Inverse Relationship Between Evolutionary Rate and Age of Mammalian Genes

M. Mar Albà* and Jose Castresana†

*Research Group on Biomedical Informatics, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain; and †Department of Physiology and Molecular Biodiversity, Institut de Biologia Molecular de Barcelona, Barcelona, Spain.

A large number of genes is shared by all living organisms, whereas many others are unique to some specific lineages, indicating their different times of origin. The availability of a growing number of eukaryotic genomes allows us to estimate which mammalian genes are novel genes and, approximately, when they arose. In this article, we classify human genes into four different age groups and estimate evolutionary rates in human and mouse orthologs. We show that older genes tend to evolve more slowly than newer ones; that is, proteins that arose earlier in evolution currently have a larger proportion of sites subjected to negative selection. Interestingly, this property is maintained when a fraction of the fastest-evolving genes is excluded or when only genes belonging to a given functional class are considered. One way to explain this relationship is by assuming that genes maintain their functional constraints along all their evolutionary history, but the nature of more recent evolutionary innovations is such that the functional constraints operating on them are increasingly weaker. Alternatively, our results would also be consistent with a scenario in which the functional constraints acting on a gene would not need to be constant through evolution. Instead, starting from weak functional constraints near the time of origin of a gene—as supported by mechanisms proposed for the origin of orphan genes—there would be a gradual increase in selective pressures with time, resulting in fewer accepted mutations in older versus more novel genes.

Introduction

The protein sequence evolutionary rate, which can be effectively measured as the number of nonsynonymous substitutions per nonsynonymous site (Ka), is indicative of the intensity of the selective forces acting on a protein. Although the observation that different types of proteins evolve at different rates is not new (Wilson, Carlson, and White 1977; Doolittle et al. 1986; Nei 1987; Li 1997), the grounds for such marked heterogeneity have remained ill defined. Wilson, Carlson, and White (1977) predicted that proteins that differ in dispensability would be subject to different levels of purifying selection. This should result in significantly different Ka values for essential and nonessential genes. However, whereas some studies support the idea that more slowly evolving proteins tend to result in more severe knockout phenotypes (Hirsh and Fraser 2001), others fail to detect a strong correlation between gene dispensability and amino acid substitution rates (Hurst and Smith 1999; Yang, Gu, and Li 2003). Other studies have revealed that proteins that show low evolutionary rates tend to be expressed at high levels in yeast (Pal, Papp, and Hurst 2001) and in vertebrates (Subramanian and Kumar 2004), are ubiquitously expressed in mammalian tissues (Duret and Mouchiroud 2000; Zhang and Li 2004), and show a low propensity to be lost during eukaryote evolution (Krylov et al. 2003).

Another aspect that deserves attention is the relationship between the antiquity of a gene and its evolutionary rate. Since an early statement by Doolittle et al. (1986) that “some of the most ancient proteins are changing very slowly,” no specific analysis on this topic has been performed. Orphan genes, which are genes that have no known homologs in the genomes of other organisms and that are presumably of very recent origin (Dujon 1996; Fischer and Eisenberg 1999), have been shown to evolve faster than nonorphan genes in bacteria (Daubin and Ochman 2004) and Drosophila (Domazet-Loso and Tautz 2003). It has also been observed that vertebrate-specific genes evolve faster than older genes (Subramanian and Kumar 2004). An interesting possibility is that these observations are the result of a more general relationship between the time of origin of a gene and its evolutionary rate. To address this question, we examine, by using evolutionary rates in human and mouse orthologous genes, whether genes classified in four different age groups, from a few hundred million to a few thousand million years, evolve at similar rates. Our results show that there are significant differences among all age groups, with a consistent increase in the number of constrained sites with the age of the gene, demonstrating an important and general feature of the molecular evolution of proteins.

Methods

Databases

Human-mouse orthologous protein pairs, their corresponding gene-coding sequences, and human protein gene ontology (GO) molecular function annotations (Ashburner et al. 2000) were retrieved from ENSEMBL (Clamp et al. 2003). These orthologs are defined, briefly, as pairs of reciprocal Blast hits as well as pairs that have high similarity and conserved gene order (Clamp et al. 2003). Genes that have ambiguous orthologous assignations, denoted by more than one potential orthologous sequence in the other genome, were eliminated. In addition, we selected those orthologous gene pairs in which the human protein was functionally annotated by at least one GO molecular function term.

