The fate of duplicated genes: loss or new function?

Andreas Wagner

Summary

Gene duplication events are important sources of novel gene functions. However, more often than not, a duplicate gene may lose its function and become a pseudogene. What is the relative frequency of these two scenarios: functional divergence versus gene loss? Given that most non-neutral mutations are deleterious, gene loss should be far more frequent than divergence. However, a recent empirical study[1] suggests that about 50% of all gene duplications will lead to functional divergence. The study infers the frequency of functional divergence from the size distribution of gene families produced by two successive genome duplications early in vertebrate evolution. Reasons for this unexpectedly high frequency of functional divergence are discussed. *BioEssays* **20**:785–788, 1998.

© 1998 John Wiley & Sons, Inc.

Predictions from population genetic theory

Gene duplication is a major force in genome evolution. It may be the predominant mechanism for the evolution of new gene functions.[2] However, the vast majority of duplicate genes may become pseudogenes through loss-of-function mutations. The argument is simply that deleterious mutations are much more frequent than advantageous mutations. Thus, as long as one duplicate gene functions normally, the other may go the more likely route of accumulating deleterious mutations.

The evolutionary fate of duplicated genes is the subject of a large body of work in population genetics, going back at least to J.B.S. Haldane.[3] However, the issue of loss versus divergence does not rank prominently in this line of work. Issues of greater concern are the maintenance of polymorphisms of functional versus nonfunctional alleles at both loci, and the distribution of times until a pseudogenes becomes fixed at one of the two loci.[4–7] Notable exceptions are a series of papers by Ohta and collaborators[8–12] regarding the evolution of large gene families by unequal crossing over, mutation, and selection. A study by Walsh[13] explicitly addresses the question of loss versus divergence after duplication of one gene. Predicted rates of gene loss are similar to those in Ohta's work.[13] Taken together, these results suggest that under mutation rates and population sizes realistic for a wide range of organisms, the rate of gene loss should be at least an order of magnitude higher than that of divergence.

Correspondence to: Andreas Wagner, Department of Biology, University of New Mexico, 167 Castetter Hall, Albuquerque, NM 87131–1091; E-mail: aw@santafe.edu

Empirical studies

Organisms evolve under complex conditions which may not be captured by abstract models. Thus, it is necessary to determine the rate of loss versus divergence by experiment or observation. The fact that most known genes belong to large families with extensive DNA sequence similarities is not informative in this context. Even if only a minute fraction of duplicated genes are retained, one will still obtain the observed distribution of gene family sizes if gene duplications are very frequent. Conversely, the observation that many gene families contain pseudogenes[14] does not say much about the relative rate of divergence either, if the fraction of pseudogenes versus functional genes can not be accurately determined. Furthermore, estimating rates of gene loss from duplication of individual genes is not feasible because the expansion and contraction rate of gene families is not an easily measured quantity. It is thus not a coincidence that the few empirical studies in this area take advantage of polyploidization events during the phylogeny of animal taxa to infer rates of gene loss.[15–17] In a polyploidization event all genes are duplicated simultaneously, yielding one large sample of duplicated genes, each of which may be lost or diverge functionally. Earlier empirical studies are based on electrophoretic separation of isozymes in salmonid and catostomid fish,[15–17] which have experienced tetraploidization events

