THE ROYAL SOCIETY

Received 26 February 2002
Accepted 20 March 2002
Published online 12 June 2002

A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes

Tal Pupko[1] **and Nicolas Galtier**[2*]

[1]*The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu Minato-ku, Tokyo 106-8569, Japan*
[2]*CNRS UMR 5000—'Génome, Populations, Interactions', Université Montpellier 2, Place E. Bataillon, 34095 Montpellier, France*

A new method for detecting site-specific variation of evolutionary rate (the so-called covarion process) from protein sequence data is proposed. It involves comparing the maximum-likelihood estimates of the replacement rate of an amino acid site in distinct subtrees of a large tree. This approach allows detection of covarion at the gene or the amino acid levels. The method is applied to mammalian–mitochondrial–protein sequences. Significant covarion-like evolution is found in the (simian) primate lineage: some amino acid positions are fast-evolving (i.e. unconstrained) in non-primate mammals but slow-evolving (i.e. highly constrained) in primates, and some show the opposite pattern. Our results indicate that the mitochondrial genome of primates reached a new peak of the adaptive landscape through positive selection.

Keywords: covarion; maximum likelihood; mitochondrial genome; positive Darwinian selection; primates

1. INTRODUCTION

Understanding species adaptation at the molecular level is a challenge in biological science. Most genomic changes occurring between species are neutral (i.e. do not affect the fitness of individuals) or slightly deleterious, and are fixed by genetic drift (Kimura 1983). Unravelling the molecular basis of adaptation, therefore, involves discriminating the few adaptive changes from the many neutral ones (Golding & Dean 1998).

Positive selection at the molecular level has been essentially detected in specific categories of genes, including genes that are involved in cell signalling (Hughes & Nei 1988) or male sex-related functions (Wyckoff *et al.* 2000). Recently, Liberles *et al.* (2001) composed a database of evolutionary families, for which positive Darwinian selection is indicated. Though not found in the database of Liberles *et al.* (2001), evidence for positive selection in the primate mitochondrial genome was recently reported. The animal mitochondrial genome is widely considered to be evolving in a neutral or nearly neutral fashion. This belief comes mainly from the fact that mitochondrial genes encode proteins involved into cellular respiration, i.e. a basic metabolic pathway common to all the eukaryotes. Respiration function appears to be quite similar between species and hardly dependent on environmental changes (at least non-extreme ones). Several authors, however, have reported a significant increase of the non-synonymous–synonymous rate of evolution for some mitochondrial genes in the simian primate lineage (Wu *et al.* 1997; Andrews *et al.* 1998; Andrews & Easteal 2000; Grossman *et al.* 2001). This was interpreted as a consequence of molecular adaptation: many advantageous non-synonymous changes would have recently been fixed in primates. An increase of the non-synonymous substitution rate, however, is also expected under the hypothesis of relaxed constraints, possibly as a consequence of reduced population size. In this study, we develop a distinct approach to the same issue.

Functional changes in a protein involve modification of its tertiary structure. When a new structure is reached, the constraints applying to specific amino acids can change. Some positions were critical in the former structure but no longer have any functional relevance in the new structure, and vice versa. This must be reflected by changes in the evolutionary rate of such amino acid sites. A position evolves slowly when it is functionally important, but changes rapidly when neutral. A structural change, therefore, can be detected from the phylogenetic history of a protein by seeking a lineage in which several amino acid positions show a significant increase or decrease in evolutionary rate. This rationale was applied to mammalian mitochondrial proteins through a new, model-based method of amino acid sequence analysis. We compared the rate of evolution of every amino acid, which were estimated separately in the primate lineage and the non-primate group. For a given amino acid site, highly different rates in the two subgroups reveal a change in the functional constraints. The significance of the difference between the two groups was assessed using a likelihood-ratio test. The notion that the evolutionary rate of a site can change in time and between lineages was called 'covarion' by Fitch (1971). The covarion process has recently received attention in the field of molecular phylogenetics (e.g. Tuffley & Steel 1998; Lopez *et al.* 1999; Galtier 2001).

2. MATERIAL AND METHODS

Twelve mitochondrial genes from 34 completely sequenced mammalian mitochondrial genomes were analysed. The

* Corresponding author (galtier@univ-montp2.fr).