Alignments and Calculation of Evolutionary Rates

Orthologous gene pairs were aligned with ClustalW (Thompson, Higgins, and Gibson 1994) at the amino acid
level, and gaps were introduced in the nucleotide sequence according to the amino acid sequence alignment. For each gene pair, we calculated the number of nonsynonymous substitutions per nonsynonymous site (Ka) and the number of synonymous substitutions per synonymous site (Ks) using the maximum-likelihood method in the Codeml program of the PAML software package (Yang 2000). Equilibrium codon frequencies of the model were used as free parameters (CodonFreq = 3). We discarded pairs with very high substitution rates (Ka ≥ 0.5 and/or Ks ≥ 5 substitutions/position). The data set contained 6,776 human-mouse gene pairs after applying this filter.

Gene Ontology Functional Annotations

For the comparison of the Ka value distribution of proteins belonging to different GO classes, we used 4,936 gene pairs. This was the result of selecting GO classes that were well represented and had a limited level of overlap among themselves. To be selected, a GO class annotation had to be present in at least 30 different proteins. As a protein may have several GO annotations, we calculated the percentage of overlap in the proteins from different GO classes. If the overlap between two classes was more than 20% of the proteins in one class and also more than 20% of the proteins in the other class, the smallest class, representing a most specific function, was kept, and the largest class was discarded. This process led to the elimination of 15 GO classes. The final selected data set contained 70 GO groups.

Statistical Tests

To detect any statistical differences among groups, we used the Kolmogorov-Smirnov test, which is a non-parametric test. Correlations were calculated with the Spearman rank correlation method.

Blast Searches and Assignment of Temporal Categories

We used the human proteins from the human-mouse orthologous pairs to identify any homologous gene product in six different eukaryotic genomes, using BlastP (Altschul et al. 1997). The genomes used were from Saccharomyces cerevisiae, Schizosaccharomyces pombe, Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, and Takifugu rubripes. Protein sequences were downloaded from the Cogent Database release 153 (Janssen et al. 2003), except T. rubripes, which was obtained from the ENSEMBL release of March 12, 2003. To avoid spurious hits caused by the presence of low-complexity sequences, we filtered this type of region from the human sequences using the SEG program (Wootton and Federhen 1993) with default parameters. We considered that a homolog of the mammalian program (Wootton and Federhen 1993) with default parameters was a very high substitution rate (Ka ≥ 0.5 and/or Ks ≥ 5 substitutions/position). The data set contained 6,776 human-mouse gene pairs after applying this filter.

Results

Evolutionary Rate Differences Among Age Groups

To analyze the relationship between protein evolutionary rate and protein age and function we used a data set of 6,776 orthologous human-mouse sequences that contained gene ontology functional annotations (Ashburner et al. 2000), downloaded from ENSEMBL (Clamp et al. 2003). For each orthologous pair, we calculated the number of nonsynonymous substitutions per nonsynonymous site (Ka) and the number of synonymous substitutions per synonymous site (Ks) using a maximum-likelihood method (Yang 2000). Pairs with very high substitution rates (Ka ≥ 0.5 and/or Ks ≥ 5 substitutions/position) were discarded. This process led to the elimination of 15 GO classes. The final selected data set contained 70 GO groups.
In addition, the differences in the mean Ka were also remarkable: the mean Ka of proteins classified in group “TETRAPODS” (0.2317 substitutions/position), corresponding to the most recent proteins, was four times larger than in group “OLD,” proteins present in all eukaryotes analyzed (0.0571 substitutions/position). Differences in Ks were smaller among age groups, whereas the ratio Ka/Ks (i.e., the rate of nonsynonymous substitutions corrected for neutral rates) showed a trend similar to Ka (table 1). Changing the sensitivity of the Blast detection method from an E-value cutoff of $10^{-24}$ to a more conservative one of $10^{-10}$ did not significantly affect our results (data not shown).

