Heterotachy, an Important Process of Protein Evolution

P. Lopez, D. Casane, and H. Philippe

Phylogénie, Bioinformatique et Génome, CNRS, Université Pierre et Marie Curie, Paris

Because of functional constraints, substitution rates vary among the positions of a protein but are usually assumed to be constant at a given site during evolution. The distribution of the rates across the sequence positions generally fits a Γ distribution. Models of sequence evolution were accordingly designed and led to improved phylogenetic reconstruction. However, it has been convincingly demonstrated that the evolutionary rate of a given position is not always constant throughout time. We called such within-site rate variations heterotachy (for ‘‘different speed’’ in Greek). Yet, heterotachy was found among homologous sequences of distantly related organisms, often with different functions. In such cases, the functional constraints are likely different, which would explain the different distribution of variable sites. To evaluate the importance of heterotachy, we focused on amino acid sequences of mitochondrial cytochrome *b,* for which the function is likely the same in all vertebrates. Using 2,038 sequences, we demonstrate that 95% of the variable positions are heterotachous, i.e., underwent dramatic variations of substitution rate among vertebrate lineages. Heterotachy even occurs at small evolutionary scale, and in these cases it is very unlikely to be related to functional changes. Since a large number of sequences are required to efficiently detect heterotachy, the extent of this phenomenon could not be estimated for all proteins yet. It could be as large as for cytochrome *b,* since this protein is not a peculiar case. The observations made here open several new avenues of research, such as the understanding of the evolution of functional constraints or the improvement of phylogenetic reconstruction methods.

Introduction

An accurate description of how biological sequences evolve is a fundamental prerequisite for comparative analyses like phylogenetics or genomics (Yang 1996). In the early days of sequence comparisons, the frequencies of change of one amino acid to another were observed to be highly heterogeneous (Dayhoff 1979) and were empirically estimated. The various resulting mutation matrices (e.g., PAM250 or BLOSUM62) are now heavily used in similarity searches of sequence databases (BLAST or FASTA) (Jones, Taylor, and Thornton 1992, 1994; Adachi and Hasegawa 1996; Adachi et al. 2000). Moreover, according to our knowledge of functional constraints acting on proteins, the number of substitutions per position was also shown to be unevenly distributed along the sequence. Some positions were found more prone to accepting mutations than others, and the distribution of the rates of substitutions could generally be fitted on a gamma (Γ) distribution (Uzzell and Corbin 1971). This latter observation was neglected until the 1990s, when models of sequence evolution were designed to account for among-site rate variation. Similarity searches greatly benefited from this better description, with the development of PSI-BLAST for example (Altschul and Koonin 1998). In the phylogenetic field, the variability of the substitution rates along the sequence was handled by rate-across-sites (RAS) models through a Γ distribution (Rzhetsky and Nei 1994; Strimmer and von Haeseler 1996; Yang 1997). This improved the reconstruction of the phylogenies (Yang 1996; Sullivan and Swofford 1997), especially when very divergent taxa were included and long-branch attraction artifacts were prevalent (Huelsenbeck 1998; Philippe and Germot 2000; Van de Peer, Ben Ali, and Meyer 2000).

RAS models postulate that the evolutionary rate of a position is constant throughout time (i.e., in all lineages), even if this rate can vary between positions, eventually leading to so-called slow and fast positions. We will call such models homotachous (from ‘‘same speed’’ in Greek). In such a static evolutionary framework, a fast evolving position will be so in any taxonomic group. However, as demonstrated by Fitch (1971), substitutions in the cytochrome *c* occur at different positions in fungi versus metazoa, which is incompatible with any homotachous model (Fitch 1971). This is, however, compatible with the covarion model (Fitch and Markowitz 1970). In this model, at a given time, only a fraction of the positions (called ‘‘c’’), the concomitantly variable codons (covarions), can accept substitutions, yet with the same probability for each of them. After a substitution, the probability of change of the covarions is 1 − ‘‘p’’ (p is called persistence). When such changes happen, a randomly chosen variable position becomes invariable and vice versa, since ‘‘c’’ is assumed to be constant. Nevertheless, studies have shown that the number of variable positions can be different between lineages (Germot and Philippe 1999), suggesting that a constant c is a limitation of the covarion model (Steel, Hudson, and Lockhart 2000). The model has been refined by including permanently variable and invariable positions (Fitch and Ye 1991). Although this framework appeared as early as 1971, it has never been shown to thoroughly explain the data. Fitch verified that, in a simulation under a covarion model, pairs of simulated sequences displayed the same amount of differences as real ones. However, this is not an extensive validation, and the authors admit that ‘‘the gamma [. . .] model is a viable alternative’’ (Miyamoto and Fitch 1995). In fact the covarion model did not receive

Key words: covarion, cytochrome *b,* molecular evolution, phylogeny, vertebrates.

