Estimates of Positive Darwinian Selection Are Inflated by Errors in Sequencing, Annotation, and Alignment

Adrian Schneider,* Alexander Souvorov,† Niv Sabath,‡ Giddy Landan,‡ Gaston H. Gonnet,* and Dan Graur‡

*ETH Zürich, Zürich, Switzerland; †National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland; and ‡Department of Biology and Biochemistry, University of Houston

Published estimates of the proportion of positively selected genes (PSGs) in human vary over three orders of magnitude. In mammals, estimates of the proportion of PSGs cover an even wider range of values. We used 2,980 orthologous protein-coding genes from human, chimpanzee, macaque, dog, cow, rat, and mouse as well as an established phylogenetic topology to infer the fraction of PSGs in all seven terminal branches. The inferred fraction of PSGs ranged from 0.9% in human through 17.5% in macaque to 23.3% in dog. We found three factors that influence the fraction of genes that exhibit telltale signs of positive selection: the quality of the sequence, the degree of misannotation, and ambiguities in the multiple sequence alignment. The inferred fraction of PSGs in sequences that are deficient in all three criteria of coverage, annotation, and alignment is 7.2 times higher than that in genes with high trace sequencing coverage, “known” annotation status, and perfect alignment scores. We conclude that some estimates on the prevalence of positive Darwinian selection in the literature may be inflated and should be treated with caution.

Introduction

A general consensus in molecular evolution is that the vast majority of protein-coding genes evolve by random genetic drift. It is also a common lore that genes exhibiting signs of positive Darwinian selection are the “interesting” ones. Much of the recent molecular evolution literature deals with detecting positive selection on a genome-wide scale, sometimes as a tool for gaining insight into function. A wide variety of methods for identifying positively selected genes (PSGs) have been introduced. One of the most commonly used methods is the nonsynonymous to synonymous ratio test pioneered by Li and Gojobori (1983) and Hill and Hastie (1987) and explicitly put forward by Hughes and Nei (1988).

Nonsynonymous changes are far more likely than synonymous changes to improve the function of a protein, that is, to be advantageous. Because advantageous mutations undergo fixation in a population much more rapidly than neutral mutations and because most synonymous substitutions may be assumed to be neutral, the rate of nonsynonymous substitution should exceed that of synonymous substitution if positive Darwinian selection plays a major role in the evolution of a protein-coding gene. Therefore, an excess of nonsynonymous substitutions over synonymous ones is frequently used as an indicator of positive selection at the molecular level.

By using the ratio of nonsynonymous to synonymous substitution (dN/dS), it has been shown that the proportion of PSGs varies widely among and within taxonomic lineages. However, the results vary widely depending on taxonomic sampling, sequence sampling, quality filtering, and estimation model. For example, the proportion of PSGs in human was estimated to be 0.02% (Gibbs et al. 2007), 0.07% (Kosiol et al. 2008), 1.1% (Bakewell et al. 2007), 1.5% (Arbiza et al. 2006), 5.2% (Jorgensen et al. 2005), and 8.7% (Clark et al. 2003): a 435-fold range. Some such results were extremely surprising in other respects. In one such study, positive selection was detected in 77% of vertebrate genes (Studer et al. 2008). In another, the claim was made that the fraction of PSGs in chimpanzees is almost twice that in humans (Bakewell et al. 2007).

Because errors in sequencing, annotation, and alignment are unaffected by codon position and subsequent effect on the amino acid sequence, they are equally likely to affect synonymous as nonsynonymous sites. Because most nonsynonymous mutations are selected against, the vast majority of genes exhibit ratios of nonsynonymous to synonymous substitution that are lower than 1. On the other hand, errors that equally affect synonymous and nonsynonymous sites add a noise component to $dN/dS = 1$. Although low levels of such “neutral” noise will not generally raise the overall ratio above 1, estimation models that allow for several categories of sites may detect the noise component as a separate category. Consequently, such models will inflate the fraction of inferred PSGs within a genome.

Here, we investigate quantitatively the effects of errors in sequencing, annotation, and alignment on inferences of positive selection in genomic studies.

