

# A simple covarion-based approach to analyse nucleotide substitution rates

J. SILTBERG & D. A. LIBERLES

Department of Biochemistry and Biophysics and Stockholm Bioinformatics Center, Stockholm University, Stockholm, Sweden

## Keywords:

adaptive evolution;  
bioinformatics;  
gene sequence;  
molecular evolution;  
protein function.

## Abstract

Using the ratio of nonsynonymous to synonymous nucleotide substitution rates ( $K_a/K_s$ ) is a common approach for detecting positive selection. However, calculation of this ratio over a whole gene combines amino acid sites that may be under positive selection with those that are highly conserved. We introduce a new covarion-based method to sample only the sites potentially under selective pressure. Using ancestral sequence reconstruction over a phylogenetic tree coupled with calculation of  $K_a/K_s$  ratios, positive selection is better detected by this simple covarion-based approach than it is using a whole gene analysis or a windowing analysis. This is demonstrated on a synthetic dataset and is tested on primate leptin, which indicates a previously undetected round of positive selection in the branch leading to *Gorilla gorilla*.

## Introduction

To understand the molecular basis for species divergence under selective pressure, it is valuable to have effective methods to detect positive selection along specific branches of phylogenetic trees (Liberles *et al.*, 2001). Detecting a ratio of nonsynonymous to synonymous nucleotide substitution rates ( $K_a/K_s$ ) significantly  $>1$  is a common method for detecting positive selection (Yang & Bielawski, 2000). To estimate this ratio on specific branches of a phylogenetic tree, ancestral sequences can be reconstructed across all nodes of the phylogenetic tree using either parsimony (Fitch, 1971) or maximum likelihood (Yang *et al.*, 1995; Koshi & Goldstein, 1996), and  $K_a/K_s$  calculated between ancestral sequences along all branches (Benner *et al.*, 1998; Liberles, 2001). Alternatively,  $K_a$  and  $K_s$  can be calculated pairwise between sequences and the values reconstructed using a distance algorithm like least squares (Zhang *et al.*, 1998), neighbour joining or unweighted pair-group method with arithmetic mean (UPGMA) (J. Siltberg and D.A. Liberles, unpublished results).

These methods for estimating  $K_a/K_s$  ratios along branches consider a whole gene. However, during the course of evolution, some amino acid sites are strictly conserved (e.g. for protein folding) whereas others are subject to positive selection (e.g. in a ligand binding pocket, where the ligand specificity is changing). To detect regions of a protein that are under positive selective pressure when the whole protein may not be, Endo *et al.* (1996) looked at windows of 20 contiguous codons in the primary sequence. This has subsequently become a common method for estimating  $K_a/K_s$ .

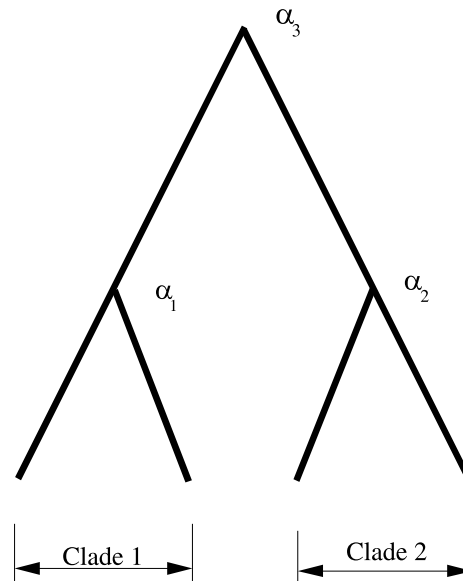
However, selection is occurring on a folded protein and residues that interact or effect function may not be proximal in primary sequence. An approach that takes this into consideration may be desirable. Miyamoto & Fitch (1995) developed a simple approach to categorizing amino acid sites in a protein into two broad categories, permanently invariable (pi) and potentially variable (pv). Potentially variable sites are composed of currently variable sites plus temporarily invariable sites among sequences analysed. This idea has been built on to develop methodologies to detect genes and sites under positive selective pressure (Gu, 1999, 2001; Gaucher *et al.*, 2001; Liberles, 2001).