The use of a sequence similarity detection method for the identification of homologs in eukaryotic genomes is expected to be reliable for proteins that evolve slowly but may present some limitations for quickly evolving proteins. It should be noted, however, that a protein can be very fast evolving because of multiple substitutions, but, if there is a group of conserved residues in close proximity, they may be sufficient for Blast to detect homology. In fact, the distribution of E-values against a specific genome indicated that, regardless of their Ka (< 0.5 substitutions/position in any case), most proteins could be confidently detected. For example, a large majority (> 77%) of the top 10% most-divergent proteins (Ka > 0.187 substitutions/position) was detected with E-values less than $10^{-10}$ in all genomes (see Supplementary Material online), indicating that, indeed, divergent proteins are detected with high confidence. To be even more conservative, we performed the statistical comparisons of the distributions with only half of the proteins, representing the best-conserved fraction (Ka < 0.051 substitutions/position), where loss of sensitivity of Blast detection is almost negligible, as measured by the small decay in the number of proteins detected with E-value less than $10^{-10}$ in these Ka intervals (see Supplementary Material online). Under these conditions, the Ka distribution differences remained statistically significant between the groups “OLD,” “METAZOANS,” and “DEUTEROSTOMES” (“TETRAPODS” could not be compared, as there were only nine proteins left in this group [data not shown]). Although the effect of Blast searches is probably not null, these data indicated the robustness of the underlying Ka differences between age groups.

**Table 1**

<table>
<thead>
<tr>
<th>Age Group</th>
<th>N</th>
<th>Ka</th>
<th>Ks</th>
<th>Ka/Ks</th>
<th>Length (Human)</th>
<th>Low-Complexity (Human)</th>
<th>% Indels</th>
<th>% GC (Human)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old</td>
<td>2982</td>
<td>0.0571</td>
<td>0.7429</td>
<td>0.0816</td>
<td>587.2</td>
<td>0.072</td>
<td>2.6</td>
<td>51.4</td>
</tr>
<tr>
<td>Metazoans</td>
<td>1075</td>
<td>0.0790</td>
<td>0.8354</td>
<td>0.1038</td>
<td>505.3</td>
<td>0.088</td>
<td>2.5</td>
<td>54.7</td>
</tr>
<tr>
<td>Deuterostomes</td>
<td>448</td>
<td>0.1350</td>
<td>0.9327</td>
<td>0.1691</td>
<td>338.9</td>
<td>0.105</td>
<td>3.2</td>
<td>55.5</td>
</tr>
<tr>
<td>Tetrapods</td>
<td>201</td>
<td>0.2317</td>
<td>0.9556</td>
<td>0.2967</td>
<td>249.6</td>
<td>0.127</td>
<td>5.2</td>
<td>54.1</td>
</tr>
</tbody>
</table>

**Note.**—N is the number of genes. Nonsynonymous (Ka) and synonymous (Ks) evolutionary rates are in substitutions per site. Length refers to amino acid length. Low-complexity content is the fraction of a protein with low-complexity sequence as determined by the SEG program.
(P < 10^{-4}, Kolmogorov-Smirnov test), with tetrapod genes being more than two times shorter than old genes (table 1). This finding would be consistent with the observation that shorter genes tend to have higher Ka values (Lipman et al. 2002). Because the length of genes could also affect their detection by Blast, we calculated evolutionary rate differences in genes shorter than 150 amino acids. In this set, there were no statistical differences in length among the four age groups, but the differences in Ka remained strong: “OLD,” “METAZOANS,” “DEUTEROSTOMES,” and “TETRAPODS” had 0.036, 0.106, 0.142, and 0.216 substitution/site, respectively. The differences in Ka among age classes were all significant, except in the “METAZOANS”/“DEUTEROSTOMES” comparison. Thus, differences in Ka among age classes were not caused by differences in length.

It is well known that some regions in protein sequences show a high degree of repetitiveness or low sequence complexity (Green and Wang 1994; Alba and Guigo 2004). Many of these regions may have been generated by DNA slippage (Tautz, Trick, and Dover 1986). We observed that the fraction of a protein occupied by low-complexity sequences, as determined by the SEG program (Wootton and Federhen 1993), also showed an inverse correlation with age (table 1). All group-to-group differences were significant (P < 10^{-3}), except for the comparison “TETRAPODS”/“DEUTEROSTOMES.”

These results are in accordance with the observation that modern eukaryotic proteins show a high degree of repetitiveness (Nishizawa and Nishizawa 1999). So, as with point mutations, the accumulation of products of slippage appears to be higher in more novel genes.