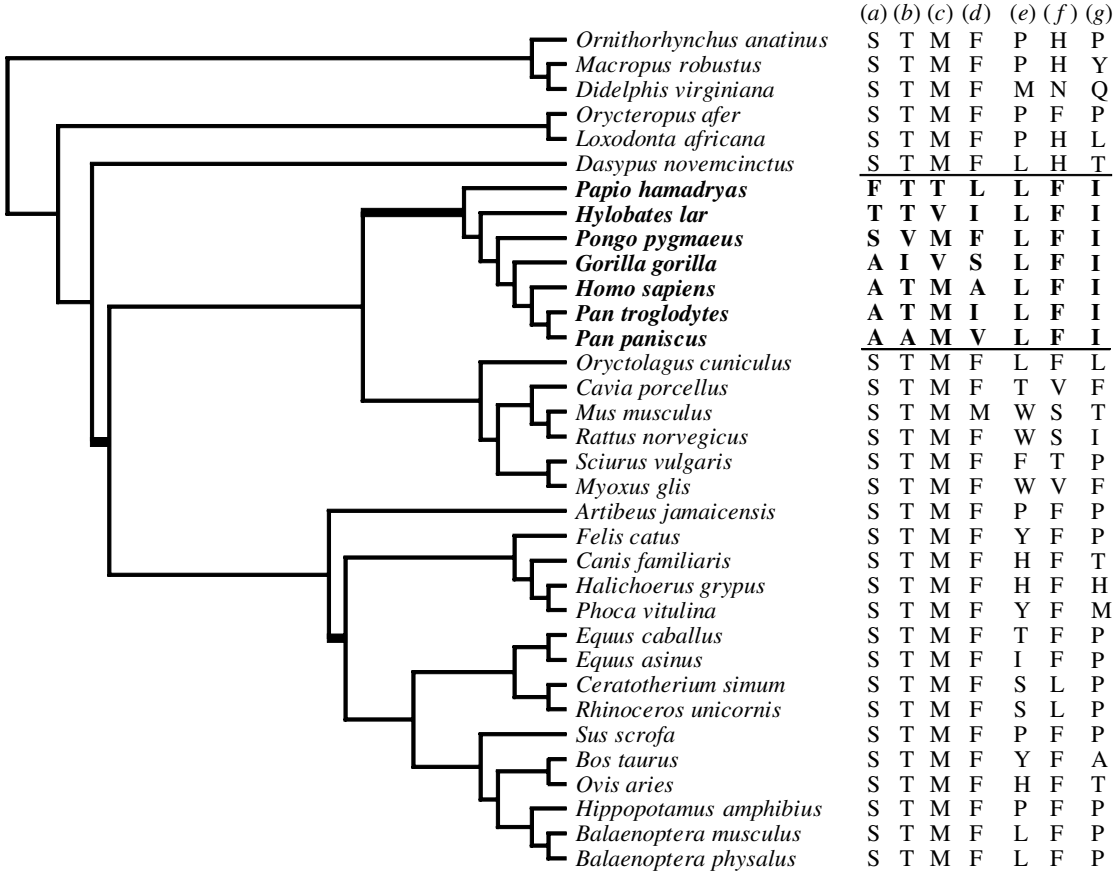

Figure 1. Phylogenetic tree of mammals used in this study (Murphy *et al.* 2001). The primate species names are in bold. Seven mitochondrial amino acid sites showing highly significant covarion are displayed as follows: (*a*) *atp6*(124); (*b*) *co3*(2); (*c*) *nd2*(187); (*d*) *nd5*(202); (*e*) *nd5*(484); (*f*) *nd5*(520); (*g*) *nd5*(537), where the numbers indicate the amino acid positions in the human sequence. Sites (*a*)–(*d*) are monomorphic in non-primate mammals but highly variable in primates, while sites (*e*)–(*g*) show the opposite. The thick internal branches mark the cutting points used for the test.

sequences were aligned using CLUSTALX (Thompson *et al.* 1997). For each gene, the branch lengths of the mammalian tree were estimated according to the maximum-likelihood method (Felsenstein 1981) under the REV + Gamma model of evolution (Cao *et al.* 2000). The reversible matrix was chosen because it was based on mitochondrial data and hence better reflects the replacement probabilities of amino acid changes in mitochondrial genomes. Estimating the branch length with the Gamma model is more accurate because it takes into account any among-site rate variation. The phylogenetic tree of figure 1 was assumed (Murphy *et al.* 2001). The dataset D was then split into two subsets, namely primate (D_1, seven species) versus non-primate mammals (D_2, 27 species). For every amino acid position i, the (relative) evolutionary rates in the primate (r_1^*) and non-primate (r_2^*) lineages at position i were estimated separately from the two datasets, using the corresponding two subtrees T_1 (primate tree) and T_2 (non-primate tree). $r_1^*(i)$ is the optimal scaling factor for tree T_1, i.e. the value of r_1 that maximizes the likelihood for site i in data D_1 computed after all the branch lengths in T_1 were multiplied by r_1 (relative branch lengths within T_1 are kept the same). Highly different $r_1^*(i)$ and $r_2^*(i)$ values reveal a change in the functional constraint at site i in the primates. The significance of the difference between $r_1^*(i)$ and $r_2^*(i)$ was assessed using a likelihood-ratio test

