Gene duplication and the origin of novel proteins

Austin L. Hughes*

Department of Biological Sciences, University of South Carolina, Columbia, SC 29205

Evolutionary biologists agree that gene duplication has played an important role in the history of life on Earth, providing a supply of novel genes that make it possible for organisms to adapt to new environments (1). The existence of diverse multigene families, particularly in eukaryotes, provides evidence that numerous events of gene duplication followed by functional diversification have shaped genomes as we know them. But it is less certain how this panoply of new functions actually arises, leaving room for ingenious speculation but not much rigor. Cases where we can reconstruct with any confidence the evolutionary steps involved in the functional diversification are relatively few. Thus the report in this issue of PNAS by Tocchini-Valentini and colleagues (2) on tRNA endonucleases of Archaea is particularly welcome as a concrete example of how new protein functions can arise.

New Proteins for Old Functions

The first hypothesis regarding the origin of new gene function was that of Ohno (3), who assumed that, after duplication, one gene copy would be entirely redundant and thus freed from all constraint. This redundant copy would become a nonfunctional pseudogene in most cases. But occasionally, Ohno postulated, such a gene would reemerge from nonfunctionality with a new function acquired as a result of chance mutations. There are a number of reasons for doubting this hypothesis. First, as the late Marianne Hughes and I (4) showed in the case of the tetraploid frog *Xenopus laevis*, duplicate genes are not in general freed from all functional constraint. Rather, purifying selection acts to eliminate deleterious nonsynonymous (amino acid-altering) mutations even in apparently redundant gene copies. Furthermore, there are a number of multigene families where there is evidence that positive Darwinian selection has acted to promote amino acid changes in functionally important regions of proteins (5). In these families, new function clearly has not arisen as a result of random mutation alone, contrary to the prediction of Ohno's model.

An alternative to Ohno's hypothesis is that both functions are already present before gene duplication. Goodman *et al.* (6), observing the homology between the α and β chains of hemoglobin, hypothesized that the ancestral hemoglobin molecule, before the gene duplication that gave rise to separate α and β chain genes, was a homotetramer. Jensen (7), observing the "substrate ambiguity" of many contemporary enzymes, proposed that after duplication of an ancestral gene encoding an enzyme of broad specificity, the daughter genes may divide the work of the ancestor, encoding enzymes of narrower specificity. Indeed, Orgel (8) suggested that sharing of an ancestral function between duplicates was the ordinary way that diversification occurred among members of multigene families.

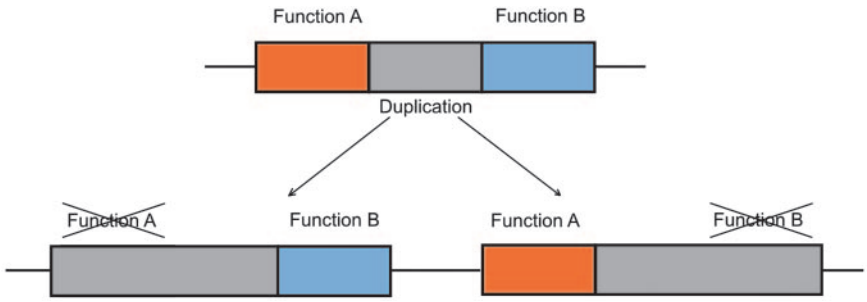

Fig. 1. Schematic illustration of subfunctionalization through complementary loss of function in two duplicate genes.

All of this remained highly speculative until the work of Piatigorsky and Wistow (9) on eye lens crystallins revealed the remarkable phenomenon of "gene sharing," whereby a single gene encodes a molecule that is a functional enzyme but, when expressed in the eye, serves as a crystallin. A wide variety of proteins have been independently recruited as lens crystallins over the course of animal evolution. Moreover, duplication of such "shared" genes has made it possible for one gene to specialize in encoding the enzyme, whereas the other specializes in encoding the crystallin. The phenomenon of gene sharing followed by duplication and subsequent specialization seemed to support the model of Jensen (7) and Orgel (8), at least in this one perhaps atypical case, but it suggested the possibility that the phenomenon might be more widespread (5, 10).