A  $\Gamma$  function can be used to describe the distribution of substitution rates across sites in a protein, where  $\alpha$  is the variable describing the shape of the function, and varies over a wide range in different protein families under

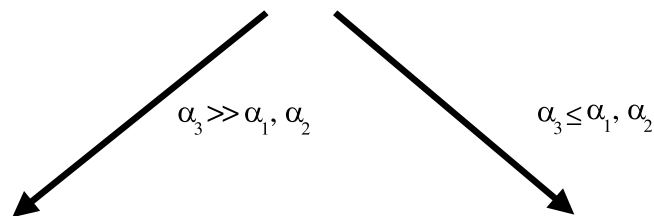
Correspondence: David A. Liberles, Department of Biochemistry and Biophysics and Stockholm Bioinformatics Center, Stockholm University, 10691 Stockholm, Sweden.  
Tel.: +46-8-55378565; fax: +46-8-55378214; e-mail: liberles@sbc.su.se

different selective pressures with different three dimensional folds (Zhang & Gu, 1998). When two subclades of a phylogenetic tree are combined without a significant jump in the  $\alpha$  value of the distribution of substitution rates across sites, as in the approaches of Gu and Gaucher *et al.* it can be taken to indicate that the clade is evolving under stationary selective pressures across sites (see Gu, 1999, 2001; Gaucher *et al.*, 2001; for a detailed descrip-

tion of these methods). In the present study, residues were divided into the pi and pv categories. The pv sites are subject to positive (and negative) selective pressures and  $K_a/K_s$  was calculated using only these residues in the approach presented here. This approach (illustrated in Fig. 1) is demonstrated here on a simulated and a real dataset and compared with the whole gene approach and the windowing approach.



Calculate  $\alpha$  for two nodes ( $\alpha_1, \alpha_2$ ) of a phylogenetic tree and the node that joins them ( $\alpha_3$ )



Evidence for divergent evolution between clade 1 and clade 2

Divide sequences into pv and pi sites under  $\alpha_3$  and calculate  $K_a/K_s$  through the tree using pv sites only to search for positive selection

**Fig. 1** An overview of the use of covarion and  $K_a/K_s$  methods to detect branches where adaptive evolution has taken place is depicted. This enables detection of such events when selective pressures on sites are nonstationary (where  $\alpha$  describing the distribution of substitution rates across sites increases significantly) using the previously reported methods of Gu (1999, 2001) and Gaucher *et al.* (2001) and when they are stationary using the novel covarion-based approach for calculating  $K_a/K_s$  described here.

## Methods

### Programs

All programs were written in the Darwin programming language (Gonnet *et al.*, 2000). Darwin language programs for calculating  $K_a/K_s$  ratios using standard methods have previously been reported (Benner *et al.*, 2000; Liberles, 2001). A program to divide a gene family into pi and pv sites and output Darwin array, entitled covarion\_mask.drw, can be downloaded from <http://www.sbc.su.se/~jessica/programs/>.

### Generating simulated datasets

Simulated data was generated in a phylogenetic context over a symmetrical binary tree (where all branch lengths are equal and no genes have been lost) of 32 sequences (five rounds of speciation). Genes (10 genes of 100 codons) were subjected to 10% mutation (10% of sites were mutated; some were fixed and some were lost) along each branch as a discrete process (multiple substitutions not allowed), where the selective pressure at each site along each branch was individually determined. Thirty-five percent of sites were fixed as absolutely invariable, as restricted by protein fold or function. A Poisson distribution of selective pressures across sites was generated randomly for the remaining sites. A Poisson distribution of branch-specific selective pressures to simulate a punctuated equilibrium model of evolution (Gould & Eldredge, 1993) was also generated randomly. At each site, the selective pressure was determined by multiplying the branch-specific selective pressure by the site-specific selective pressure, and random mutations were either fixed or eliminated based upon this combined selective pressure.