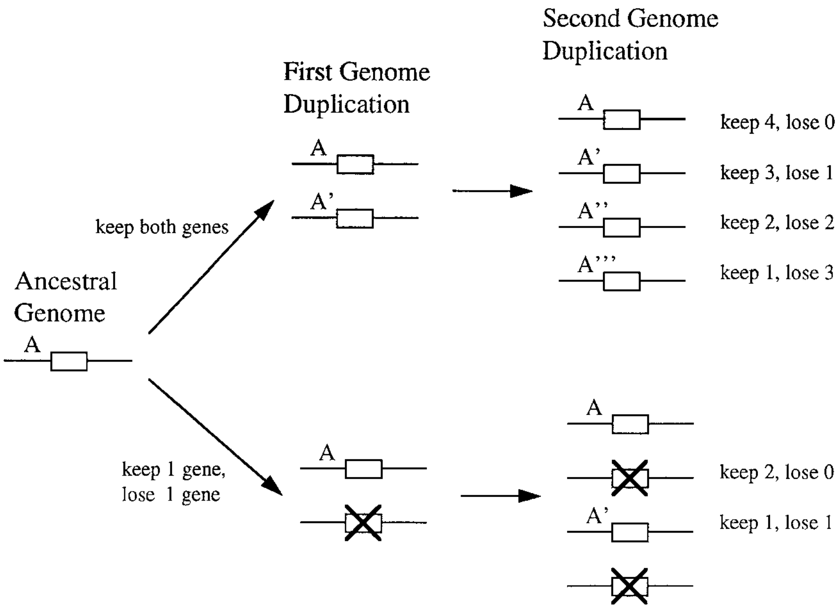

Figure 1. Alternative fates of genes following successive genome duplications, under the assumption that at least one member of each gene family must be retained. Individual gene family members, denoted by A, A', A'', and A''' can have identical, overlapping, or unique functions. After Figure 1 in ref. 1.

approximately 50 Myr ago. These studies suggest that approximately 50% of duplicate enzyme loci become silenced, a figure substantially higher than theoretical predictions. However, one can not strictly exclude the possibility that gene loss after polyploidization in these species is still an ongoing process. The critical parameter in this regard is the ''half-life'' of duplicated genes, estimates of which vary widely for vertebrates, from less than 1 Myr to 50 Myr.[18,19]

To circumvent this problem, one needs to study genome duplications that are significantly older than the half-life of most duplicated genes, but where the traces of the genome duplications can still be identified. A recent study by Nadeau and Sankoff.[1] analyzes such a case, the remnants of two consecutive vertebrate genome duplications that occurred more than 250 Myr ago. Several alternative fates are possible for each gene family created from one ancestral gene by two such duplications (Fig. 1). Because the probability of gene loss implies a probability for each of the scenarios depicted in Figure 1, one can make inferences on the probability of gene loss from an observed distribution of gene family sizes.

Nadeau and Sankoff's[1] study does precisely that. It is based on 276 human and 176 mouse gene families of two to four members, all of which are likely to have arisen in these two genome duplications. The investigators take advantage of the fact that genes duplicated in a genome duplication are distinguished by their chromosomal context. That is, they will be part of duplicated chromosome segments that reflect the arrangement of genes in the ancestral genome. This is in contrast to tandem duplication, where multiple copies of a gene are created at the same site in the genome, or replicative transposition, which may move gene copies to unrelated sites in the genome. Genes found in only one copy are excluded from the study, because they may have multiple origins besides being the only remaining members of a gene family. For example, they may have originated after the last genome duplication, they may have diverged rapidly, or insufficient effort may have been made to uncover similar genes. (The authors argue that the latter problem is less pertinent in multigene families, because the discovery of two members of a family motivates the search for further members.) Excluding single- copy genes, the rate of gene loss can still be inferred from the ratio of two-, three-, and four-gene families.

The statistical analysis carried out by the authors revolves around the estimation of the parameter of interest, the (cumulative) probability ψ that a duplicated gene eventually loses its function, i.e., after a sufficient amount of time has passed after the last duplication event. Two major scenarios are conceivable for each duplicated gene, depending on the (unknown) amount of time separating the two duplications. In the first, simpler, scenario, the time interval between the two duplications is so short that neither gene loss nor divergence can occur. In this case, each gene before the duplication gives rise to a family of four genes after the duplication.

Assuming no deleterious effects of changing gene dosage, up to three members of this family can suffer a loss-of-function. Given ψ, one can then calculate the probabilities that four, three, or two genes survive long after the second duplication. These quantities are calculated as $(1 - \psi)^4/(1 - \psi^4)$, $4\psi(1 - \psi)^3/(1 - \psi^4)$, and $6\psi^2(1 - \psi)^2/(1 - \psi^4)$, respectively.[1] In the second scenario, sufficient time has passed between the duplications for gene loss or divergence of duplicated genes. Thus, the two genes resulting from the first duplication may have completely diverged in function. In this case, the second duplication will produce two two-gene families with different functions. One gene within each family can be lost, so that after sufficient time has elapsed after the second duplication, two, three, or four genes remain. Again, one can express the probability of each of these three outcomes as a function of ψ. Finally, the authors consider two further scenarios for long interduplication intervals that might best be described as ‘‘mixed’’ models. These models incorporate more complications, e.g., allowing for the possibility that not all genes diverge at the same rate. Again, expressions for the probability to observe a gene family of two, three, or four genes can be derived given ψ.