Data and Methods

Sequence Data

Genomic data from seven eutherian mammalian species were downloaded from Ensembl (Hubbard et al. 2007) and converted to the Darwin database format (Gonnet, Hallett, et al. 2000). The National Center for Biotechnology Information (NCBI) version numbers were 36 for Homo sapiens, NCBI m36 for Mus musculus, PanTro2.1 for Pan troglodytes, Mmul 1.0 for Macaca mulatta, RGSC 3.4 for Rattus norvegicus, CanFam 2.0 for Canis familiaris, and Btau 3.0 for Bos taurus. The reason for restricting the data to these species is that their relative phylogenetic position in the eutherian tree is unambiguous (Murphy et al. 2001). Adding more eutherian or noneutherian mammals would have been phylogenetically contentious (e.g., Cannarozzi et al. 2007).

Orthologous Protein-Coding Sequences

Orthologous sequences were taken from the mammalian data set in Orthologs Matrix project (OMA; Dessimoz...
et al. 2005; Schneider et al. 2007). Putative orthologous proteins for the seven species under consideration were found in 9,942 OMA groups. Of these, 1,227 contained a coding sequence (CDS) of less than 200 codons and in 5,089 groups, the longest sequence was more than 5% longer than the shortest one. Given the stochastic errors associated with short sequences and the uncertainties of alignment due to indels, we decided not to use these sequences. (These groups may also possibly contain fragments of actual proteins and were, therefore, excluded from further analysis.) Additionally, six groups were excluded because at least one of the CDSs had more than 1% of its bases listed as missing or unknown. The proteins in each group were aligned using the Darwin multiple sequence alignment package (Gonnet, Hallett, et al. 2000; Gonnet, Korostensky, and Benner 2000). The coding DNA was, then, mapped to the aligned amino acids in order to construct codon-wise alignments of the CDSs.

Identification of Positive Selection

The CODEML program from version 3.14b of PAML (Yang 1997) was used to estimate branch lengths and nonsynonymous (dN) and synonymous (dS) rates for each branch of the tree. All analysis was performed on the tree shown in figure 1. The codon frequencies were derived from the average nucleotide frequencies at the three-codon position (F3 × 4 model), and the model was chosen to compute one ratio of nonsynonymous over synonymous substitution (dN/dS) per branch but fixed for all sites (model = 1 and NSsites = 0). In 758 cases (orthologous groups), at least one branch had a length of 0. Such sequences are not suitable for ratio calculations and, hence, were excluded, leaving us with a final data set of 2,980 orthologous trees on which the subsequent analyses were performed. The multiply aligned data are available as Supplementary Material (see Supplementary Material online for alignments).

PSGs were then identified using the improved branch-site model as implemented in CODEML. We performed “test 2” from Zhang et al. (2005) under “model A,” where for the null hypothesis it is assumed that some sites are selected (ω < 1) and some are neutral (ω = 1), whereas for the alternative hypothesis, some sites are allowed to experience positive selection (ω > 1) on one or more a priori–specified “foreground branches.” Each of the seven terminal branches was set up as foreground separately, whereas all other branches were designated as background at that time. Following the recommendations from PAML, CODEML was run three times for the null hypothesis and three times for each alternative with the initial branches set to a random value between 0.8 and 1.2 times the previously estimated branch lengths, allowing for three randomized runs of the optimization process, which increases the chances of reaching the global optimum. The highest likelihood for each alternative was saved for the likelihood ratio test (LRT). Finally, PSGs were identified by comparing the LRT values to the distribution. Rom (1990) correction for multiple testing and a false discovery rate of 5% were applied. Rom (1990) correction has been recently shown to be the very accurate (Anisimova and Yang 2007), but other corrections yielded essentially the same results (data not shown).

Quality of Sequences

The quality of each CDS was assessed by the degree of coverage in the trace database (Trace Archive, http://www.ncbi.nlm.nih.gov/Traces/home/). Trace data for the human genome have unfortunately not been kept. Each CDS was Blasted against its appropriate trace archive. Only hits with identities of 95% or higher were considered. We experimented with thresholds other than 95%, but the
Quality of the Alignment

The quality of the multiple sequence alignment was determined with the Heads or Tails (HoT) algorithm (Landan and Graur 2007). Briefly, this methodology is based upon the a priori expectation that sequence alignment results should be independent of the orientation of the input sequences. Thus, for totally unambiguous cases, reversing residue order prior to alignment should yield an exact reversed alignment of that obtained by using the unreversed sequences. The degree of agreement between these two alignments may be used to assess the reliability of the sequence alignment. In this study, we divided the sequences into unambiguous (100% agreement between the head and tail alignments) and ambiguous alignments.