### Examining the leptin gene family

The leptin orthologue gene family was taken from the Master Catalogue (Benner *et al.*, 2000) and the phylogenetic tree for these species (*Homo sapiens*, *Pan troglodytes*, *Gorilla gorilla*, *Pongo pygmaeus*, *Macaca mulatta*, *Mus musculus*, *Rattus norvegicus*, *Canis familiaris*, *Sus scrofa*, *Ovis aries*, *Bos taurus*, *Sminthopsis crassicaudata*, *Gallus gallus*, and *Meleagris gallopavo*) from a standard tree as used in Liberles (2001).

### Calculation of alpha

The  $\alpha$  shape parameter of the  $\Gamma$  distribution of substitution rates across sites in the leptin genes from primates was calculated using the program Diverge by measuring the effect of incrementally including more distantly related sequences in the leptin phylogeny (Gu, 1999).

### Analysis of nucleotide substitution rates

$K_a/K_s$  was calculated using a standard method (Li *et al.*, 1985; Pamilo & Bianchi, 1993) encoded in the Darwin programming language (Liberles, 2001) drawn from ancestral sequences that were calculated using a parsimony-based approach. Whereas the method of Li was used for calculation of  $K_a/K_s$  ratios, maximum likelihood methods like that of Yang & Nielsen (2000) can also be combined with the approach presented here.

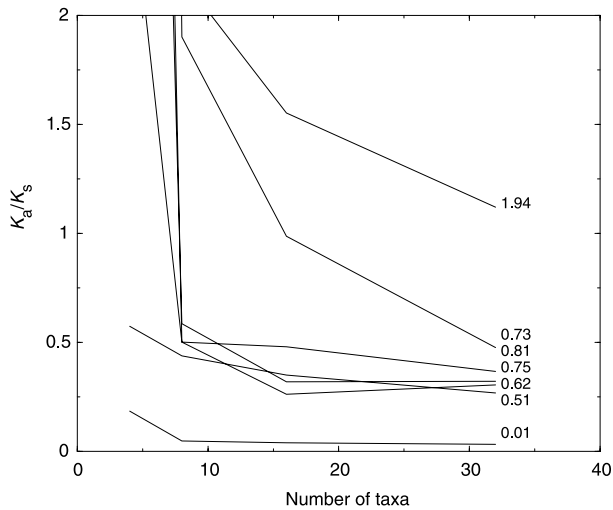
Three different approaches were taken to partition positions within a gene, whole gene averaging (no partitioning), windowing and covarion-based. Windowing recorded the highest value (with a minimum mutational threshold of 10% of residues in the window, or two mutations) for the branch from examination of 20 contiguous codons. The covarion-based approach categorized all sites from every sequence below a given specified node from the phylogenetic tree into the two categories, pi and pv.  $K_a/K_s$  was calculated considering only the pv sites. As a control, the  $K_a/K_s$  ratios of the known ancestors were also compared.

## Results

Using programs written in the Darwin programming language (Gonnet *et al.*, 2000), a simulated dataset was generated as described in the Methods section. Using a new parsimony-based approach for ancestral sequence reconstruction coupled with a standard approach for  $K_a/K_s$  ratio calculation (Liberles, 2001),  $K_a/K_s$  ratios were calculated from ancestral sequences along branches of a phylogenetic tree across whole genes, or across subsets of genes determined either by conservation across species, or by windows in the primary sequence.

When restricting measurement of substitution ratios to pv sites according to a covarion model of evolution, it is important that the subtree under examination is governed by stationary substitution trends, where the selective pressures dictating the pv and pi categories are not changing. In our simulated dataset this was fixed. The next concern is that of taxon sampling. As one narrows into an appropriate subclade, the number of sequences and their divergence decreases, and the number of pv sites is underestimated, over-weighting sites that do have changes in the calculation of a  $K_a/K_s$  ratio. This effect was observed for several branches in Fig. 2. Whereas some variation is seen between branches, most  $K_a/K_s$  values are fairly stable with clades containing eight sequences. This is of course dependent upon the divergence of the sequences, and fewer more diverse sequences will be required than the most extreme case seen in Fig. 2.