The data set used by Nadeau and Sankoff to estimate the value of ψ is the number of genes in each of the 276 human gene families (21, 67, and 188 families of four, three, and two genes, respectively) and the number of genes in each of the 176 human gene families (11, 45, and 120 families). They estimate ψ via a maximum likelihood procedure for each of the four scenarios, and for the mouse and human data set separately. Comparing the maximum likelihoods obtained for each of the four scenarios should also allow one to decide which of the four scenarios is best supported by the data. Perhaps not surprisingly, the available data were not sufficient to answer this question, except to say that the scenario involving very short interduplication times is least supported. Most importantly, however, regardless of which scenario was used to estimate ψ, and regardless of whether the mouse or human data were used, all maximum likelihood estimates of ψ ranged between 0.4 and 0.6. This robustness of estimates for ψ lends itself to the conclusion that the probability $1 - \psi$ of eventually becoming diversified in function may well be of the order of 50%, as estimated in the earlier studies. There are a number of implicit assumptions behind the analysis, such as that of a constant ψ, and the assumption that loss of a gene function is an irreversible all-or-none process (see below and ref. 18). However, the consistency of the main result with earlier, independent work is reassuring, and the study clearly points toward avenues of further investigation.

Why the low rate of gene loss?

Assuming that population sizes and mutation rates for the lineages under study are not atypical, one is lead to conclude that population genetic theory seriously underestimates the probability of functional divergence of duplicated genes. Why does it not predict higher diversification rates? The reasons may lie in the particular perspective that most population genetic models take on genes, and in the very simplification that make population genetics a powerful body of theory. In most population genetic models, alleles at a gene locus enter a model only through their frequencies in a population and through their contribution to an individual’s fitness. This also holds for the few models that have addressed the problem of functional divergence: Mutations in duplicate genes that lead to a loss-of-function are neutral or deleterious, and mutations that lead to divergence are neutral or beneficial. This is the only distinction made between the two outcomes, and no substructure is superimposed on the evolving genes.

This perspective, in which the gene is the ‘‘atom’’ of the evolutionary process originated at a time where the biochemical nature of genes was poorly understood. It may be an adequate level of resolution for many purposes. However, molecular biology has demonstrated that genes and their products have a complex substructure that most population genetic modeling has not yet captured. Two observations are particularly germane. First, there is an increasing number of genes that appear to encode multifunctional proteins, whose functions may be affected separately by mutation. Here, function can be defined both biologically, e.g., acting in two different developmental pathways,[20] or biochemically, e.g., regulating both transcription and translation.[21] Second, some of these functions may be redundant.[22,23] There may be groups of genes whose range of biochemical functions overlaps greatly, and such redundancy may provide an effective buffer against otherwise deleterious mutations.[24,25] Notably, the loci Nadeau and Sankoff[1] used in their study include not only enzyme loci, but also regulatory loci with demonstrated partial redundancies (e.g., *engrailed*[26]).

A duplicated gene may lose some of its functions while retaining others. If some of the retained functions are unique and essential, the gene will remain in the population. If other retained functions are redundant, they may be free to evolve new and unique functions. To complicate matters further, genes are usually embedded in complex metabolic and regulatory networks, and the network itself may buffer many mutations that would otherwise be deleterious.[27,28] Thus, the potential of neutral evolution to generate a reservoir of new functions that later becomes useful may currently be greatly underestimated. It is these features, the rich substructure of genes and their embedding into superstructures—genetic networks—that the theory may need to capture to explain the abundance of functional diversification. A deeper understanding of evolutionary dynamics in this area is sorely needed.

Outlook

Any study of the sort discussed here, using a small data set (less than 1% of the genomic gene content) assembled from multiple primary sources, raises unavoidable questions about potential biases in the data. For example, could some of the genes have been part of small gene families before the first duplication? Given that the genome duplications in question are very old, could some duplicate members of a family have evolved beyond recognition? Do the genetic and biochemical methods used to identify the studied genes generate bias in the size distribution of gene families? For example, a genetic screen will more readily reveal a gene with a partially redundant copy (a two-gene family) than a member of a four-gene family with one or more completely redundant members. Such biases may be unavoidable given the state of our knowledge of vertebrate gene families. However, the ongoing eukaryotic genome projects will improve this situation. The torrent of information, both genomic and expressed (ESTs), produced by large scale sequencing will provide much larger and bias-free samples of existing gene families. Such samples will certainly help decide the question at issue here. They may also help date particular genome duplication events which shaped not only the genomes of higher vertebrates, but may also have enabled the manifold morphological innovations found in this taxonomic group.