We also calculated the percentage of nucleotides involved in internal indels (i.e., after discarding terminal gaps) in the alignments of human and mouse proteins of different ages. As expected, the proportion of indels increased in newer proteins, from 2.6% in “OLD” proteins to 5.2% in “TETRAPODS” (table 1). Differences between groups were significant except in the “OLD”/“METAZOANS” comparison.

The GC content was similar for the three most recent groups, whereas it was significantly smaller in the oldest group (table 1). Thus, it seems that old genes have low GC, or they tend to be located in regions (isochores) of low-GC content, but this effect does not change with age. The differences observed in Ks between genes of different ages (table 1) could be a consequence of the known correlation between Ka and Ks, which, in turn, could be caused by tandem substitutions and by a fraction of amino acids that evolve neutrally, among other possible causes (Wolfe and Sharp 1993; Lercher, Chamary, and Hurst 2004). Alternatively, because genes in different chromosome regions have been shown to have different underlying Ks values (Lercher, Williams, and Hurst 2001; Castresana 2002), it could also reflect clustering of genes of different ages. To study this possibility, we analyzed the physical position in the human genome of the 201 tetrapod genes, which are the newest ones and could be the most affected by a biased chromosome distribution. However, these genes are distributed in a similar manner in all chromosomes, and we found no evidence of clustering in any particular region (data not shown).

Protein Age and Function

To get further insight into the connection between protein evolutionary rate and age, we analyzed the relationship between these variables in light of the function of the protein. For this purpose, we first compared the Ka values of proteins associated with different molecular function GO annotations. The data set we used contained 4,936 different orthologous pairs and 70 GO classes with minimum overlap (see Methods). As observed in table 2, almost one order of magnitude separates the mean Ka values of proteins annotated as “pre-mRNA splicing factor” (mean Ka 0.021 substitutions/position), the most slowly evolving type, and proteins annotated as “lectin” (mean Ka 0.186 substitutions/position), the most-divergent type. Thus, our data indicated that, on the one hand, ancestral proteins tended to have lower evolutionary rates than do novel proteins, and, on the other hand, significant differences in Ka values were detectable among different GO functional classes. This finding raised the interesting possibility that functional types of proteins that showed a high degree of sequence conservation were of a more ancestral character. To investigate this possibility, for each human protein GO class we plotted the fraction of proteins that had at least one homolog in a given eukaryotic genome (as a measure of its degree of antiquity) versus the mean Ka value of the GO class. For the six eukaryotic genomes analyzed, there was a significant negative correlation between these two variables (figure 2 shows the results obtained using the S. cerevisiae and C. elegans proteomes). Therefore, the differences in the mean Ka values of proteins under different GO functional annotations were indeed related to the degree of overall ancestry of such function.

We also analyzed whether the effect of the age of proteins on the distribution of evolutionary rates was caused by a few abundant functions or whether this occurred across different functions. An analysis of frequencies of the most abundant GO functions (> 30 genes) in different age classes showed that most functions were present in all age classes, but some of them were underrepresented or overrepresented in particular classes. A subsequent correspondence analysis of the contingency table (data not shown) indicated that the class “METAZOANS” had the most biased distribution of GO functions, with “extracellular ligand-gated ion channel,” “steroid hormone receptor,” “trypsin,” “chymotrypsin,” “G-protein–coupled receptor,” and “rhodopsin-like receptor” being highly overrepresented as proteins of metazoan origin. This is likely the result of the high number of innovations that occurred just before the explosive radiation of metazoa. Other GO classes with biased distributions in the correspondence analysis were “cytokine” overrepresented in “TETRAPODS,” and “growth factor” overrepresented in both “METAZOANS” and “TETRAPODS.” Elimination of the 398 genes that contained any of these eight biased GO functions did not affect the inverse correlation between Ka and gene age. In addition, the inverse relationship between protein age and evolutionary rate observed in the complete data set (fig. 1) also existed within specific functional classes that had at
least two representatives in the four age groups (fig. 3). In almost all GO classes, there was a clear progression in nonsynonymous rates from the "OLD" to the "TETRAPODS" groups. Furthermore, although the sample size is small for many functions, at least for those better represented ("DNA binding," "RNA binding," "calcium ion binding," "receptor," and "transcription factor"), differences in Ka were statistically significant between the "OLD" and "TETRAPODS" classes. Thus, the higher proportion of accelerated proteins among the most recent genes is not the result of some specific functional classes; rather it affects a whole range of different functions.

