Differential loss of ancestral gene families as a source of genomic divergence in animals

Austin L. Hughes* and Robert Friedman

Department of Biological Sciences, University of South Carolina,
Coker Life Sciences Building, 700 Sumter Street, Columbia, SC 29208,
USA

* Author for correspondence (austin@biol.sc.edu).

Recd 17.09.03; Accptd 13.10.03; Online 12.11.03

A phylogenetic approach was used to reconstruct the pattern of an apparent loss of 2106 ancestral gene families in four animal genomes (*Caenorhabditis elegans*, *Drosophila melanogaster*, human and fugu). Substantially higher rates of loss of ancestral gene families were found in the invertebrates than in the vertebrates. These results indicate that the differential loss of ancestral gene families can be a significant factor in the evolutionary diversification of organisms.

Keywords: animal phylogeny; gene content; gene deletion; genome evolution

1. INTRODUCTION

Comparison of eukaryotic genomes has revealed that both duplication and deletion of genes play important roles in the evolutionary divergence between different lineages of organisms (Lynch & Conery 2000; Lynch 2002; Hughes & Friedman 2003). Several recent studies have addressed patterns of gene duplication, including the duplication of genomic segments (Friedman & Hughes 2001a; Bailey *et al.* 2002) and patterns of duplication of individual genes in different lineages (Friedman & Hughes 2001b, 2003; Gu *et al.* 2002; McLysaght *et al.* 2002; Hughes & Friedman 2003). The role of gene deletion has been studied through phylogenetic analysis of gene families undergoing the process of 'birth-and-death' evolution, whereby members of a family duplicate and are deleted differentially in different lineages (Hughes & Nei 1989; Nei *et al.* 1997; Piontkivska *et al.* 2002).

However, little is known regarding the evolutionary role played by gene deletions that lead to the loss of an entire gene family from a given lineage (Aravind *et al.* 2000; Roelofs & Van Haastert 2001). Here, we use a phylogenetic approach to address this question in three major groups of animals: the nematodes, the insects and the vertebrates. We reconstructed a set of protein-coding gene families present in the common ancestor of these animal species and then reconstructed the pattern of apparent family loss in each lineage since that ancestor.

2. METHODS

To reconstruct the set of ancestral gene families in animals, we used complete sets of predicted protein translations (proteomes) for the following organisms (downloaded from <http://iubio.bio.indiana.edu:8089/> except where indicated otherwise): (i) two fungi,

Saccharomyces cerevisiae (v. 06/24/2002) and *Schizosaccharomyces pombe* (http://www.sanger.ac.uk/Projects/S_pombe/, v. 02/25/2003); (ii) a plant *Arabidopsis thaliana* (v. 06/24/2002); (iii) *Caenorhabditis elegans* (v. 06/24/2002), belonging to the pseudocoelomate animal phylum Nematoda (nematode worms); (iv) *Drosophila melanogaster* (v. 06/24/2002), an insect belonging to the phylum Arthropoda (arthropods), a coelomate protostome animal phylum; and (v) two vertebrates, human *Homo sapiens* (v. 06/24/2002) and fugu *Takifugu rubripes* (<http://genome.jgi-psf.org/fugu6/fugu6.home.html>, v. 3.0), belonging to Chordata, a coelomate deuterostome animal phylum. Protein families were constructed from the above proteomes using the BLASTCLUST computer program available in the BLAST tools (Altschul *et al.* 1997), which establishes families by BLASTP homology search and the single-linkage method.

We used a value of 10^{-6} for the *E* parameter (the probability that a score as high as that observed between two sequences will be found by chance in a database of the size examined) of the BLAST algorithm. We assembled protein families using two sets of criteria to count a match between a pair of sequences: (i) that 20% of amino acids be identical and 30% of aligned sites be shared; and (ii) that 30% of amino acids be identical and 50% of aligned amino acid sites be shared. We used these two sets of criteria as a way of controlling for the possibility that, given strict match criteria, a family would be scored as having been lost over the course of evolution if it had merely diverged substantially in sequence. Using the stricter criteria, more families were identified because larger families were broken up into separate families (data not shown). However, the overall pattern was qualitatively very similar with the two sets of criteria (data not shown). Here, we present only the results using the less strict criteria, because on these criteria it was less likely that a family would be scored as having been lost when it had merely diverged in sequence.

The phylogenetic relationships of the species were reconstructed by the maximum-parsimony (MP) method applying the branch-and-bound algorithm (Swofford 2002) to a data matrix in which each of 3786 families present in at least two of the genomes was treated as a clastic character (scored 'present' or 'absent'). There were 642 of these families present in all genomes analysed; 267 were parsimony-uninformative and 2877 were parsimony-informative. The reliability of branching patterns in the MP tree was tested by bootstrapping (Felsenstein 1985); 1000 bootstrap samples were used.