In Fig. 3, the general correlation between branch-specific selective pressure incorporated into the simulation and the measured  $K_a/K_s$  ratio is observed. The covarion-based  $K_a/K_s$  ratios show a better correlation



**Fig. 2** The effect of taxon sampling is shown. As one progresses towards the termini of a tree (going from the root through the various nodes towards the leaves) and includes fewer sequences, the effect on the measured  $K_a/K_s$  value using the covarion-based approach is demonstrated. Actual applied branch pressures are shown to the right of the curves.

with the input branch pressure than both the whole gene measured  $K_a/K_s$  and the actual  $K_a/K_s$  calculated over the whole gene from the ancestral sequences generated in the simulation. The improvement in performance can be seen in Table 1, where the covarion method identified 35% of the branches under positive selection ( $K_a > K_s$ ) compared with only 6% with the whole gene method. The percentage of those identified as under positive selection that actually were under positive selective pressure was only slightly worse (84 vs. 89%). The windowing method identified 85% of the branches under positive selective pressure, but only 56% of those identified as under positive selective pressure actually were.

However, different synthetic datasets will show different behaviours and the ratio of false positives to true positives may change in other models. For example, different protein families show different distributions of mutation rates and different models may be appropriate to model these differences of behaviour (Zhang & Gu, 1998). It should also be noted as a caveat for this method that the accuracy of ancestral sequence reconstruction shows a strong branch length dependence (Koshi & Goldstein, 1996). With this in mind, it is perhaps best to analyse the performance of this approach on a real dataset as well.

Turning to a real dataset, leptin in primates (*H. sapiens*, *P. troglodytes*, *G. gorilla*, *P. pygmaeus*, and *M. mulatta*) has previously been identified as a gene under positive selective pressure (Benner *et al.*, 1998, 2000; Liberles *et al.*, 2001; Liberles, 2001) implying a change in the

function of leptin compared with that in other eutherian mammals and has been used to explain differences in the role of leptin in obesity in humans vs. mice (Benner *et al.*, 1998). A phylogenetic tree for the leptin sequences found in Master Catalogue family 9614 is shown in Fig. 4 (Benner *et al.*, 2000).

