phylogenies? Mol. Biol. Evol. 14:105–108.

YANG, Z., AND S. KUMAR. 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol. Biol. Evol. 13:650–659.

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., R. NIELSEN, AND M. HASEGAWA. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. 15:1600–1611.

ZUCKERKANDL, E., AND L. PAULING. 1965. Evolutionary divergence and convergence in proteins. Pages 97–166 *in* Evolving genes and proteins (V. Bryson, and H. J. Vogel, eds.). Academic Press, New York.

Received 10 July 1998; accepted 4 November 1999
Associate Editor: C. Simon