We rooted the phylogenetic tree of animals using the plant and fungal species as outgroups, and we used the rooted tree to reconstruct patterns of loss of gene families in animals; a family reconstructed as present in an ancestor but absent in a descendant was scored as 'lost'. Because we reconstructed the set of ancestral families only for animals, this result does depend on resolution of the still controversial question of whether plants or fungi is the sister group to animals (Hedges 2002).

3. RESULTS

MP analysis yielded a single most parsimonious tree, in which all branches received 100% bootstrap support (figure 1a). If the phylogeny of animals was rooted with the plant and fungal sequences, the phylogenetic tree supported the hypothesis that the phylum Nematoda constitutes an outgroup to the coelomate phyla Arthropoda (including the insects) and Chordata (including the vertebrates) (figure 1a). This relationship is contrary to the 'Ecdysozoa' tree that has been proposed recently (Aguinaldo *et al.* 1997). An extensive phylogenetic analysis based on protein sequences (Blair *et al.* 2002) supported the same relationships revealed by our analysis. Assuming this phylogenetic tree, 2106 families were inferred to have been present in the last common ancestor of all four animal species analysed; in other words, the last common ancestor of coelomates (including insects and vertebrates) and nematodes.

Out of the 2106 families present in the last common ancestor of coelomates and nematodes, 400 (19.0%) were lost in the lineage leading to *C. elegans* (figure 1b). 428 (20.3%) of these families were lost in the lineage leading to *Drosophila*. Out of these 428 families, 15 were lost between the common ancestor of coelomates and nematodes and the common ancestor of deuterostomes and protostomes (figure 1b). The remaining 413 were lost after

Figure 1. (a) The most parsimonious tree based on the presence or absence of 3786 gene families, with a total length of 3984 character state changes (consistency index: 0.77). Numbers on branches represent the percentage of bootstrap samples supporting the branch. (b) Reconstructed numbers of gene families (circled) at ancestral nodes of animal phylogeny and numbers of families lost on each branch. (c) Numbers of ancestral families unique to each node in the phylogeny of the four animal genomes.

the divergence of the protostome lineage (figure 1b). By contrast, only 198 (9.4%) out of the 2942 ancestral families were lost on the lineage leading to *fugu*, and only 154 (7.3%) on the lineage leading to humans. Thus the proportion of ancestral gene families lost in vertebrates was less than half of that observed in the invertebrates. As a result of the differential levels of gene family loss, there were only 15 ancestral families unique to *C. elegans* and only 10 unique to *Drosophila*, as opposed to 167 unique to vertebrates (figure 1c).

4. DISCUSSION

Each of the four animal genomes that we analysed was characterized by possessing a unique subset of the 2106 families inferred to have been present in their common ancestor. This in turn suggests that specialization of genomes through differential loss of ancestral gene families has played a role in the differentiation of the animal phyla. It is possible, of course, that in some cases gene families have diverged so far that they were not detectable by homology search, even using the liberal criteria (20% amino acid identity and 30% of aligned sites shared) applied here. However, this was unlikely because the genes

included in this analysis were, by definition, conserved proteins. By contrast, known rapidly evolving proteins generally belong to taxon-specific families, such as the immune system gene families of vertebrates (Murphy 1993; Hughes 1997). In addition, essentially the same pattern of reconstructed gene family loss (data not shown) was seen using stricter search criteria (30% amino acid identity and 50% of aligned sites shared), suggesting that the overall pattern was not affected by non-detection of a certain proportion of homologues.

The human genome has been estimated to include *ca.* 30 000–40 000 protein-coding genes, roughly 2–3 times as many as *Drosophila* and 1.5–2 times as many as *C. elegans* (Claverie 2001). One factor that has contributed to the larger genome size in vertebrates than in these two invertebrates has undoubtedly been an increased rate of gene duplication in the vertebrates (Friedman & Hughes 2001b, 2003; Gu *et al.* 2002; McLysaght *et al.* 2002). The present results show that an additional contributing factor has been a reduced rate of loss of ancestral gene families in the vertebrate lineage in comparison with the *C. elegans* and *Drosophila* lineages.

The last common ancestor of deuterostomes and protostomes has been estimated to have occurred *ca.* 830 Myr ago (Nei *et al.* 2001). Assuming this divergence time, our results (figure 1b) yield an estimate of 2.4×10^{-10} loss family⁻¹ yr⁻¹ for the rate of loss of ancestral gene families since the deuterostome–protostome ancestor in the lineage leading to *Drosophila*. However, we estimate the rate of loss of ancestral gene families over the same period in the lineage leading to humans to be 0.9×10^{-10} loss family⁻¹ yr⁻¹. Interestingly, the ratio of these two values approximates the inverse of the ratios of estimated gene numbers in these two genomes.