Address for correspondence and reprints: Hervé Philippe, Phylogénie, Bioinformatique et Génome, UMR 7622 CNRS, Université Pierre et Marie Curie, 9, quai St. Bernard, 75005 Paris, France. E-mail: herve.philippe@snv.jussieu.fr.

Mol. Biol. Evol. 19(1):1–7. 2002

much attention until recently (Lockhart et al. 1998; Tuffley and Steel 1998; Lopez, Forterre, and Philippe 1999; Galtier 2001; Gaucher, Miyamoto, and Benner 2001).

Many proteins display global substitution rates of their positions that fit a Γ law (Uzzell and Corbin 1971; Yang 1996), explaining the current success of the RAS model. Until recently, the assumption that these rates are constant within a position has not been tested. Thanks to different statistical tools (Philippe et al. 1996; Lockhart et al. 1998; Gu 1999; Lopez, Forterre, and Philippe 1999; Gaucher, Miyamoto, and Benner 2001), it now has been convincingly demonstrated that the evolutionary rate of a given position is not always constant throughout time. These findings invalidate homotachous models but do not validate the covarion model either as a sufficient explanation of sequence evolution. For this reason, we coined the word heterotachous to describe such positions (Philippe and Lopez 2001), rather than the previous term covarion-like (Lockhart et al. 1998; Lopez, Forterre, and Philippe 1999). The need for a new term was also because of the possible confusion between covarion and covariation, which could be completely unrelated to heterotachy.

The rejection of homotachous models was always achieved with very divergent orthologs (archaea vs. eukaryota, plastids vs. cyanobacteria, or animals vs. plants [Fitch 1971; Miyamoto and Fitch 1995; Lockhart et al. 1998]), or between paralogs of different functions (Gu 1999; Lopez, Forterre, and Philippe 1999; Naylor and Gerstein 2000). In such cases, the functional constraints are likely different, which would explain the different distribution of variable sites. It has even been suggested that ‘‘the covarion theory can be treated as a special case of functional divergence’’ (Gu 1999). We instead think that heterotachy (e.g., the covarion theory) is more widely relevant. Therefore, we investigated the possible rejection of the homotachous models when functional changes were presumably ruled out.

As an accurate estimation of evolutionary rates requires a great amount of data, we focused on vertebrate cytochrome *b* for which more than 3,000 almost complete sequences are available. As the metabolism of the mitochondrion is homogeneous among vertebrates, the proteins of our data set could be reliably considered devoid of functional changes. It appeared that almost all variable positions in the cytochrome *b* are heterotachous, although there is likely no functional shift. We investigated the localization of heterotachous positions with respect to the three-dimensional structure of the protein, and did not find any clear relationships.

Methods

Data Collection and Taxon Sampling

We collected 2,744 amino acid sequences of vertebrate cytochrome *b* from the data banks. We then sampled 2,038 sequences out of these, corresponding to 32 large monophyletic groups, which were indisputably defined on both morphological and molecular data basis. Aligned sequences are available from the authors upon request. These following 32 groups were used. The number of species contained in each group and the parsimony length of the optimal tree relating these species were: Anseriformes (45 sp., 92), Arvicolinaea (103 sp., 264), Bathyergidae (27 sp., 292), Bovinae (45 sp., 139), Caprinae (35 sp., 153), Carnivora (95 sp., 490), Cervoidea (34 sp., 129), Cetacea (64 sp., 258), Chiroptera (23 sp., 156), Chondrichtyes (41 sp., 577), Ciconiiformes (83 sp., 479), Ctenomyidae (28 sp., 121), Cyprininae (47 sp., 148), Echimyidae (28 sp., 251), Galliformes (45 sp., 251), Geomyidae (65 sp., 435), Insectivora (52 sp., 268), Labroidei (68 sp., 289), Lagomorpha (29 sp., 120), Lepidosauria (80 sp., 1,169), Leuciscinae (102 sp., 311), Metatheria (102 sp., 946), Murinae (113 sp., 564), Nesomyinae (53 sp., 309), Passeriformes (232 sp., 1,165), Primates (75 sp., 765), Procellariiformes (78 sp., 257), Sciuridae (153 sp., 483), Sigmodontinae (118 sp., 672), Syngnathidae (45 sp., 127), Trogoniformes (20 sp., 159).