$$\mathrm{LR}(i) = 2\log[\max(L_1(r)L_2(r))] - 2\log[\max(L_1(r_1))\max(L_2(r_2))], \quad (2.1)$$

where L_1 is the likelihood for dataset D_1, L_2 is the likelihood for dataset D_2 and the maximizations are over r, r_1 and r_2, respectively. The first term on the right-hand side of equation (2.1) is twice the logarithm of the maximum likelihood for site i under the assumption of a common rate in subtrees T_1 and T_2. The second term in equation (2.1) is twice the logarithm of the maximum likelihood for site i assuming distinct rates in the two subtrees. The statistical measure LR(i) is asymptotically χ^2-distributed (1 d.f.) under the hypothesis of a constant evolution rate at site i. If the gain in likelihood that was obtained by relaxing the constant-rate assumption is high enough, then position i is considered to be evolving in a covarion-like manner. If m independent sites are examined in this way (say, the m sites of a given gene), then the number of sites showing a significant departure from the χ^2-distribution at the 1% level follow a binomial distribution B(m, 0.01), making it possible to test the site-specific rate constancy at gene level.

3. RESULTS

Out of the 3578 mitochondrial amino acid sites analysed, 62 showed a significant (1% level) departure from the hypothesis of a constant evolutionary rate in primates and non-primates, compared with 35.78 ± 5.95 that would be expected just by chance ($p < 0.001$, table 1). When genes were considered separately, a significant

Table 1. Covarion test for 12 mitochondrial protein coding genes in mammals.

gene	length[a]	site test[b]
atp6	226	8 (2.26)[c]
atp8	63	1 (0.63)
co1	513	3 (5.13)
co2	227	3 (2.27)
co3	261	3 (2.61)
cytb	378	2 (3.78)
nd1	314	8 (3.14)[c]
nd2	343	7 (3.43)[c]
nd3	113	1 (1.13)
nd4	459	5 (4.59)
nd4L	97	2 (0.97)
nd5	584	19 (5.84)[c]
all	3578	62 (35.78)[c]

[a] After sites including at least one gap have been removed.
[b] Number of sites showing a significant change of evolutionary rate (1% level). Numbers in parentheses show expected number under a constant rate in each site.
[c] A significant difference.

excess of such sites was found in four genes out of 12, namely *atp6*, *nd1*, *nd2* and *nd5* (table 1). The discrepancy between the primate and non-primate rates was sometimes striking, as shown in figure 1. Site 124 of gene *atp6* is conserved as a serine in every non-primate species, including two marsupials and a monotreme, which diverged *ca*. 160 Myr ago (Kumar & Hedges 1998). Surprisingly, this site shows four distinct states in the seven primate species analysed, which diverged less than 40 Myr ago (Cao *et al.* 2000). Conversely, site 484 of *nd5* is highly variable in non-primates (for example, a horse and a donkey have distinct states, as do different seal species), whereas all seven primates share a common leucine. About five changes (4.62 ± 2.09) would be expected in the primate subtree if this site had evolved at the rate observed in non-primates. The functional requirements of these sites are obviously different in primates and non-primates, strongly indicating that the structure of mitochondrial proteins is different in these two groups.

Several controls were performed to check the robustness of these results. Results were essentially unchanged when a different phylogenetic tree (Reyes *et al.* 2000) was used (not shown). The analysis was re-conducted using distinct 'cuts' in the mammalian tree (marked by thick branches in figure 1). Only the primate versus non-primate split yielded a significant amount of site-specific rate variation. The result appears to be specific to primates. Finally, different models of amino acid replacement, such as DAY and JTT (Jones *et al.* 1992), gave very similar results.