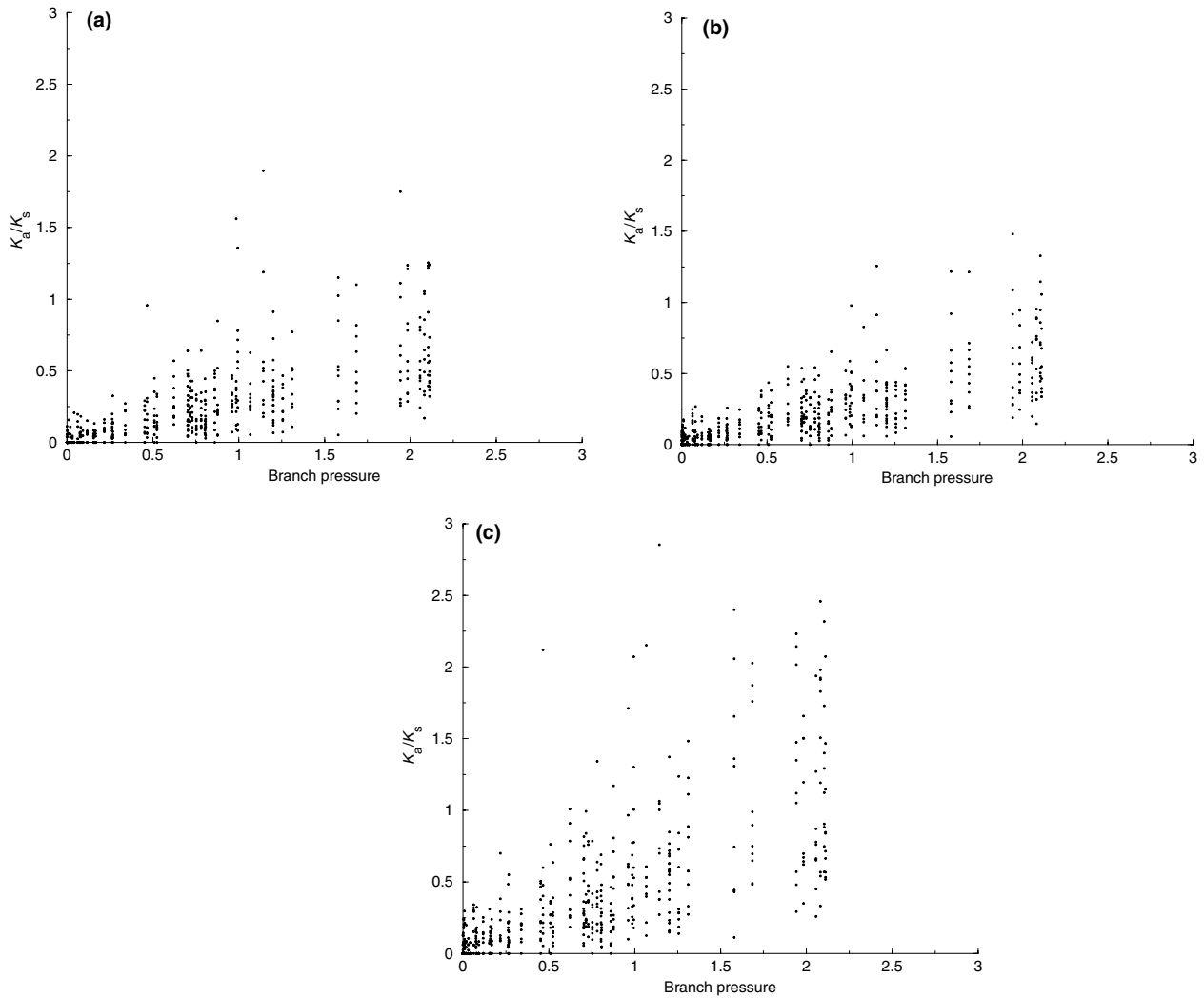
As a first control, before applying covarion-based  $K_a/K_s$  calculations, clades where substitution rates across sites are stable, where pi sites are likely to be constant, need to be identified. This was performed using the program Diverge (Gu, 1999). As can be seen in Table 2, a significant jump in  $\alpha$  values is seen both when rhesus monkey is added to the hominid clade and when nonprimate eutherians are added to the primate clade (significance can be assessed using the statistical models of, for example, Gaucher *et al.*, 2001). Otherwise, stationary substitution trends seem to operate in the tree.

In Table 3,  $K_a/K_s$  ratios for the various branches in the primate leptin tree (Fig. 4) were calculated using a whole gene analysis, an analysis of pv sites from the whole tree, and pv sites within the primate clade. The primate clade contained 21 pv sites and 124 pi sites, whereas the whole tree contained 69 pv sites and 76 pi sites. Potentially variable sites from the whole tree do not reflect stationary substitution trends and the number of pv sites may be overestimated for this reason. Because of sampling issues, as a comprehensive set of sequences across amniotes does not exist, some pv sites may not have been sampled and may have been included in the pi category. The number of such miscounted sites is expected to be small, given the large number of species compared with the results in Fig. 2. This analysis shows a new branch under positive selective pressure in the lineage leading to gorilla that was not previously detected as having been under positive selective pressure using the whole gene averaging approach.