Test of the Distribution of the Number of Substitutions

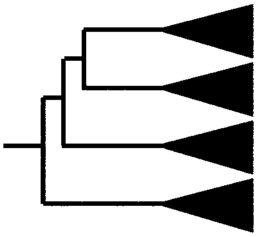
The distribution of the number of substitutions is compared to the distribution produced by the best-fitting Γ distribution (estimated by PAML [Yang 1997]). If they are not significantly different, as assessed by a 5% level chi-square test, then we consider the Γ distribution a good fit.

Test of Heterotachy

If the substitution rate is constant for a position, then the substitutions should be more or less evenly dispatched along the tree. As the data set is divided into monophyletic groups, a tree is computed for each of them. A position is then described by its profile, i.e., the number of changes it undergoes in every group. If a given position is homotachous, its profile should be proportional to the size (in steps) of the groups, which we test with a modified chi-square test (Lopez, Forterre, and Philippe 1999). Our method thus allows for determining how many positions significantly reject a homotachous behavior. The number of substitutions was inferred for each position either by maximum parsimony (MP), using PAUP 3.1 software (Swofford 1993) or by maximum likelihood (ML), with the help of GZ-AA software (Gu and Zhang 1997). A sufficient number of substitutions are necessary to yield significant statistical results. Therefore, a position is considered testable when undergoing a number of substitutions greater than half the number of groups.

Simulations

Sets of sequences were simulated on a template tree, obtained from the MP reconstruction of 200 sequences of vertebrate cytochrome *b* (four monophyletic groups of 50 sequences). Simulations under a homotachous model were performed by pSeq-Gen (Rambaut and Grassly 1997), and simulations under a covarion model were performed by simtree (Fitch and Ye 1991).



		Steps	Pos 50	Pos 57	Pos 71	Pos 247	Pos 251	Pos 325	Pos 360	Pos 364	Pos 374
Arvicolinae	**0.16**	**264**	*0*	*1*	**12**	*0*	*0*	**8**	**7**	**9**	3
Sigmodontinae	**0.21**	**242**	7	*1*	3	*0*	1	1	**7**	2	**6**
Murinae	**0.16**	**234**	5	**8**	6	6	1	*0*	*0*	1	1
Nesomyinae	**0.18**	**241**	**8**	6	*1*	**7**	**10**	1	2	*0*	*0*

FIG. 1.—Some heterotachous positions in the cytochrome *b* amino acid sequences of murids. Four monophyletic groups of Muridae, whose relationships are displayed on the left, are used to build this data set. A MP tree is computed for each group, and the numbers of steps of each tree are similar. Similarly, the α parameter for the distribution of their evolutionary rates is the same for each group. Each position is described by the number of substitutions it undergoes in each tree. As this description is not proportional to the lengths of the trees, a constant substitution rate can be rejected for the positions displayed here. These positions were chosen among the most variable ones, so that the rejection is highly significant. Note that the most variable group (displayed in bold face here) is not always the same. Heterotachy is thus different from a global evolutionary acceleration that would occur in a given group, and is unrelated to molecular clock issues (Kimura 1987). It should also be noted that these positions display different substitution profiles, suggesting there is no obvious covariation between them.

Results and Discussion

Heterotachy at Small Evolutionary Scale

Since an accurate estimation of site-by-site substitution rates requires a great amount of data, we focused on vertebrate cytochrome *b* for which more than 3,000 almost complete sequences are available. This protein could be reliably considered devoid of major functional changes, even if function is not a well-defined concept. Cytochrome *b* is always implicated in the respiratory chain complex involved in oxidative phosphorylation that is highly conserved among vertebrates. To strengthen our assumption, we first studied sequences of murids (rats and mice), a group of rodents that diversified about 50 MYA. The distribution of the number of substitutions per position fitted a Γ distribution quite well for the complete murid data set (data not shown). The Γ distribution shape parameter α was 0.21, indicating that a few positions accumulated most of the changes, whereas the majority did not. When we divided the data set into four monophyletic groups, each group showed a similar value for α (fig. 1), as expected under a homotachous model (which assumes that the substitution rate of a position is constant throughout time, even if it varies between positions). Yet, substitutions were not evenly distributed among the four groups (fig. 1). The rate was significantly heterotachous ($P < 0.05$) in 27% of the testable positions (see *Methods*). Even though similar shape parameters for the Γ distribution were observed for the data, a constant substitution rate for many positions can be confidently rejected. It has been suggested that, when the subgroups and the whole group have different α values, the data set should display heterotachy (Gu 1999; Gaucher, Miyamoto, and Benner 2001), but interestingly, this condition is not even necessary.