References

1 Nadeau JH, Sankoff D (1997) Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics* 147:1259–1266.
2 Ohno S (1970) *Evolution by Gene Duplication*. New York: Springer.
3 Haldane JBS (1933) The part played by recurrent mutation in evolution. *Am Nat* 67:5–19.
4 Kimura M, King JL (1979) Fixation of a deleterious allele at one of two “duplicate” loci by mutation pressure and random drift. *Proc Natl Acad Sci USA* 76:2858–2861.
5 Maruyama T, Takahata N (1981) Numerical studies of the frequency trajectories in the process of fixation of null genes at duplicated loci. *Heredity* 46:49–57.
6 Clark AG (1994) Invasion and maintenance of a gene duplication. *Proc Natl Acad Sci USA* 91:2950–2954.
7 Watterson GA (1983) On the time for gene silencing at duplicate loci. *Genetics* 105:745–766.
8 Nei M, Roychoudhury AK (1973) Probability of fixation of nonfunctional genes at duplicate loci. *Am Nat* 107:362–372.
9 Baston CJ, Ohta T (1992) Simulation study of a multigene family, with special reference to the evolution of compensatory mutations. *Genetics* 13:247–252
10 Ohta T (1987) Simulating evolution by gene duplication. *Genetics* 115:207–213.
11 Ohta T (1988) Further simulation studies on evolution by gene duplication. *Evolution* 42:375–386.
12 Ohta T (1988) Time for acquiring a new gene by duplication. *Proc Natl Acad Sci USA* 85:3509–3512
13 Walsh JB (1995) How often do duplicated genes evolve new functions? *Genetics* 139:421–428.
14 Li E-H (1997) *Molecular Evolution*. Sunderland, Massachusetts: Sinauer.
15 Ferris SD, Whitt GS (1976) Loss of duplicate gene expression after polyploidisation. *Nature* 265:258–260.
16 Ferris SD, Whitt GS (1979) Evolution of the differential regulation of duplicate genes after polyploidization. *J Mol Evol* 12:267–317.
17 Allendorf FW, Utter FM, May BP (1975) Gene duplication within the family Salmonidae. II. Detection and determination of the genetic control of duplicate loci through inheritance studies and the examination of populations. In Markert CL (ed): *Isozymes*. Vol IV: *Genetics and Evolution*. New York: Academic Press, pp 415–432.
18 Marshall CR, Raff EC, Raff RA (1994) Dollo’s law and the death and resurrection of genes. *Proc Natl Acad Sci USA* 91:12283–12287.
19 Ohno S (1985) Dispensable genes. *Trends Genet* 1:160–164.
20 Doe CQ, Hiromi Y, Gehring WJ, Goodman, CS (1988) Expression and function of the segmentation gene fushi tarazu during *Drosophila* neurogenesis. *Science* 239:170–175.
21 Ladomery M (1997) Multifunctional proteins suggest connections between transcriptional and post-transcriptional processes. *BioEssays* 19:903–909.
22 Tautz D (1992) Redundancies, development and the flow of information. *BioEssays* 14:263–266.
23 Thomas JH (1993) Thinking about genetic redundancy. *Trends Genet* 9:395–398.
24 Nowak MA, Boerlijst MC, Cooke J, Maynard-Smith J (1997) Evolution of genetic redundancy. *Nature* 388:167–171.
25 Wagner A (1998) Redundant gene functions and natural selection. *J Evol Biol* (in press).
26 Joyner AL, Herrup K, Auerbach BA, Davis CA, Rossant J (1991) Subtle cerebellar phenotype in mice homozygous for a targeted deletion of the en-2 homeobox. *Science* 251:1239–1243.
27 Hartl DL, Dykhuizen DE, Dean AM (1985) Limits of adaptation: The evolution of selective neutrality. *Genetics* 111:655–674.
28 Wagner A (1996) Does evolutionary plasticity evolve? *Evolution* 50:1008–1023.