Strictly speaking, the appropriate analysis uses the pv sites in the hominid clade only, which is the most ancient node under which stationary substitution trends operate. However, because of the limited divergence time in these species and the small number of species sampled, this analysis was not possible. Even considering the primate clade, as shown in Table 3, the data should not be considered with high confidence levels because of the possible bias towards counting sites known to have mutated along any given branch. Still, the values seen in Table 3 may indeed reflect large periods of positive selection in primates and further sequencing of this gene from other primates may illuminate the evolutionary processes occurring here.

## Discussion

A new and more sensitive approach to detect branches of a gene family tree suffering from positive selective pressures has been introduced here. First, clades within the tree where stationary substitution trends are



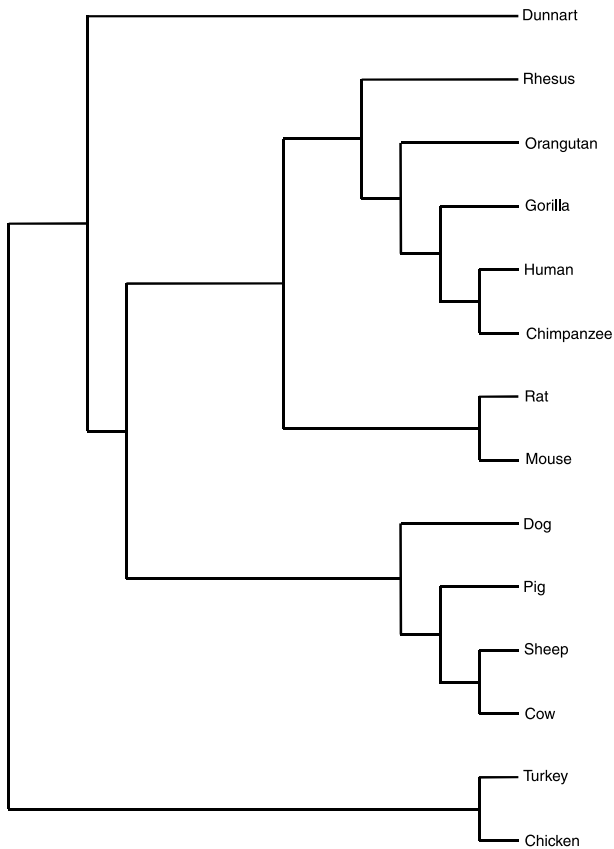
**Fig. 3** The relationship between the applied branch pressure and the measured  $K_a/K_s$  ratio is shown for all branches based upon (a)  $K_a/K_s$  calculated using the actual ancestors as whole genes (b)  $K_a/K_s$  calculation averaging over the whole gene from reconstructed ancestral sequences, and (c) using the covarion-based  $K_a/K_s$  approach from the top node of the phylogenetic tree on reconstructed ancestors.

**Table 1** From the synthetic dataset, and  $K_a/K_s > 1$  as a benchmark, the correlation between the applied branch specific pressure and the actual  $K_a/K_s$  calculated from the known ancestors as well as that measured using the whole gene averaging (wga), covarion-based and windowing methodologies is shown. True positives have a branch pressure  $> 1$  and  $K_a/K_s > 1$ , true negatives have both  $< 1$ , false positives have  $K_a/K_s > 1$  whereas the branch pressure  $< 1$ , and false negatives have  $K_a/K_s < 1$  whereas branch pressure  $> 1$ . Reported error ranges are the standard error of the mean.

Method	True positives	False negatives	True negatives	False positives
Actual	17 ± 4.5	123 ± 4.5	478 ± 1.3	2 ± 1.3
wga	8 ± 2.9	132 ± 2.9	480 ± 0.0	0 ± 0.0
Covarion	49 ± 5.0	91 ± 5.0	471 ± 2.8	9 ± 2.8
Windowing	119 ± 2.8	21 ± 2.8	385 ± 8.2	95 ± 8.2

operating are identified. From this, using a covarion-based approach, sites that are subject to molecular adaptation are counted, whereas those that are absolutely conserved are excluded. This approach appears to be more sensitive to detecting positive selection in a simulated dataset when compared with whole gene averaging. Although not as sensitive as the windowing approach of Endo *et al.* (1996), it does not identify the large number of false positives identified through the windowing approach.