Quantification of Heterotachy and the Number of Sequences

Until the present study, the great extent to which homologous sequences can alter their distribution of variable sites during their divergence has not been obvious, because one has to infer many substitutions at a given position for a significant result to be obtained (i.e., a very large data set is needed). For instance, with two groups of 10 species, the observation of two substitutions in one group and none in the other shall be inconclusive. In contrast, with groups of 100 species, a distribution of 18 and 2 substitutions shall be significantly heterogeneous. Observing large numbers of substitutions can be achieved mainly by increasing the number of species, but also by using more efficient inference methods to detect multiple changes. As shown below, both approaches converge on the existence of an extensive heterotachy in cytochrome *b*.

First, we increased the number of species with a constant number of groups. For three groups (birds, mammals, and fishes), the number of significantly heterotachous positions steadily increased with the number of sequences, until an asymptotic value was reached (fig. 2*a*). With 10 sequences in each group (a common case for most markers), only 9% of the positions are detected as heterotachous, a value close to the 5% expected by chance. However, up to 47% are found with 300 sequences in each group. Second, we increased the number of species by increasing the number of groups. The number of significantly heterotachous positions increased rather linearly and converged to the 81% value observed with 32 groups. Contrary to the previous case (fig. 2*a*), the shape of the curve (fig. 2*b*) suggests that saturation is not reached, and that adding more groups will still allow the detection of more heterotachous positions, likely up to 100%. Third, large numbers of changes were underestimated by the MP method, which we used to reduce computational time, and this underestimation might reduce the detection of heterotachy. Whenever ML methods were applied, the percentage of heterotachous positions increased, up to 88% for our complete data set, rendering our MP-based estimations (81%) conservative. Finally, we considered testable the positions that underwent at least an average of 0.5 substitutions per group, for which the resolving power of our method can be weak. If this threshold is raised, an increase of the percentage of heterotachous positions is always observed. For instance, with 32 groups, at least

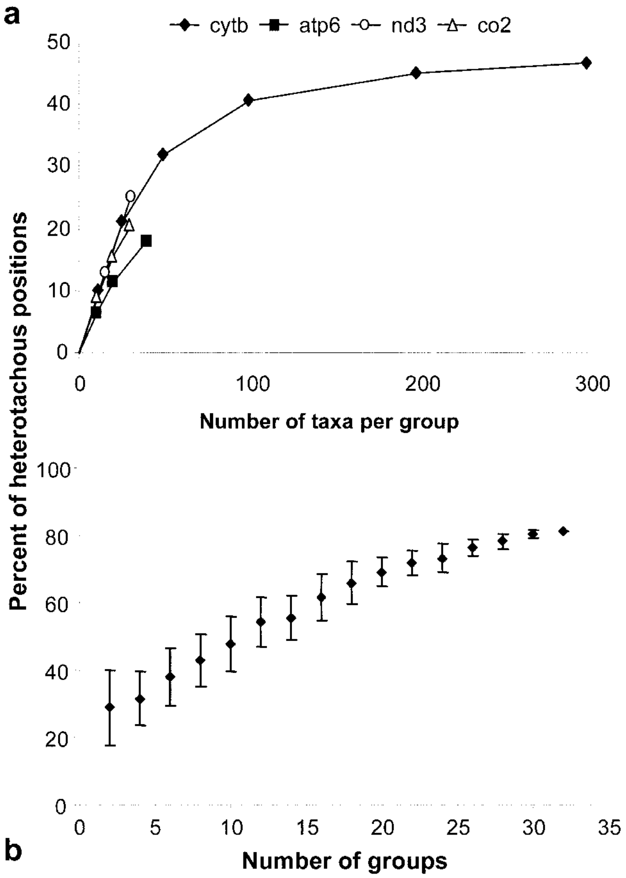


FIG. 2.—Evolution of the percentage of heterotachous positions with the number of taxa and the number of groups. *a,* The average percentage of heterotachous positions in the cytochrome *b* is computed for different numbers of randomly selected taxa chosen among three groups (fishes, birds, and mammals). Each result is obtained from 100 replicates, and the standard error is given as error bars. *b,* The average percentage of heterotachous positions in the cytochrome *b* is computed for data sets of 2–32 randomly selected groups with the same protocol.

one substitution per group and ML estimates, 95% of the positions are heterotachous.

Heterotachy is clearly a major evolutionary feature of vertebrate cytochrome *b.* Such an observation is currently difficult to generalize as very few data sets are rich enough to be characterized in the manner described here. Considering the three monophyletic groups (fishes, birds, and mammals) defined above for cytochrome *b,* we nevertheless investigated a few other mitochondrial proteins, and all argued for significant numbers of heterotachous sites when estimated under parsimony (ATP6 29%, ND3 39%, and CO_2 20% with 126, 197, and 161 species, respectively). The heterotachy levels were 37%, 44%, and 28% when ML estimation was used to infer the number of substitutions per site. These percentages were somewhat lower than for cytochrome *b,* but this might be expected because of the smaller number of available species, preventing estimation of the asymptotic value (fig. 2*a*). These results, along with the fact that cytochrome *b* is unlikely to be an exception, strongly suggest that heterotachy is a common feature of protein evolution, even when no functional changes occur.

