LETTERS

The "Inverse Relationship Between Evolutionary Rate and Age of Mammalian Genes" Is an Artifact of Increased Genetic Distance with Rate of Evolution and Time of Divergence

Eran Elhaik,[1] *Niv Sabath,*[1] *and Dan Graur*

Department of Biology and Biochemistry, University of Houston

It has recently been claimed that older genes tend to evolve more slowly than newer ones (Alba and Castresana 2005). By simulation of genes of equal age, we show that the inverse correlation between age and rate is an artifact caused by our inability to detect homology when evolutionary distances are large. Since evolutionary distance increases with time of divergence and rate of evolution, homologs of fast-evolving genes are frequently undetected in distantly related taxa and are, hence, misclassified as "new." This misclassification causes the mean genetic distance of 'new' genes to be overestimated and the mean genetic distance of "old" genes to be underestimated.

Introduction

Alba and Castresana (2005) have recently reported a negative correlation between the "age" of genes and the rate of evolution. They proposed two alternative explanations for this relationship. The first was that functional constraint remained constant throughout the evolutionary history of each gene, but that newer genes are less constrained than older genes. The second explanation invokes a scenario whereby functional constraints are not constant, rather they are weak at the time of origin of a gene and they become progressively more stringent with age.

In the study of Alba and Castresana (2005), the rate of evolution was calculated from human-mouse "orthologous" gene pairs. Orthology was defined operationally, rather than evolutionarily, through reciprocal BlastP (Altschul et al. 1997) hits. For each pair of human-mouse genes, the number of nonsynonymous substitutions per nonsynonymous site (K_A) was calculated. Alba and Castresana (2005) determined the age of each human-mouse gene pair by the phylogenetic distribution of their homologs among the genomes of six model organisms: *Takifugu rubripes*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*. If homologous genes were present in all these six genomes, the human-mouse gene pair was assigned to the OLD group. If homologous genes were found in *C. elegans*, *D. melanogaster*, and *T. rubripes* but absent from *S. cerevisiae*, *S. pombe*, and *A. thaliana*, then the human-mouse pair was classified in the METAZOANS group. If homologous genes were present in *T. rubripes* but absent from the other five genomes, then the pair was classified in the DEUTEROSTOMES group. If homologous genes were absent from all six genomes, then the pair was classified in the TETRAPODS group. A BlastP hit with an expectation value of less than 10^{-4} was deemed to be indicative of "presence." A negative correlation was found between the rate of substitution and age. The inferred K_A values were 0.06, 0.08, 0.14, and 0.23 for OLD, METAZOANS, DEUTEROSTOMES, and TETRAPODS, respectively.

In this note, we show by simulation that the inverse relationship between evolutionary rate and gene age is an artifact caused by our inability to detect similarity when genetic distances are large.

Methods

The DNA Assembly with Gaps (DAWG) simulation program (Cartwright 2005) was used to generate terminal sequences whose phylogenetic relationships are shown in figure 1. In a manner analogous to Alba and Castresana (2005), A and B may be regarded as the human and mouse orthologous genes, while C, D, and E represent homologous genes from increasingly more distant taxa. All the genes originated in the common ancestor of A, B, C, D, and E and are, thus, of equal age. We used Kimura's two parameters model (Kimura 1980) with a ratio of transitions to transversions of 1.2. To simulate different evolutionary rates, the ratios of the branch lengths to one another were kept constant, and in each simulation we multiplied the branch lengths by an evolutionary constant that ranged from 1 to 5 in 0.04 intervals (for a total of 101 different rates). For each constant, we simulated 50 sets of DNA sequences, each 1,000 nt long. The simplest distance measure between genes A and B (genetic distances) was calculated with ClustalW (Thompson, Higgins, and Gibson 1994). We used Blast 2 sequences (Tatiana and Madden 1999) to detect homology between gene A and genes C, D, and E.