This approach has been extended to the leptin dataset, a gene of potential pharmaceutical interest for its role in obesity. Previously, three branches in the leptin orthologue gene tree with high  $K_a/K_s$  values were identified (see Liberles, 2001). Using a conservative



**Fig. 4** A phylogenetic tree for the leptin gene is shown. It was adapted in Liberles (2001) from Master Catalogue family 9614 (Benner *et al.*, 2000). The hominid species are human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), and orangutan (*Pongo pygmaeus*). The additional primate is rhesus monkey (*Macaca mulatta*). The additional eutherians are mouse (*Mus musculus*), rat (*Rattus norvegicus*), dog (*Canis familiaris*), pig (*Sus scrofa*), sheep (*Ovis aries*), and cow (*Bos taurus*). The additional mammal is dunnart (*Sminthopsis crassicaudata*). The additional amniotes are chicken (*Gallus gallus*) and turkey (*Meleagris gallopavo*).

covariation-based approach, a branch where positive selection was not previously detected was now identified as a candidate for positive selection, that leading to

**Table 2**  $\alpha$  Values of the  $\Gamma$  distribution of amino acid substitution rates were calculated using the program Diverge (Gu, 1999) over the tree in Fig. 3. Different clades were analysed and the  $\alpha$  value reflects changes as additional sequences were included.

Clade analysed	$\alpha$ -Value
Hominids	0.06
Primates	0.54
Eutherians	0.94
Mammals	0.83
Amniotes	0.81

gorilla. It may be possible that leptin is undergoing substantial evolution throughout the primate lineage. This may correlate with changes of gene expression in leptin in primates, where the insertion of a MER11 transposable element into an enhancer region has altered the expression level and tissue distribution in humans and other primates since the last common ancestor with mouse (Bi *et al.*, 1997). Further, a retroviral long-terminal repeat may also have inserted into the leptin receptor gene in the primate lineage (Kapitonov & Jurka, 1999), and there may be evidence for further transposition around the leptin and leptin receptor genes in other lineages.

One possible model for the ecological basis for such rapid evolution in leptin may be dietary changes in the evolution of primates from their last common ancestor with other eutherian mammals. Another possible model is a divergent role in pregnancy, correlating with differences between any of the several aspects of the reproductive cycle between primates and other eutherian mammals. A fuller picture of the evolution of leptin in primates will emerge as the gene is sequenced from additional primate species and its precise function is better defined physiologically.

## Acknowledgments

This research was supported by a centre grant from the Swedish Foundation for Strategic Research. We are grateful to Jens Lagergren, Bengt Sennblad and Alexander Roth for helpful discussions.

**Table 3**  $K_a/K_s$  values for the various branches of the primate leptin tree (Fig. 3) are calculated using whole gene averaging (wga), the covariation-based approach taken from the amniote node at the top of the tree (amn), and the covariation-based approach taken from the most ancient primate node (pri). Values that are  $>1$  are shown in bold.  $K_a/K_s$  1 refers to the value along branch 1, whereas  $K_a/K_s$  2 refers to the value along branch 2. It should be noted that the low degree of variation and small number of sequences in the primate lineage generates large amounts of uncertainty for that dataset. Hgc refers to human, gorilla and chimpanzee.

Branch 1	Branch 2	wga $K_a/K_s$ 1	wga $K_a/K_s$ 2	amn $K_a/K_s$ 1	amn $K_a/K_s$ 2	pri $K_a/K_s$ 1	pri $K_a/K_s$ 2
Rhesus	Hominids	<b>1.25</b>	<b>1.09</b>	<b>4.32</b>	<b>1.52</b>	<b>22.66</b>	<b>2.76</b>
Orangutan	hgc	0.36	<b>1.02</b>	0.74	<b>2.71</b>	<b>2.35</b>	<b>9.00</b>
Gorilla	hc	0.55	0.18	<b>1.11</b>	0.37	<b>3.55</b>	<b>1.16</b>
Human	Chimp.	0.14	0.15	0.26	0.37	0.21	<b>1.16</b>