4. DISCUSSION

Our results indicate that the three dimensional (3D) structure of primate mitochondrial proteins is different from that of other mammals. The evolutionary significance of this structural change is an exciting issue and leads to the question of why primates achieve respiration 'differently' from other mammals. It is plausible that a neutral (or slightly deleterious) structural change happened by chance through genetic drift, for example, in the context of reduced population size in primates. Note, however, that a reduced population size should have resulted in the acceleration of the genomic non-synonymous substitution rate in primates (reduced efficacy of purifying selection in small populations)—a pattern not detected from nuclear data (Madsen *et al.* 2001). Alternatively, it is possible that the new mitochondrial structure is advantageous for primates, but not for other mammals. At first sight, this appears very unlikely given the physiological closeness among eutherian orders. We propose, however, a selective hypothesis involving adaptation to variable tissue-specific constraints. Optimizing cellular respiration in different tissues may require distinct structures of mitochondrial proteins, because of the differential expression of interacting nuclear genes. Unfortunately, a species must choose: every individual carries a single mitochondrial genome and it is now possible that primates have made a 'choice' that is different from other species'. The fittest mitochondria for a primate might be mitochondria that are optimal in certain tissues (i.e. nerves), while other mammals favour mitochondria that work well in other tissues (i.e. muscle). Wu *et al.* (1997), followed by Grossman *et al.* (2001), indicated that the larger neocortex is the evolutionary pressure responsible for the high rate of evolution of the aerobic metabolism in anthropoid primates. This scenario is highly speculative and we leave it as a working hypothesis for future (experimental) work.

Whatever the evolutionary mechanism (adaptation or drift) that underlies the structural changes in primate mitochondrial proteins, we argue that this change must have generated an episode of adaptation at the sequence level. This is because the new advantageous, equally fit or even disadvantageous stable structure cannot have appeared via a single mutation. Let us consider the many sites that appear to be constrained in primates (new structure), but unconstrained in other mammals (former structure). Not all of these sites can have reached their new optimal state simultaneously, nor by chance through random drift. An 'optimization' stage must have occurred, during which the mitochondrial proteins have 'climbed' the newly reached peak of the adaptive landscape through the fixation of favourable alleles at several amino acid sites, in an ancestral primate population.

Mitochondrial DNA is by far the most popular marker for phylogenetic and population genetics studies (Avise 1994). The reason most often invoked for using it is its near neutrality: the (claimed) lack of selective effects makes it a suitable tool for reconstructing species and population history. Our results cast serious doubts on this rationale. Several sites have evidently undergone an adaptive episode when optimizing the newly reached structure. Such events might still be occurring in some present-day species, thus distorting the historical signal.

The standard approach for detecting positive selection involves comparing non-synonymous (K_a) and synonymous (K_s) evolutionary rates. This method is powerful when positive selection is a recurrent process, making the 'equilibrium' K_a:K_s ratio high. Adaptation, however, might occur through short episodes of positive selection followed by long periods of purifying selection. The K_a:K_s ratios calculated from present-day sequences can be used to give the average value of the effects of positive and

negative selection over long periods of time, making the detection of adaptive episodes difficult, especially in the case of old ones. Another limitation is the saturation of synonymous positions, which precludes any comparison with divergent genes. Saturation at the third-codon position was found to be prominent even for within-order comparisons of cytochrome *b* sequences in mammals (Kenneth & Robinson 1999).

The covarion approach used in this study seems to avoid most of the problems mentioned in the previous section. This approach does not aim to recover the signal produced during short selective episodes. Rather, it compares long-term evolutionary processes before and after a putative structural change. Hence we do not expect lower power against old vs recent episodes of structural change. In addition, neither a global relaxation of constraints nor a change in the mutation rate should be falsely detected because the relative, not absolute, rates of distinct sites are measured. The method, however, has some requirements. First, a prior guess about the relevant cutting point (i.e. the putative phylogenetic location of an episode of positive selection) is needed. Second, each of the two subtrees must include a reasonable number of taxa. With a few sequences, the site-specific rate estimation is inaccurate, thus decreasing the power of the test (while two sequences can be enough to detect a high K_a:K_s ratio). Third, as already mentioned, the test cannot be used for the hypothesis of recurrent positive selection, by contrast with methods involving K_a:K_s comparisons. The two methods appear complementary and should be combined for an optimal characterization of molecular adaptation.

Finally, we would like to emphasize again that detecting positive selection at the sequence level (either by K_a:K_s comparisons or using the covarion approach) does not imply that there was adaptation at the structural (and functional) level. A protein might neutrally switch between equally efficient 3D structures, i.e. equally high peaks of the adaptive landscape. If peaks are 'separated' by a large number of amino acid changes (this study), this will imply a stage of molecular 'optimization', i.e. adaptation at the sequence level. The substitutions occurring during such optimization stages have a major impact on molecular evolution studies, as they:

(i) contribute to the total number of adaptive changes measured from sequence comparisons (e.g. Fay *et al*. 2001);
(ii) influence the long-term pattern of sequence evolution; and
(iii) locally distort the signal about the population and species history.