Presence or absence of a homolog in C, D, or E was determined by Blast 2 sequences. If the E value was less than 10^{-4}, a homolog was assumed to be present; otherwise, we inferred absence. Similar results were obtained with cutoffs that varied from 10^{-2} to 10^{-20}. All cutoffs are, of course, arbitrary.

In a manner analogous to OLD, METAZOANS, DEUTEROSTOMES, and TETRAPODS of Alba and Castresana (2005), each simulated A-B pair was assigned to one of four age groups: SENIORS, ADULTS, TEENAGERS, and TODDLERS.

[1] These authors contributed equally to this work and are to be considered co-first authors.

Key words: nonsynonymous substitutions, novel genes, divergence times.

E-mail: dgraur@uh.edu.

Mol. Biol. Evol. 23(1):1–3. 2006
doi:10.1093/molbev/msj006
Advance Access publication September 8, 2005

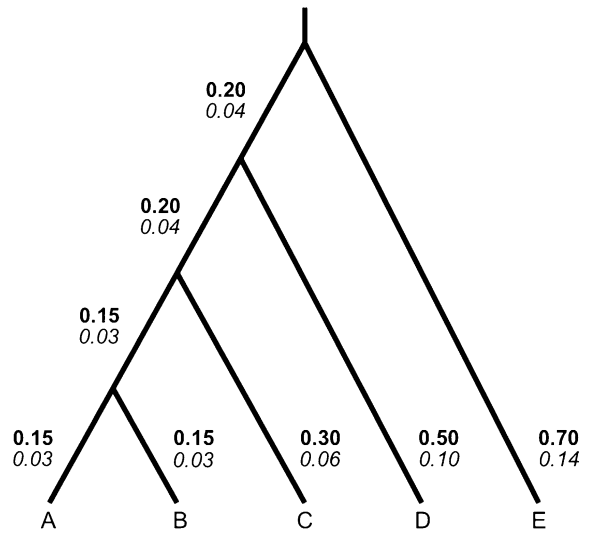

FIG. 1.—A rooted phylogenetic tree used in the simulation. The ratios of the branch lengths to one another were kept constant, and in each simulation we multiplied the branch lengths by an evolutionary constant that ranged from 1 to 5 in 0.04 intervals (for a total of 100 different rates). For each branch, we show the smallest (italic) and the largest (bold) branch lengths.

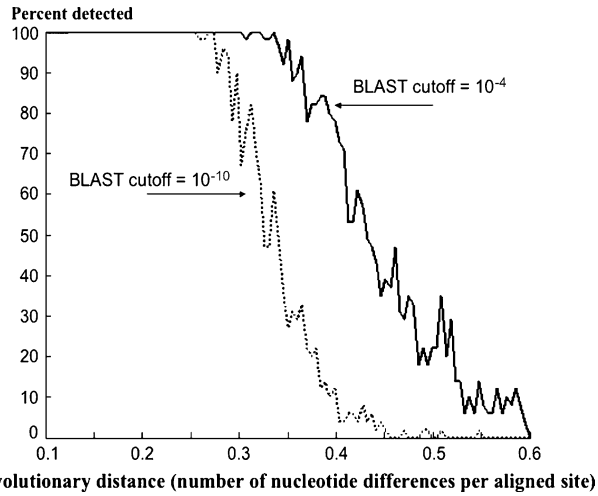

FIG. 3.—Fraction of genes in lineage C with homology to a gene in lineage A that are detectable by Blast as a function of the evolutionary distance between A and C. Two cutoff values for Blast are shown.

Results and Discussion

The distribution of the genetic distances and, by implication, the evolutionary rates are shown in figure 2. The SENIORS category consists of 753 sequence pairs. The corresponding values for ADULTS, TEENAGERS, and TODDLERS were 803, 1,766, and 1,592, respectively. The mean genetic distance for SENIORS was 0.08, and the corresponding values for ADULTS, TEENAGERS, and TODDLERS were 0.11, 0.17, and 0.22, respectively.