Table 2

<table>
<thead>
<tr>
<th>Gene Ontology Class</th>
<th>N</th>
<th>Mean Ka</th>
<th>Gene Ontology Class</th>
<th>N</th>
<th>Mean Ka</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-mRNA splicing factor</td>
<td>46</td>
<td>0.0215</td>
<td>Kinase</td>
<td>86</td>
<td>0.0688</td>
</tr>
<tr>
<td>GTP binding</td>
<td>151</td>
<td>0.0323</td>
<td>Structural constituent of ribosome</td>
<td>111</td>
<td>0.0696</td>
</tr>
<tr>
<td>Ubiquitin conjugating enzyme</td>
<td>37</td>
<td>0.0361</td>
<td>Transcription coactivator</td>
<td>95</td>
<td>0.0698</td>
</tr>
<tr>
<td>Small monomeric GTPase</td>
<td>41</td>
<td>0.0374</td>
<td>Isomerase</td>
<td>55</td>
<td>0.0699</td>
</tr>
<tr>
<td>RAB small monomeric GTPase</td>
<td>43</td>
<td>0.0386</td>
<td>Ubiquitin C-terminal hydrolase</td>
<td>36</td>
<td>0.0701</td>
</tr>
<tr>
<td>Heat shock protein</td>
<td>30</td>
<td>0.0415</td>
<td>Structural molecule</td>
<td>112</td>
<td>0.0706</td>
</tr>
<tr>
<td>cAMP-dependent protein kinase</td>
<td>38</td>
<td>0.0417</td>
<td>Lyase</td>
<td>73</td>
<td>0.0708</td>
</tr>
<tr>
<td>Protein kinase CK2</td>
<td>38</td>
<td>0.0417</td>
<td>ATP-binding cassette (ABC) transporter</td>
<td>30</td>
<td>0.0721</td>
</tr>
<tr>
<td>Voltage-gated potassium channel</td>
<td>54</td>
<td>0.0422</td>
<td>Nucleic acid binding</td>
<td>189</td>
<td>0.0734</td>
</tr>
<tr>
<td>Translation initiation factor</td>
<td>38</td>
<td>0.0428</td>
<td>Tumor suppressor</td>
<td>103</td>
<td>0.0768</td>
</tr>
<tr>
<td>Adenosinetrifosphatase</td>
<td>41</td>
<td>0.0443</td>
<td>Transporter</td>
<td>155</td>
<td>0.0770</td>
</tr>
<tr>
<td>Structural constituent of cytoskeleton</td>
<td>38</td>
<td>0.0464</td>
<td>DNA binding</td>
<td>532</td>
<td>0.0773</td>
</tr>
<tr>
<td>RNA binding</td>
<td>239</td>
<td>0.0464</td>
<td>Acytransferase</td>
<td>45</td>
<td>0.0780</td>
</tr>
<tr>
<td>Calmodulin binding</td>
<td>60</td>
<td>0.0471</td>
<td>Extracellular matrix structural protein</td>
<td>46</td>
<td>0.0784</td>
</tr>
<tr>
<td>Motor</td>
<td>47</td>
<td>0.0495</td>
<td>Enzyme</td>
<td>98</td>
<td>0.0796</td>
</tr>
<tr>
<td>Extracellular ligand-gated ion channel</td>
<td>30</td>
<td>0.0512</td>
<td>Hydrolase</td>
<td>491</td>
<td>0.0813</td>
</tr>
<tr>
<td>GTPase activator</td>
<td>32</td>
<td>0.0513</td>