Degree of Certainty in Gene Annotation

We used the gene annotation status in Ensembl to divide the genes into two categories: known genes and novel genes. “Ensembl known genes” are predicted on the basis of species-specific database entries from manually curated UniProt/Swiss-Prot, partially manually curated RefSeq, and UniProt/TrEMBL databases. Predictions of “Ensembl novel genes” are based on other experimental evidence such as protein and cDNA sequence information from related species (http://www.ensembl.org/Homo_sapiens/glossaryview). We, therefore, have more confidence in the veracity of the annotation of the “known” category than that in the “novel” (or “inferred”) category.

Results

We first used the trace database to divide the 17,880 genes from six species (trace data for the human genome are not available) into ones that have poor sequences regions (less than three times coverage) and genes in which all parts have been sequenced at least three times. The results (not shown) did not affect the conclusions. For each position in the CDS, the number of hits overlapping this position were counted and used as a measure of sequencing coverage. We divided the sequences into two categories: those in which parts were covered by less than three trace sequences and those in which the whole sequence was covered at least three times.

Table 1

Inferred Percentage of PSGs as a Function of Sequencing Coverage

| Coverage ≥3× | Coverage <3× |
|--------------|--------------|
| Total | PSG % PSG | Total | PSG % PSG | $P(\chi^2)$ |
| Human | N/A | N/A | N/A | N/A | N/A |
| Chimp | 1,144 | 9 | 0.8 | 1,836 | 74 | 4.0 | $9.6 \times 10^{-8}$ |
| Macaque | 896 | 32 | 3.6 | 2,084 | 488 | 23.4 | $8.1 \times 10^{-46}$ |
| Mouse | 2,493 | 77 | 3.1 | 487 | 37 | 7.6 | $1.3 \times 10^{-6}$ |
| Rat | 1,841 | 93 | 5.1 | 1,139 | 217 | 19.1 | $1.3 \times 10^{-6}$ |
| Dog | 1,568 | 212 | 13.5 | 1,412 | 481 | 34.1 | $1.7 \times 10^{-49}$ |
| Cow | 1,086 | 54 | 5.0 | 1,894 | 249 | 13.1 | $1.7 \times 10^{-49}$ |
| Total | 9,028 | 477 | 5.3 | 8,852 | 1,546 | 17.5 | $2.3 \times 10^{-106}$ |

*Note:* N/A, Not available.

Table 2

Inferred Percentage of PSGs as a Function of Annotation Status

| Known Genes | Inferred Genes |
|-------------|----------------|
| Total | PSG % PSG | Total | PSG % PSG | $P(\chi^2)$ |
| Human | 2,833 | 21 | 0.7 | 95 | 5 | 5.3 | $3.5 \times 10^{-8}$ |
| Chimp | 318 | 17 | 5.3 | 2,653 | 66 | 2.5 | 0.0031 |
| Macaque | 139 | 13 | 9.4 | 2,840 | 507 | 17.9 | 0.0082 |
| Mouse | 2,924 | 108 | 3.7 | 56 | 6 | 10.7 | 0.0063 |
| Rat | 2,696 | 224 | 8.3 | 284 | 86 | 30.3 | $8.9 \times 10^{-43}$ |
| Dog | 2,568 | 491 | 19.1 | 412 | 202 | 49.0 | $7.6 \times 10^{-46}$ |
| Cow | 2,670 | 206 | 7.7 | 309 | 97 | 31.4 | $1.8 \times 10^{-41}$ |
| Total | 14,148 | 1,080 | 7.6 | 6,649 | 969 | 14.6 | $1.4 \times 10^{-6}$ |

*The sums of the two total numbers of genes in each species do not always add up to 2,980 because some genes lack annotation status in the database.*

Table 3

Inferred Percentage of PSGs as a Function of Alignment Quality (HoT)

| HoT Score = 100% | HoT Score < 100% |
|------------------|------------------|
| Total | PSG % PSG | Total | PSG % PSG | $P(\chi^2)$ |
| Human | 2,805 | 23 | 0.8 | 175 | 3 | 1.7 | 0.22 |
| Chimp | 2,805 | 78 | 2.8 | 175 | 5 | 2.9 | 0.95 |
| Macaque | 2,805 | 468 | 16.7 | 175 | 52 | 29.7 | $5.6 \times 10^{-6}$ |
| Mouse | 2,805 | 99 | 3.5 | 175 | 15 | 8.6 | $6.3 \times 10^{-4}$ |
| Rat | 2,805 | 284 | 10.1 | 175 | 26 | 14.9 | 0.042 |
| Dog | 2,805 | 630 | 22.5 | 175 | 63 | 36.0 | $2.1 \times 10^{-3}$ |
| Cow | 2,805 | 276 | 9.8 | 175 | 27 | 15.4 | $2.0 \times 10^{-3}$ |
| Total | 19,635 | 1,858 | 9.5 | 1,225 | 191 | 15.6 | $8.2 \times 10^{-13}$ |