Two articles by Lynch and colleagues (11, 12) added further theoretical wrinkles to the discussion. Not only did these articles introduce a widely used term, "subfunctionalization," for the process by which daughter genes come to divide up the functions of an ancestral gene, but also they pointed out a mechanism by which subfunctionalization might occur. This mechanism involves complementary loss of function by the daughter genes. Imagine an ancestral gene that performs two functions, A and B (Fig. 1). After gene duplication giving rise to two new genes (genes 1 and 2; Fig. 1), it may happen, for example, that gene 1 loses function A but retains function B, whereas gene 2 loses function B but retains function A. Such a complementary loss of function can be a powerful factor in promoting the retention of duplicate genes in the genome. Once this process has occurred, deletion of either duplicate gene will be selectively disadvantageous, because both genes are required to perform the functions once performed by the ancestral gene.

Archael tRNA Endonucleases

Tocchini-Valentini *et al.*'s study (2) provides two apparent examples of subfunctionalization by this general scenario. They show that in Archaea there are three distinct forms of tRNA endonuclease, the enzyme responsible for cleaving the intron from pre-tRNA and pre-rRNA. Crystal structures of this enzyme are known from two members of the Euarcheota, *Archaeoglobus fulgidus* and *Methanococcus jannaschii*. In *A. fulgidus*, the enzyme is a homodimer. Each subunit contains two similar repeating domains, but the same domain performs a different function in each of the two subunits. The C-terminal repeat is the catalytic domain, whereas the N-terminal repeat acts to stabilize the dimer. There is a similar functional subdivision in *M. jannaschii*, whose tRNA

See companion article on page 8933.

*E-mail: austin@biol.sc.edu.

endonuclease is a homotetramer. In the latter species, two of the subunits play a catalytic role, whereas the other two serve merely to maintain the structure of the homodimer.

Because tRNA endonucleases had not been described from the Crenoarcheota, the other main subdivision of Archaea, Tocchini-Valentini *et al.* (2) used bioinformatic techniques to search the genome sequence of the crenoarcheote *Sulfolobus solfataricus* for homologs to the *M. jannaschii* protein. They found two such sequences in *S. solfataricus*, which they named α and β.

Presumably the ancestral situation in Archaea was similar to that in *M. jannashii*, with a single gene encoding a single protein that can serve as either catalytic unit or structural spacer as circumstances dictate. In the evolution of *A. fulgidus*, the gene evidently duplicated; the two duplicate genes fused, while subfunctionalization occurred between the two halves of the fused protein, with the C-terminal half retaining the catalytic function and the N-terminal half specializing as a spacer. Tocchini-Valentini *et al.* (2) tested this hypothesis by an experiment that attempted to reverse the evolutionary process. They cut the *A. fulgidus* tRNA endonuclease gene into two segments, expressed each segment separately, and showed that they could combine to form a functional enzyme.

In *S. solfataricus*, Tocchini-Valentini *et al.* (2) hypothesized that subfunctionalization occurred after gene duplication. One of the two genes has specialized for encoding the catalytic subunit and the other for encoding the spacer, and the two are believed to form a heterotetramer consisting of two α subunits and two β subunits. Tocchini-Valentini *et al.* provided evidence that both α and β subunits are necessary for substrate cleavage.

The Role of Natural Selection

One theoretically attractive feature of their model of subfunctionalization, as pointed out by Lynch and colleagues (12), is that it can occur without the need for positive Darwinian selection, which is thought to be relatively rare at the molecular level (1). In the simple scenario illustrated in Fig. 1, a mutation eliminating function A in gene 1 might become fixed in a population by genetic drift. Once this happens, conservative or purifying natural selection will act against any mutation that eliminates function A from gene 2. Conversely, a loss of function B in gene 2 can become fixed by drift as long as gene 1 retains function B. Of course, the fixation of these loss-of-function mutations will be facilitated if the effective population size is not very large.