Heterotachy and Function-structure Analysis

The fact that the functional constraints on a position do not stay the same throughout time should not be an unexpected outcome, both for intrinsic (protein structure) or extrinsic reasons (protein interactions) (Spiller et al. 1999). First, a single mutation can change the ensuing mutation probabilities of other positions (Fitch and Markowitz 1970). Second, the environment of the protein will necessarily change, especially because of substitutions occurring in interacting proteins. This is of course more relevant for proteins belonging to large complexes, like the cytochrome *b,* that have many interactions with other molecules (Xia et al. 1997). As an example of such environmental changes, the repopulation of mouse mitochondrial DNA–less cells with rat mitochondria restores translation but not respiratory functions in the mitochondrion (Yamaoka et al. 2000). Even if mouse and rat proteins are highly similar, it demonstrates that some mitochondrion-encoded rat proteins (e.g., cytochrome *b*) cannot interact properly with mouse nucleus–encoded proteins (e.g., cytochrome *c,* cytochrome oxidase IV), because few independent modifications of interacting proteins were enough to severely disturb the function of the complex. This result is in full agreement with our finding of heterotachy in murid cytochrome *b* (fig. 1). The same observations were also made on primates (Kenyon and Moraes 1997) and on yeast (Spirek et al. 2000). These experiments show how, in different lineages, coevolution of proteins canalizes the evolution of a protein in different directions, explaining why heterotachy can be found when function remains the same.

Since the crystal structure of the cytochrome *bc1* complex from bovine heart mitochondria has been solved (Xia et al. 1997), we have mapped the heterotachous sites on the three-dimensional structure (fig. 3). For clarity, the two subunits of cytochrome *b* that show opposite sides of the molecule are displayed, so that the eight transmembrane helices are easily visible (fig. 3*A*). When considering all vertebrates, patterns identifying constant and heterotachous positions can be observed to be evenly distributed across the structure. Patterns identifying homotachous positions are scarce. The distribution of the different pattern types could not be easily interpreted in terms of the three-dimensional structure. We suspect that the reason for this is the considerable evolutionary divergence separating the groups studied. We therefore focused on smaller taxonomic groups (rodents, fig. 3*B;* birds, fig. 3*C;* cetartiodactyles, fig. 3*D*). In these analyses, many homotachous positions did appear, but seemed also evenly distributed along the structure. From the naive idea that the α helices mainly serve to anchor cytochrome *b* in the membrane, one would expect that functional constraints acting on them remain the same throughout time. The presence of many heterotachous positions in these helices however demonstrated that their function may be more complex than simple anchoring.

As no clear pattern emerged from three-dimensional structure, we have investigated the secondary struc-