## References

- Benner, S.A., Chamberlin, S.G., Liberles, D.A., Govindarajan, S. & Knecht, L. 2000. Functional inferences from reconstructed evolutionary biology involving rectified databases – an evolutionarily grounded approach to functional genomics. *Res. Microbiol.* **151**: 97–106.
- Benner, S.A., Trabesinger, N. & Schreiber, D. 1998. Post genomic science: converting primary sequence into physiological function. *Adv. Enzyme Regul.* **38**: 155–180.
- Bi, S., Gavrilova, O., Gong, D.W., Mason, M.M. & Reitman, M. 1997. Identification of a placental enhancer for the human leptin gene. *J. Biol. Chem.* **272**: 30 583–30 588.
- Endo, T., Ikeo, K. & Gojobori, T. 1996. Large scale search for genes on which positive selection may operate. *Mol. Biol. Evol.* **13**: 685–690.
- Fitch, W.M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.* **20**: 406–416.
- Gaucher, E.A., Miyamoto, M.M. & Benner, S.A. 2001. Function-structure analysis of proteins using covarion-based evolutionary approaches: elongation factors. *Proc. Natl. Acad. Sci., USA* **98**: 548–552.
- Gonnet, G.H., Hallett, M.T., Korostensky, C. & Bernardin, L. 2000. Darwin v. 2.0: an interpreted computer language for the biosciences. *Bioinformatics* **16**: 101–103.
- Gould, S.J. & Eldredge, N. 1993. Punctuated equilibrium comes of age. *Nature* **366**: 223–227.
- Gu, X. 1999. Statistical methods for testing functional divergence after duplication. *Mol. Biol. Evol.* **16**: 1664–1674.
- Gu, X. 2001. Maximum-likelihood approach for gene family evolution under functional divergence. *Mol. Biol. Evol.* **18**: 453–464.
- Kapitonov, V.V. & Jurka, J. 1999. The long terminal repeat of an endogenous retrovirus induces alternative splicing and encodes an additional carboxy-terminal sequence in the human leptin receptor. *J. Mol. Evol.* **48**: 248–251.
- Koshi, J.M. & Goldstein, R.A. 1996. Probabilistic reconstruction of ancestral protein sequences. *J. Mol. Evol.* **42**: 313–320.
- Li, W.H., Wu, C.I. & Luo, C.C. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2**: 150–174.
- Liberles, D.A. 2001. Evaluation of methods for determination of a reconstructed history of gene sequence evolution. *Mol. Biol. Evol.* **18**: 2040–2047.
- Liberles, D.A., Schreiber, D.R., Govindarajan, S., Chamberlin, S.G. & Benner, S.A. 2001. The adaptive evolution database (TAED). *Genome Biol.* **2**: research0028.1–research0028.6.
- Miyamoto, M.M. & Fitch, W.M. 1995. Testing the covarion hypothesis of molecular evolution. *Mol. Biol. Evol.* **12**: 503–513.
- Pamilo, P. & Bianchi, N.O. 1993. Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between genes. *Mol. Biol. Evol.* **10**: 271–281.
- Yang, Z. & Bielawski, J.P. 2000. Statistical methods for detecting adaptive evolution. *TREE* **15**: 496–502.
- Yang, Z., Kumar, S. & Nei, M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**: 1641–1650.
- Yang, Z. & Nielsen, R. 2000. Estimating synonymous and nonsynonymous nucleotide substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**: 32–43.
- Zhang, J. & Gu, X. 1998. Correlation between the substitution rate and rate variation among sites in protein evolution. *Genetics* **149**: 1615–1625.
- Zhang, J., Rosenberg, H.F. & Nei, M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci., USA* **95**: 3708–3713.

Received 7 January 2002; revised 6 March 2002; accepted 9 March 2002