<td>Electron transfer flavoprotein</td>
<td>46</td>
<td>0.0826</td>
</tr>
<tr>
<td>CTD phosphatase</td>
<td>32</td>
<td>0.0522</td>
<td>Calcium ion binding</td>
<td>315</td>
<td>0.0844</td>
</tr>
<tr>
<td>Potassium channel</td>
<td>32</td>
<td>0.0540</td>
<td>Zinc binding</td>
<td>137</td>
<td>0.0864</td>
</tr>
<tr>
<td>Protein tyrosine kinase</td>
<td>144</td>
<td>0.0541</td>
<td>Rhodopsin-like receptor</td>
<td>185</td>
<td>0.0868</td>
</tr>
<tr>
<td>Actin binding</td>
<td>81</td>
<td>0.0541</td>
<td>Cysteine-type endopeptidase</td>
<td>40</td>
<td>0.0880</td>
</tr>
<tr>
<td>RNA polymerase II transcription factor</td>
<td>99</td>
<td>0.0555</td>
<td>Electron transporter</td>
<td>93</td>
<td>0.0924</td>
</tr>
<tr>
<td>Guanylnucleotide exchange factor</td>
<td>39</td>
<td>0.0566</td>
<td>G-protein-coupled receptor</td>
<td>43</td>
<td>0.0941</td>
</tr>
<tr>
<td>Ion channel</td>
<td>45</td>
<td>0.0569</td>
<td>Metalloendopeptidase</td>
<td>37</td>
<td>0.1054</td>
</tr>
<tr>
<td>SH3/SH2 adaptor protein</td>
<td>37</td>
<td>0.0595</td>
<td>DNA repair protein</td>
<td>38</td>
<td>0.1114</td>
</tr>
<tr>
<td>Transcription factor</td>
<td>460</td>
<td>0.0609</td>
<td>Receptor binding</td>
<td>34</td>
<td>0.1168</td>
</tr>
<tr>
<td>Protein binding</td>
<td>287</td>
<td>0.0619</td>
<td>Receptor</td>
<td>408</td>
<td>0.1172</td>
</tr>
<tr>
<td>Magnesium binding</td>
<td>76</td>
<td>0.0628</td>
<td>Receptor-signaling protein</td>
<td>30</td>
<td>0.1225</td>
</tr>
<tr>
<td>Transcription corepressor</td>
<td>68</td>
<td>0.0638</td>
<td>Apoptosis regulator</td>
<td>30</td>
<td>0.1262</td>
</tr>
<tr>
<td>Steroid hormone receptor</td>
<td>37</td>
<td>0.0643</td>
<td>Chymotrypsin</td>
<td>30</td>
<td>0.1312</td>
</tr>
<tr>
<td>ATP-dependent helicase</td>
<td>47</td>
<td>0.0652</td>
<td>Trypsin</td>
<td>39</td>
<td>0.1318</td>
</tr>
<tr>
<td>Protein kinase</td>
<td>54</td>
<td>0.0676</td>
<td>Serine protease inhibitor</td>
<td>31</td>
<td>0.1341</td>
</tr>
<tr>
<td>Peptidase</td>
<td>61</td>
<td>0.0682</td>
<td>Transmembrane receptor</td>
<td>61</td>
<td>0.1540</td>
</tr>
<tr>
<td>Chaperone</td>
<td>70</td>
<td>0.0686</td>
<td>Cytokine</td>
<td>50</td>
<td>0.1630</td>
</tr>
<tr>
<td>Signal transducer</td>
<td>145</td>
<td>0.0687</td>
<td>Lectin</td>
<td>39</td>
<td>0.1860</td>
</tr>
</tbody>
</table>