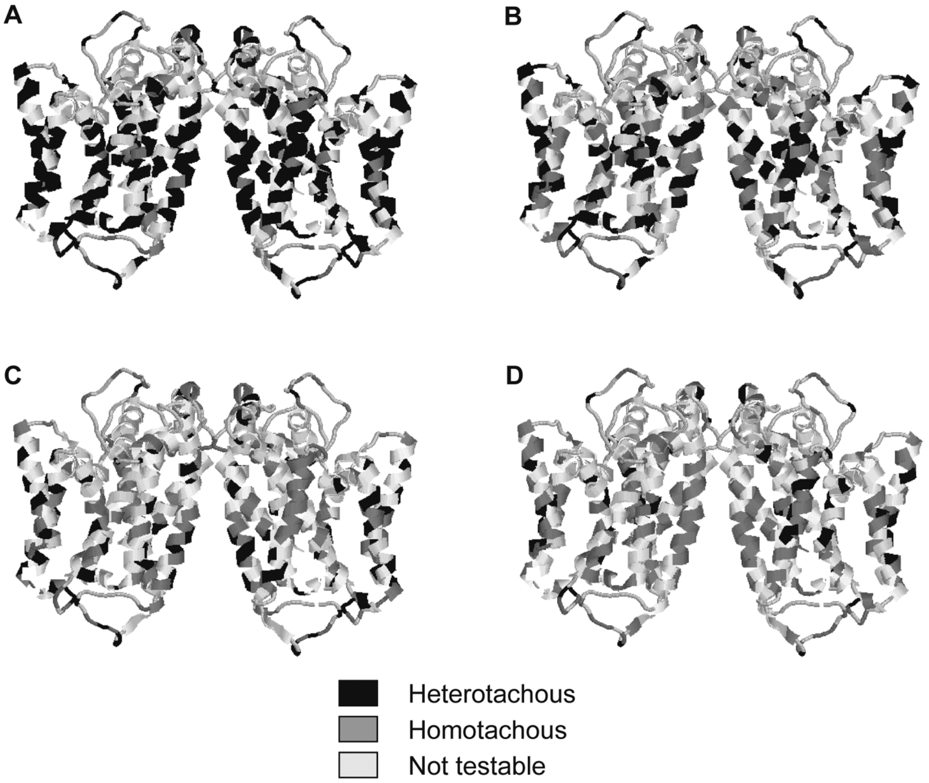


FIG. 3.—Heterotachous positions on the three-dimensional structure of bovine cytochrome *b.* The figure has been restricted to chains O and C of the PDB entry so that opposite sides of the molecule can be displayed. Different data sets have been used to detect heterotachy in cytochrome *b. A,* Vertebrate data set (32 groups). *B,* Rodents data set (nine groups). *C,* Birds data set (six groups). *D,* Cetartiodactyles data set (four groups) (for definition of groups see *Methods*). Positions are displayed in black when found heterotachous for the considered data set, whereas homotachous ones are displayed in gray. White positions did not undergo enough substitutions to be tested and are nearly constant. As there were too many partial sequences, the extremities of the protein were not included in the analysis and are thus not displayed here.

ture of cytochrome *b*. To have a sufficient number of positions per class, we have considered two classes only: α-helices and the rest (i.e., 3/10 helices, bends and hydrogen-bounded turns). Strikingly, the constant, homotachous and heterotachous positions were homogeneously distributed between the two classes (table 1), for both vertebrates and rodents. Chi-square tests of homogeneity indeed showed very high *P* values (0.89 and 0.85, respectively), meaning that the secondary structure had no impact on the probability for a position to be heterotachous. Such comparative analyses would certainly be more relevant if the structures of cytochrome *b* were known in different organisms. However, as three-dimensional structures are much more conserved than primary sequences, our results should likely remain valid when structures other than the bovine one appear.

Table 1
Distribution of Heterotachy in the Secondary Structure of Cytochrome *b*

	VERTEBRATES		RODENTS	
	α-Helix	Other	α-Helix	Other
Constant	106	65	112	77
Homotachous	18	9	50	30
Heterotachous.	79	48	41	25

NOTE.—Only two classes of secondary structures were considered because α-helices are overabundant in this molecule. Chi-square tests of homogeneity show that each class of position is equally represented in helices and in other structures (*P*-values of 0.89 and 0.85 for vertebrates and rodents, respectively).

The poor correlation of heterotachy and structure-function for cytochrome *b* (fig. 3 and table 1) is in agreement with the little success of structure-based rational design of proteins in directed evolution experiments (Tobin, Gustafsson, and Huisman 2000). In fact, the three-dimensional structure of proteins can be altered, even slightly, by any mutation, to such an extent that a functional change from aspartate to valine aminotransferase can happen after 17 mutations, all but one occurring outside of the active sites (Oue et al. 1999). Such observations suggest a new paradigm for understanding protein evolution. This would be ‘‘dispersed substitutions that act synergistically improve enzyme properties and function’’ (Tobin, Gustafsson, and Huisman 2000). The high levels of heterotachy we observed were likely to be caused by the fact that any position can become invariable (i.e., function critical) at a given time (i.e., within a given lineage). The importance of heterotachy provides an evolutionary explanation to the relative success of random design (e.g., error-prone PCR or DNA shuffling) with respect to structure-based design (Tobin, Gustafsson, and Huisman 2000).

Heterotachy and the Covarion Model

Considering the extent of heterotachy in protein evolution, sufficiently descriptive models of sequence

Table 2
Simulations of Sequence Evolution Under Different Models

	Observed α Parameter	Percentage of Heterotachy
Cytochrome *b*.	0.3	45
Γ model (α = 0.3).	0.3	7
Covarion model (p = 0.8, c = 0.1) . . .	0.3	78
Covarion model (p = 0.1, c = 0.35) . .	0.3	44

NOTE.—A 200 taxa cytochrome *b* data set is described by its alpha shape parameter and its percentage of heterotachous positions. The corresponding MP tree was used to simulate data sets under different models. Gamma RAS models are unable to reproduce the observations made on real data, as the percentage of heterotachous positions is much too low. Similarly, Fitch's covarion model (p stands for persistence and c for proportion of covarions) is unable to simultaneously reproduce both observations, leading to two suboptimal solutions.