Alba and Castresana (2005) realized that the use of a sequence-similarity detection method for the identification of homologs may be problematic as far as fast-evolving genes are concerned. However, they claimed that a local alignment search tool, such as Blast, will be able to detect homology of fast-evolving genes by identifying evolutionarily conserved residues. To support this claim, they showed that the majority of fast-evolving genes in all genomes were found with a lower cutoff of 10^{-10}. This "control" analysis is misleading in that Alba and Castresana (2005) could only count genes that were identified as homologous by their protocol. They may have, thus, failed to spot the vast majority of homologs from among the fastest evolving genes.

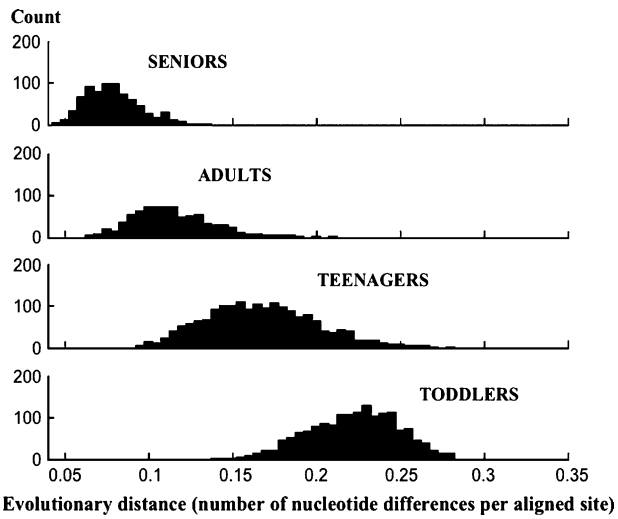

FIG. 2.—Observed distribution of evolutionary distances in four inferred age groups. Because all genes in the simulation are of equal age, the division into four age groups is an artifact of the inference protocol. The evolutionary distance was calculated between lineages A and B in figure 1.

Our simulation clearly demonstrates the inherent inability to detect homology of the fastest evolving genes (fig. 3). One can clearly see that detectability decreases rapidly with evolutionary distance. In other words, as far as the fastest evolving genes are concerned, the vast majority of them are undetectable even when the cutoffs are extremely permissive.

One explanation given by Alba and Castresana (2005) for the formation of "young" genes is that some genes have originated *de novo*. They supported this claim by noting that the genes deemed by their method to be young were shorter on average than the genes deemed by their method to be old. We note that there is yet no known mechanism for the formation of *de novo* genes. Thus, the smaller mean gene length may be an artifact of Blast, which is a local alignment search tool.

Because all our simulated genes have the same evolutionary age and because our results are similar to those obtained by Alba and Castresana (2005), we conclude that the inverse relationship between evolutionary rate and gene age is an artifact caused by our inability to detect similarity when genetic distances are large. Since genetic distance increases with time of divergence and rate of evolution, it is difficult to identify homologs of fast-evolving genes in distantly related taxa. Thus, fast-evolving genes may be misclassified as new. The only conclusion that can be

drawn from the study of Alba and Castresana (2005) is that slowly evolving genes evolve slowly.

Acknowledgments

We wish to thank Reed A. Cartwight for his help with DAWG and Jose Castresana for providing us access to the raw data.

Literature Cited

Alba, M. M., and J. Castresana. 2005. Inverse relationship between evolutionary rate and age of mammalian genes. Mol. Biol. Evol. **22**:598–606.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped Blast and PSI-Blast: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

Cartwight, R. A. 2005. DAWG: DNA Assembly with Gaps. (http://scit.us/projects/dawg/).

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

Tatiana, A. T., and T. L. Madden. 1999. Blast 2 Sequences—a new tool for comparing protein and nucleotide sequences. FEMS Microbiol. Lett. **174**:247–250.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Manolo Gouy, Associate Editor

Accepted August 30, 2005