Note.—Number of proteins (N) and mean Ka value (in substitutions/site) for 70 different human protein gene ontology (GO) classes.

Fig. 2.—Relationship between the fraction of human proteins annotated under a given gene ontology (GO) class that have at least one homolog in the Saccharomyces cerevisiae or the Caenorhabditis elegans genome (homology fraction) and the mean nonsynonymous substitution rate (mean Ka). Data points represent the 70 GO classes in table 2. Linear regression fit is shown. There is a significant negative correlation between the fraction of homologs and the mean Ka (S. cerevisiae: $r = -0.53$, $P < 0.0001$; C. elegans: $r = -0.55$, $P < 0.0001$). The correlation is also significant for the comparisons with the other analyzed genomes (Schizosaccharomyces pombe: $r = -0.52$, $P < 0.0001$; Arabidopsis thaliana: $r = -0.48$, $P < 0.0001$; Drosophila melanogaster: $r = -0.53$, $P < 0.0001$; Takifugu rubripes: $r = -0.26$, $P = 0.0288$).
Discussion

Our results show that, according to the classification of genes in different age groups as derived from Blast hits, old proteins evolve more slowly than new ones. Although we are measuring evolutionary rates in orthologous human and mouse genes (separated approximately 80 MYA), we observe that these rates are related to the time of origin of genes that may be traced back several hundred million or a few thousand million years ago. Thus, genes that are exclusive of tetrapods (found in mammals but not in Fugu) evolve faster than genes of deuterostome distribution (found in mammals and Fugu but not in other metazoans); deuterostome genes show higher substitution rates than metazoan genes (found in all metazoans analyzed but not in plants or yeasts); and old genes (found in all these lineages) are the most conserved. This finding indicates that evolutionary rates progressively diminish with the age of a gene. Genes classified as “OLD” in our study are common to all eukaryotes, and many of them were probably present in the first cellular organisms, so they are likely to perform essential housekeeping cellular functions, which may explain that this class of genes is the most conserved one (Zhang and Li 2004). However, the progression in the degree of variability on the three classes of newly arisen genes (“METAZOANS,” “DEU- TEROSTOMES,” and “TETRAPODS”) is not obvious and requires a more specific explanation. To this end, it may be first useful to consider possible mechanisms for the origin of such new genes.

According to current knowledge, most novel genes probably originated from gene duplications. Normally, sequence changes are relatively fast in one of the copies during the first few million years after the duplication, and, after this initial neutral phase, genes that are not silenced start a period of strong purifying selection (Lynch and Conery 2000; Long et al. 2003). In some cases, however, there may be so many changes in the initial neutral phase that the similarity is virtually erased along all the sequence. As a result, the new duplicate can no longer be recognized by Blast as homologous to the original copy (Schmid and Tautz 1997; Schmid and Aquadro 2001; Domazet-Loso and Tautz 2003). This would lead to what it is normally considered a new gene from the sequence point of view. In addition, such novel genes are likely to contribute to completely or partially new functions for the organism. Thus, the main feature of these novel genes, in contrast to other duplicated genes where sequence similarity is not lost, would be the existence of almost no constraint during the phase of fast evolution and the consequent lack of detection by Blast of the original gene because of sequence changes along all the sequence. Of course, after the initial phase of rapid sequence diversification, these new genes must undergo a subsequent phase of purifying selection, or otherwise they would be rapidly silenced. Therefore, at least for a period of time after a gene has originated, there is a progression towards increased selective pressure, measurable as increasingly lower evolutionary rates. This mechanism, proposed for the origin of genes of very restricted phylogenetic distribution or orphan genes (Domazet-Loso and Tautz 2003), may also apply to the earlier origin of “METAZOAN,” “DEUEROSTOME,” and “TETRAPOD” genes. Although sequence similarity is lost, for some of these proteins it might be possible to detect structurally related proteins that could be related to the original copy of the duplicated gene (Mueller et al. 2004).

Although it is likely that this type of duplications is a common mechanism for the formation of new genes with
nondetectable homologs in other lineages, another possibility is that some genes or part of their sequences originated de novo. The smaller mean length of newer genes is in agreement with this, because de novo formation of small genes should be easier than formation of larger ones. In our data set, low-complexity sequences are more abundant in novel proteins, which suggests that the changes induced by DNA slippage may be better tolerated in this type of protein. For example, 20.7% of the new tetrapod genes, but only 7% of the genes present in all eukaryotes, show a very high content in low-complexity regions (>20% of the protein). Thus, it seems plausible that repeat expansion by the action of slippage has contributed to the formation of new sequence regions in novel genes and, exceptionally, to the formation of novel short genes.