*Note:* The quality is determined for the whole multiple sequence alignment, so the numbers of unambiguous and ambiguous alignments is the same for all species.
Discussion

Our results indicate that the worse the quality of the sequence is, the greater the proportion of PSGs will appear to be. We note that recent developments in technology allow for genomes to be cheaply and rapidly sequenced and that assembly programs are now able to infer contigs from partial trace sequences with little overlap between neighboring ones. As a result, although the quantity of genomic sequences accumulating in databanks is growing exponentially, the quality of the sequences may be concomitantly deteriorating.

We have also shown that uncertainties in alignment and annotation contribute to overestimating the ratio of nonsynonymous to synonymous substitution rates and, consequently, to overestimating the fraction of PSGs. Problems in sequence quality, annotation, and alignment are not independent factors as shown by the fact that their effects are not cumulative. However, they all work in the same direction, inflating estimates of positive selection. Other factors, such as purifying selection on synonymous sites or the quality of the sequences may be concomitantly deteriorating.

We conclude that estimates on the prevalence of positive Darwinian selection may be inflated and should be treated with caution.

Funding

This work was supported by National Science Foundation grant [DBI-0543342 to D.G.].

Supplementary Material

Supplementary Alignments are available at Genome Biology and Evolution online (http://www.oxfordjournals.org/our_journals/gbe/).

Acknowledgments

We thank Maria Anisimova for advice on multiple testing corrections and Gina Cannarozzi for helpful discussions.

Literature Cited

Anisimova M, Yang Z. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol. 24:1219–1228.

Arbizu L, Dopazo J, Dopazo H. 2006. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. PLoS Comput Biol. 2:e38.

Bakewell MA, Shi P, Zhang J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc Natl Acad Sci USA. 104:7489–7494.

Cannarozzi G, Schneider A, Gonnet G. 2007. A phylogenomic study of human, dog, and mouse. PLoS Comput Biol. 3:e2.

Clark AG, et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science. 302:1960–1963.

Dessimoz C, et al. 2005. OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: introduction and first achievements. In: McLysath A, Huson D, editors. Lecture notes in computer science. Berlin (Germany): Springer-Verlag. p. 61–72.

Gibbs RA, et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science. 316:222–234.

Gonnet GH, Hallett MT, Korostensky C, Bernardin L. 2000. Darwin v. 2.0: an interpreted computer language for the biosciences. Bioinformatics. 16:101–103.

Gonnet GH, Korostensky C, Benner S. 2000. Evaluation measures of multiple sequence alignments. J Comput Biol. 7:261–276.

Hill RE, Hastie ND. 1987. Accelerated evolution in the reactive centre regions of serine protease inhibitors. Nature. 326:96–99.
Hubbard TJ, et al. 2007. Ensembl 2007. Nucleic Acids Res. 35:D610–D617.
Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals over-dominant selection. Nature. 335:167–170.
Jorgensen FG, et al. 2005. Comparative analysis of protein coding sequences from human, mouse and the domesticated pig. BMC Biol. 3:2.
Kosiol C, et al. 2008. Patterns of positive selection in six mammalian genomes. PLoS Genet. 4:e1000144.
Landan G, Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol. 24:1380–1383.
Li W-H, Gojobori T. 1983. Rapid evolution of goat and sheep globin genes following gene duplication. Mol Biol Evol. 1:94–108.
Murphy WJ, et al. 2001. Molecular phylogenetics and the origins of placental mammals. Nature. 409:614–618.
Rom DM. 1990. A sequentially rejective test procedure based on a modified Bonferroni inequality. Biometrika. 77:663–665.
Schneider A, Dessimoz C, Gonnet GH. 2007. OMA Browser—exploring orthologous relations across 352 complete genomes. Bioinformatics. 23:2180–2182.
Studer RA, Penel S, Duret L, Robinson-Rechavi M. 2008. Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. Genome Res. 18:1393–1402.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–556.
Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22:2472–2479.

William Martin, Associate Editor
Accepted May 21, 2009