On the other hand, some of the best-documented examples of positive Darwinian selection at the molecular level involve functional diversification among members of multigene families, for example, Ig V region genes (13) and defensins (14). It may often be as true of molecules as it is of human beings that "a jack of all trades is master of none." In such cases, positive selection may actually favor the loss of one function in a bifunctional molecule if a duplicate gene is able to take up the slack. In our simple scenario, if gene 1 loses function A it may be better able to perform function B, whereas if gene 2 loses function B it may be better able to perform function A. We might call this the "Babe Ruth effect." As baseball fans know, Ruth was a great pitcher early in his career (1915–1917) and a great outfielder and hitter later (1919–1934). The transitional year, 1918, when Ruth played both positions for the Boston Red Sox, was a mediocre one, at least by Ruth's standards.

In the case of archaeal tRNA endonucleases, there is no direct evidence whether drift alone gave rise to subfunctionalization or whether positive selection played a role. These events occurred in the distant past; thus, the most convincing signal of positive selection, an accelerated rate of nonsynonymous nucleotide substitution (10) is not obtainable, being obscured by numerous subsequent neutral changes. However, the fact that subfunctionalization has occurred twice independently and by different pathways in the same gene family suggests that positive selection may indeed have been involved. Perhaps, in the high-temperature environments occupied by these archaeal species, there is something less than optimal about the homotetrameric type of tRNA endonuclease, where the same polypeptide does double duty as a catalytic subunit and a spacer.

Evolutionary System Biology

If we have learned anything at all in a century and a half of evolutionary biology, it is that facile generalizations are dangerous. The evolutionary process finds a way to create exceptions to every model we propose. Thus, it seems unwise to expect that functional diversification after gene duplication follows the same pathway every time. Sometimes, subfunctionalization may occur by drift alone. On other occasions, as we know, positive selection is involved. Perhaps there are even cases where a new function has arisen by Ohno's model of resuscitation of a dead gene (3).

Whatever the mechanism by which subfunctionalization arose, Tocchini-Valentini *et al.*'s (2) study confirms the pioneering insight of Jensen (7): when bifunctionality or multifunctionality precedes gene duplication, it is straightforward for duplicates to specialize by sharing the ancestral function. In fact, as recent data on gene expression and protein–protein interaction networks make clear, all genes are multifunctional. Even in its infancy, systems biology makes clear that protein functions are complex processes existing in multiple dimensions (15). It thus seems a reasonable extension of Jensen's original insight to propose that new protein functions arise as the multidimensional space of functional interactions is parceled out in new ways, new links in biological networks are formed, and old links are broken.

Testing this hypothesis will require work at the interface of molecular evolutionary genetics and systems biology. We will need to be able to understand the diversification of gene duplicates in terms of the totality of each gene's role in cellular processes. It is a tall order given our present knowledge, but this kind of evolutionary systems biology not only will increase our understanding of how new protein functions evolve but also will shed essential light on why biological systems work the way they do.

1. Kimura, M. & Ohta, T. (1974) *Proc. Natl. Acad. Sci. USA* **75,** 6168–6171.
2. Tocchini-Valentini, G. D., Fruscolini, P. & Tocchini-Valentini, G. P. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 8933–8938.
3. Ohno, S. (1973) *Nature* **244,** 259–262.
4. Hughes, M. K. & Hughes, A. L. (1993) *Mol. Biol. Evol.* **10,** 1360–1369.
5. Hughes, A. L. (1994) *Proc. R. Soc. London Ser. B* **256,** 119–124.
6. Goodman, M., Moore, G. W. & Matsuda, G. (1975) *Nature* **253,** 603–608.
7. Jensen, R. A. (1976) *Annu. Rev. Microbiol.* **30,** 409–425.
8. Orgel, L. E. (1977) *J. Theor. Biol.* **67,** 773.
9. Piatigorsky, J. & Wistow, G. (1991) *Science* **252,** 1078–1079.
10. Hughes, A. L. (1999) *Adaptive Evolution of Genes and Genomes* (Oxford Univ. Press, New York).
11. Lynch, M. & Force, A. (2000) *Genetics* **154,** 459–473.
12. Lynch, M., O'Hely, M., Walsh, B. & Force, A. (2001) *Genetics* **159,** 1789–1804.
13. Tanaka, T. & Nei, M. (1989) *Mol. Biol. Evol.* **6,** 447–459.
14. Hughes, A. L. (1999) *Cell. Mol. Life Sci.* **56,** 94–103.
15. Uetz, P. & Finley, R. L., Jr. (2005) *FEBS Lett.* **579,** 1821–1827.