These changes, however, can occur independently of any functional innovation, and this might hinder progress in connecting molecular evolution to phenotype evolution.

We thank two anonymous referees for their helpful suggestions and discussion. T.P. is supported by a JSPS fellowship. N.G. is supported by the 'Génopole Montpellier Languedoc Roussillon'.

REFERENCES

Andrews, T. D. & Easteal, S. 2000 Evolutionary rate acceleration of cytochrome *c* oxidase subunit I in simian primates. *J. Mol. Evol.* **50**, 562–568.

Andrews, T. D., Jermiin, L. S. & Easteal, S. 1998 Accelerated evolution of cytochrome *b* in simian primates: adaptive evolution in concert with other mitochondrial proteins? *J. Mol. Evol.* **43**, 249–257.

Avise, J. C. 1994 *Molecular markers, natural history and evolution*. New York: Chapman & Hall.

Cao, Y., Fujiwara, M., Nikaido, M., Okada, N. & Hasegawa, M. 2000 Interordinal relationships and timescale of eutherian evolution as inferred from mitochondrial genome data. *Gene* **259**, 149–158.

Fay, J., Wycoff, G. J. & Wu, C. I. 2001 Positive and negative selection on the human genome. *Genetics* **158**, 1227–1234.

Felsenstein, J. 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376.

Fitch, W. M. 1971 Rate of change of concomitantly variable codons. *J. Mol. Evol.* **1**, 84–96.

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

Golding, G. B. & Dean, A. M. 1998 The structural basis of molecular adaptation. *Mol. Biol. Evol.* **15**, 355–369.

Grossman, L. I., Schmidt, T. R., Wildman, D. E. & Goodman, M. 2001 Molecular evolution of aerobic energy metabolism in primates. *Mol. Phylogenet. Evol.* **18**, 26–36.

Hughes, A. L. & Nei, M. 1988 Nucleotide substitution at major histocompatibility complex loci reveals overdominant selection. *Nature* **335**, 167–170.

Jones, D. T., Taylor, W. R. & Thornton, J. M. 1992 The rapid generation of mutation data matrices from protein sequences. *Comp. Appl. Biosci.* **8**, 275–282.

Kenneth, M. H. & Robinson, T. J. 1999 Multiple substitutions affect the phylogenetic utility of cytochrome *b* and 12 rDNA data: examining a rapid radiation in Leporid (Lagomorpha) evolution. *J. Mol. Evol.* **48**, 369–379.

Kimura, M. 1983 *The neutral theory of molecular evolution*. Cambridge University Press.

Kumar, S. & Hedges, S. 1998 A molecular timescale for vertebrate evolution. *Nature* **392**, 917–920.

Liberles, D. A., Schreiber, D. R., Govindarajan, S., Chamberlin, S. G. & Benner, S. A. 2001 The adaptive evolution database. *Genome Biol.* **2**, 1–6.

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

Madsen, O., Scally, M., Donady, C. J., Kao, C. J., De Brys, R. W., Adkins, R., Amrine, H. M., Stanhope, M. J., de Jong, W. W. & Springer, M. S. 2001 Parallel adaptive radiations in two major clades of placental mammals. *Nature* **409**, 610–614.

Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A. & O'Brien, S. J. 2001 Molecular phylogenetics and the origin of placental mammals. *Nature* **409**, 614–618.

Reyes, A., Gissi, C., Pesole, G., Catzeflis, F. M. & Saccone, C. 2000 Where do rodents fit? Evidence from the complete mitochondrial genome of *Sciurus vulgaris*. *Mol. Biol. Evol.* **17**, 979–983.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. 1997 The CLUSTALX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acid Res.* **24**, 4876–4882.

Tuffley, C. & Steel, M. A. 1998 Modelling the covarion hypothesis of nucleotide substitution. *Math. Biosci.* **147**, 63–91.

Wu, W., Goodman, M., Lomax, M. I. & Grossman, L. I. 1997 Molecular evolution of cytochrome *c* oxidase subunit IV: evidence for positive selection in simian primates. *J. Mol. Evol.* **44**, 477–491.

Wyckoff, G. J., Wang, W. & Wu, C.-I. 2000 Rapid evolution of male reproductive genes in the descent of man. *Nature* **403**, 304–309.