Our results have implications for the controversial question of horizontal gene transfer (HGT) from prokaryotes to eukaryotes (Roelofs & Van Haastert 2001; Salzberg *et al.* 2001). The occurrence of certain gene families in bacterial genomes and in the human genome, but not in other sequenced eukaryotic genomes, led the International Human Genome Sequencing Consortium (2001) to conclude that there have been multiple events of HGT from bacteria to humans. Wolf *et al.* (2000) previously made a similar inference regarding *C. elegans*. In both cases, the argument for HGT was based on the assumption that parallel loss of the same gene family in multiple eukaryotic lineages is highly unlikely. However, our results suggest that this reasoning is faulty. Consider five hypothetical eukaryotic lineages with a common ancestor 1 billion years ago. Assuming a rate of gene loss similar to that computed above for the human lineage, the probability that a given gene family would be lost in four out of the five lineages is in the order of 10^{-4} . This would seem to be a much higher probability than that of HGT from a prokaryote to a eukaryote (excluding organelle to nuclear gene transfer). Furthermore, the probability would be even higher in cases of parallel gene family loss, as observed between *C. elegans* and *Drosophila*.

Acknowledgements

This research was supported by grant GM066710 from the National Institutes of Health to A.L.H.

- Aguinaldo, A. M., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A. & Lake, J. A. 1997 Evidence for a clade of nematodes, arthropods, and other moulting animals. *Nature* **397**, 489–493.
- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
- Aravind, L., Watanabe, H., Lipman, D. J. & Koonon, E. V. 2000 Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl Acad. Sci. USA* **97**, 11 319–11 324.
- Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W. & Eichler, E. E. 2002 Recent segmental duplications in the human genome. *Science* **297**, 1003–1007.
- Blair, J. E., Ikeo, I., Gojobori, T. & Hedges, S. B. 2002 The evolutionary position of nematodes. *BioMed Central Evol. Biol.* **2002**, 2: 7. (See http://www2.shu.ac.uk/services/ic/infoservices/eris/info.cfm?db_id=590)
- Claverie, J. M. 2001 What if there are only 30 000 human genes? *Science* **291**, 1255–1257.
- Felsenstein, J. 1985 Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791.
- Friedman, R. & Hughes, A. L. 2001a Pattern and timing of gene duplication in animal genomes. *Genome Res.* **11**, 1842–1847.
- Friedman, R. & Hughes, A. L. 2001b Gene duplication and the structure of eukaryotic genomes. *Genome Res.* **11**, 373–381.
- Friedman, R. & Hughes, A. L. 2003 The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Mol. Biol. Evol.* **29**, 154–161.
- Gu, X., Wang, Y. & Gu, J. 2002 Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nature Genet.* **31**, 205–209.
- Hedges, S. B. 2002 The origin and evolution of model organisms. *Nature Rev. Genet.* **3**, 838–849.
- Hughes, A. L. 1997 Rapid evolution of immunoglobulin superfamily C2 domains expressed in immune system cells. *Mol. Biol. Evol.* **14**, 1–5.
- Hughes, A. L. & Nei, M. 1989 Evolution of the major histocompatibility complex: independent origin of nonclassical class I genes in different groups of mammals. *Mol. Biol. Evol.* **6**, 559–579.
- Hughes, A. L. & Friedman, R. 2003 Parallel evolution by gene duplication in the genomes of two unicellular fungi. *Genome Res.* **13**, 794–799.
- International Human Genome Sequencing Consortium 2001 Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
- Lynch, M. 2002 Gene duplication and evolution. *Science* **297**, 945–947.
- Lynch, M. & Conery, J. S. 2000 The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155.
- McLysaght, A., Hokamp, K. & Wolfe, K. H. 2002 Extensive genomic duplication during early chordate evolution. *Nature Genet.* **31**, 200–204.
- Murphy, P. M. 1993 Molecular mimicry and the generation of host defense protein diversity. *Cell* **72**, 823–826.
- Nei, M., Gu, X. & Sitnikova, T. 1997 Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc. Natl Acad. Sci. USA* **94**, 7799–7806.
- Nei, M., Xu, P. & Glazko, G. 2001 Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proc. Natl Acad. Sci. USA* **98**, 2497–2502.
- Piontkivska, H., Rooney, A. P. & Nei, M. 2002 Purifying selection and birth-and-death evolution in the histone H4 gene family. *Mol. Biol. Evol.* **19**, 689–697.
- Roelofs, J. & Van Haastert, J. P. 2001 Genes lost during evolution. *Nature* **411**, 1013–1014.
- Salzberg, S. L., White, O., Peterson, J. & Eisen, J. A. 2001 Microbial genes in the human genome: lateral transfer or gene loss? *Science* **292**, 1903–1906.
- Swofford, D. L. 2000 *PAUP*: phylogenetic analysis using parsimony (*and other methods)*. Sunderland, MA: Sinauer.
- Wolf, Y. I., Kondrashov, F. A. & Koonin, E. V. 2000 No footprints of primordial introns in a eukaryotic genome. *Trends Genet.* **16**, 333–334.