evolution need to reproduce this feature. The models presently implemented in phylogenetic reconstruction (e.g., the Γ law model [Strimmer and von Haeseler 1996]) are homotachous, i.e., they assume that the substitution rate of a position is constant throughout time. Such a model seems at first glance appropriate for vertebrate cytochrome *b*, as the total number of substitutions for vertebrate cytochrome *b* fits a binomial negative distribution quiet well (data not shown), as is expected if the rate of substitutions is distributed according to a Γ law. But, we verified that sequences simulated under this model only display a level of heterotachy close to 5%, which is the level expected by chance (table 2). A single heterotachous model, the covarion one, has been proposed in 1970 by Fitch and Markowitz. In this model, at a given time, only a fraction of the positions, the covarions, can accept substitutions, yet with the same probability for each of them. After a substitution, the covarion pool has a fixed probability to change. When such changes happen, a randomly chosen variable position becomes invariable and vice versa. In order to know whether the covarion model explains our observations, we have conducted extensive simulations of sequence evolution (see *Supplementary Material*). In brief, the covarion model was able to generate heterotachous positions and binomial negative distribution of the number of substitutions or both, depending on the values of the two free parameters of the model. Unfortunately, no values (table 2) can simultaneously reproduce the observations made on cytochrome *b* (too many heterotachous positions for the correct α parameter or too high an α parameter for the correct fraction of heterotachous positions). Thus, our observations are not explained by any current model of sequence evolution.

The wealth of sequence data recently produced allowed us to demonstrate the extent of a major process of protein evolution, heterotachy. This feature opens new avenues of research. For example, sequence similarity searches like PSI-BLAST would greatly benefit from taking into account the fact that a position can be invariable only during most of its history. This might improve the detection of very distantly related homologs, which are for now only detected through the comparison of three-dimensional structures (Brenner, Chothia, and Hubbard 1998). In function-structure analyses, why do substitution-accepting positions differ in two related proteins, whereas their functions are the same? Do substitution-accepting positions differ more in two homologous proteins with different functions, as recently suggested (Gu 1999; Naylor and Gerstein 2000; Gaucher, Miyamoto, and Benner 2001)? What are the implications of heterotachy to sequence analysis-base protein fold predictions or to modeling of directed protein evolution? Another major issue would be to find a model of sequence evolution that explains heterotachy, and to implement it in phylogenetic reconstruction methods (Galtier 2001). Such an improvement should help in solving some highly debated phylogenetic questions, especially the ancient ones.

Acknowledgments

We thank Andrew Roger and David Moreira for careful reading of the manuscript.

LITERATURE CITED

ADACHI, J., and M. HASEGAWA. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. **42**:459–468.

ADACHI, J., P. J. WADDELL, W. MARTIN, and M. HASEGAWA. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J. Mol. Evol. **50**:348–358.

ALTSCHUL, S. F., and E. V. KOONIN. 1998. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. Trends Biochem. Sci. **23**:444–447.

BRENNER, S. E., C. CHOTHIA, and T. J. HUBBARD. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc. Natl. Acad. Sci. USA **95**:6073–6078.

DAYHOFF, M. O. 1979. Atlas of protein sequence and structure. Vol. 5, Suppl. 3, 1978. National Biomedical Research Foundation, Washington, D.C.

FITCH, W. M. 1971. The nonidentity of invariable positions in the cytochromes *c* of different species. Biochem. Genet. **5**: 231–241.

FITCH, W. M., and E. MARKOWITZ. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. **4**:579–593.

FITCH, W. M., and J. YE. 1991. Weighted parsimony: does it work? Pp. 147–154 *in* M. M. MIYAMOTO and J. CRACRAFT, eds. Phylogenetic analysis of DNA sequences. Oxford University Press, New York.

GALTIER, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. **18**:866–873.

GAUCHER, E. A., M. M. MIYAMOTO, and S. A. BENNER. 2001. Function–structure analysis of proteins using covarion-based evolutionary approaches: elongation factors. Proc. Natl. Acad. Sci. USA **98**:548–552.

GERMOT, A., and H. PHILIPPE. 1999. Critical analysis of eukaryotic phylogeny: a case study based on the HSP70 family. J. Eukaryot. Microbiol. **46**:116–124.

GU, X. 1999. Statistical methods for testing functional divergence after gene duplication. Mol. Biol. Evol. **16**:1664–1674.

GU, X., and J. ZHANG. 1997. A simple method for estimating the parameter of substitution rate variation among sites. Mol. Biol. Evol. **14**:1106–1113.

Huelsenbeck, J. P. 1998. Systematic bias in phylogenetic analysis: is the Strepsiptera problem solved? Syst. Biol. **47**: 519–537.