As stated above, our results indicate that there is an inverse relationship between gene age and evolutionary rate. In light of the most likely mechanisms operating at the time of origin of novel or orphan genes, one possible explanation for this effect is that the increase in the number of constrained sites after an initial phase of fast evolution after the gene duplication is not limited to a short period of time, but the trend applies to proteins long after they have originated. For example, it may be that in the case of older proteins, a larger part of the protein is directly involved in function as a result of many interacting partners or multiple functions that have accumulated through evolution, making each substitution less likely to be neutral or advantageous. In this respect, it has been observed that there is a positive relationship between the antiquity of protein folds and the number of interacting partners (Park and Bolser 2001), and, independently, proteins that have many interactions with other proteins tend to have lower than average evolutionary rates (Fraser et al. 2002; Teichmann 2002 [but see Jordan, Wolf, and Koonin {2003}]). Also in relation to this hypothesis, older genes are more likely to be functional in many different tissues and broadly expressed genes have been shown to evolve more slowly (Duret and Mouchiroud 2000; Zhang and Li 2004). Thus, the few constraints at the time of origin of a gene and the gradual accumulation of functional or structural sites since that time may explain the tendency of older proteins to have lower substitution rates.

A second possibility that could explain the relationship between gene age and evolutionary rate is that novel genes may have maintained their degree of functional constraint along all or most of their evolutionary history. Under this hypothesis, old evolutionary innovations (e.g., multicellularity, signal transduction, or motility) would have given rise to genes coding for proteins with more functional sites than genes appeared later in the deuterostomes or tetrapods, which would be less likely to contribute to essential cellular functions. Alternatively, rate constancy of genes through evolution and the appearance of a similar proportion of essential and nonessential genes at all stages of evolution would be possible if we consider the differential elimination of genes with different evolutionary rates from the genome (Krylov et al. 2003). Genes that are quickly evolving are less likely to be essential, and a deletion or other mutation that eliminates them from the genome may not be deleterious. Thus, the fastest genes would be proportionally more abundant in the younger categories because fast genes that originated long ago are more likely to have been eliminated from the genome. A difficulty with this hypothesis of constant constraints through evolution is that proteins of relatively old origin and which show strong constraints (for example “META-ZOAN” proteins) should have been very constrained since their time of origin. It may be possible, however, that a mechanism similar to the one proposed for the origin of orphan genes but followed by a sudden period of strong positive selection may lead to the rapid appearance of novel genes with many functional sites.

Phylogenetic tree reconstruction methods with evolutionary models that include a variable molecular clock along the tree may help decide among the different hypotheses for explaining the relationship between the age and rate of genes. Actually, it has been repeatedly observed that different protein residues switch in substitution rate over time, giving rise to the so-called covarion or covarion-like models of protein evolution (Fitch and Markowitz 1970; Fitch 1971; Miyamoto and Fitch 1995; Lopez, Forterre, and Philippe 1999; Galtier 2001; Penny et al. 2001; Huelsenbeck 2002; Philippe et al. 2003). When overall evolutionary rates rather than single protein residues are considered, it is also evident that different parts of the tree have different substitution rates (Gillespie 1991; Sanderson 1997; Thorne, Kishino, and Painter 1998; Huelsenbeck, Larget, and Swofford 2000; Aris-Brosou and Yang 2002; Sanderson 2002; Aris-Brosou and Yang 2003; Seo, Kishino, and Thorne 2004). If proteins have approximately maintained their degree of constraints since their time of origin, the number of substitutions will be evenly distributed over the tree (and rate changes will be in both directions, toward rate increase and toward rate decrease, in all parts of the tree). On the contrary, if proteins that are currently evolving slowly were faster at the time of their origin, the number of substitutions will be higher at the base of the tree and lower towards the tips. This autocorrelation of rates would only be detected in trees that include a broad representation of lineages and, thus, where the base of the tree is close to the origin of the gene. An analysis of evolutionary rates in different parts of the trees of some metazoan-specific proteins is consistent with the latter model (Iwabe, Kuma, and Miyata 1996; Miyata and Suga 2001). It will be interesting to know which of these two models of evolution, a “constant constraint” or an “increasing constraint” model, has been followed by the different genes of the mammalian genome.

Acknowledgments

M.M.A. is recipient of a Ramón y Cajal contract of the Spanish Ministerio de Educación y Ciencia (MEC). M.M.A and J.C. are supported by grant numbers BIO2002-04426-C02-01 and BIO2002-04426-C02-02, respectively, from the Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (I+D+I) of the MEC, cofinanced with FEDER funds. We thank Roderic Guigó, Jaume Bertranpetit, Noura Dabbouseh, Martin Lercher, and two anonymous reviewers for making useful suggestions.
Literature Cited


Michele Vendruscolo, Associate Editor

Accepted November 2, 2004