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.

———. 1994. A mutation data matrix for transmembrane proteins. FEBS Lett. **339**:269–275.

Kenyon, L., and C. T. Moraes. 1997. Expanding the functional human mitochondrial DNA database by the establishment of primate xenomitochondrial cybrids. Proc. Natl. Acad. Sci. USA **94**:9,131–9,135.

Kimura, M. 1987. Molecular evolutionary clock and the neutral theory. J. Mol. Evol. **26**:24–33.

Lockhart, P. J., M. A. Steel, A. C. Barbrook, D. Huson, M. A. Charleston, and C. J. Howe. 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol. Biol. Evol. **15**:1183–1188.

Lopez, P., P. Forterre, and H. Philippe. 1999. The root of the tree of life in the light of the covarion model. J. Mol. Evol. **49**:496–508.

Miyamoto, M. M., and W. M. Fitch. 1995. Testing the covarion hypothesis of molecular evolution. Mol. Biol. Evol. **12**:503–513.

Naylor, G. J., and M. Gerstein. 2000. Measuring shifts in function and evolutionary opportunity using variability profiles: a case study of the globins. J. Mol. Evol. **51**:223–233.

Oue, S., A. Okamoto, T. Yano, and H. Kagamiyama. 1999. Redesigning the substrate specificity of an enzyme by cumulative effects of the mutations of non-active site residues. J. Biol. Chem. **274**:2344–2349.

Philippe, H., and A. Germot. 2000. Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. Mol. Biol. Evol. **17**:830–834.

Philippe, H., G. Lecointre, H. L. V. Lê, and H. Le Guyader. 1996. A critical study of homoplasy in molecular data with the use of a morphologically based cladogram. Mol. Biol. Evol. **13**:1174–1186.

Philippe, H., and P. Lopez. 2001. On the conservation of protein sequences in evolution. Trends Biochem. Sci. **26**:414–416.

Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. **13**:235–238.

Rzhetsky, A., and M. Nei. 1994. Unbiased estimates of the number of nucleotide substitutions when substitution rate varies among different sites. J. Mol. Evol. **38**:295–299.

Spiller, B., A. Gershenson, F. H. Arnold, and R. C. Stevens. 1999. A structural view of evolutionary divergence. Proc. Natl. Acad. Sci. USA **96**:12305–12310.

Spirek, M., A. Horvath, J. Piskur, and P. Sulo. 2000. Functional co-operation between the nuclei of *Saccharomyces cerevisiae* and mitochondria from other yeast species. Curr. Genet. **38**:202–207.

Steel, M., D. Huson, and P. J. Lockhart. 2000. Invariable sites models and their use in phylogeny reconstruction. Syst. Biol. **49**:225–232.

Strimmer, K., and A. von Haeseler. 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. **13**:964–969.

Sullivan, J., and D. L. Swofford. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J. Mamm. Evol. **4**:77–86.

Swofford, D. L. 1993. PAUP: phylogenetic analysis using parsimony. Version 3.1.1. Illinois Natural History Survey, Champaign.

Tobin, M. B., C. Gustafsson, and G. W. Huisman. 2000. Directed evolution: the ‘rational’ basis for ‘irrational’ design. Curr. Opin. Struct. Biol. **10**:421–427.

Tuffley, C., and M. Steel. 1998. Modeling the covarion hypothesis of nucleotide substitution. Math. Biosci. **147**:63–91.

Uzzell, T., and K. W. Corbin. 1971. Fitting discrete probability distributions to evolutionary events. Science **172**: 1089–1096.

Van de Peer, Y., A. Ben Ali, and A. Meyer. 2000. Microsporidia: accumulating molecular evidence that a group of amitochondriate and suspectedly primitive eukaryotes are just curious fungi. Gene **246**:1–8.

Xia, D., C. A. Yu, H. Kim, J. Z. Xia, A. M. Kachurin, L. Zhang, L. Yu, and J. Deisenhofer. 1997. Crystal structure of the cytochrome *bc1* complex from bovine heart mitochondria. Science **277**:60–66.

Yamaoka, M., K. Isobe, H. Shitara, H. Yonekawa, S. Miyabayashi, and J. I. Hayashi. 2000. Complete repopulation of mouse mitochondrial DNA-less cells with rat mitochondrial DNA restores mitochondrial translation but not mitochondrial respiratory function. Genetics **155**:301–307.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. **11**:367–370.

———. 1997. PAML: phylogenetic analysis by maximum likelihood. Version 1.3. Department of Integrative Biology, University of California at Berkeley.

Franz Lang, reviewing editor

Accepted July 20, 2001