Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 7506–7509, August 1994
Evolution

Evidence for intron capture: An unusual path for the evolution of proteins

G. BRIAN GOLDING*, NORA TSAO†, AND RONALD E. PEARLMAN†

*Department of Biology, McMaster University, Hamilton, ON Canada L8S 4K1; and †Department of Biology, York University, North York, ON Canada M3J 1P3

Communicated by John R. Preer, Jr., May 2, 1994

ABSTRACT **Most new genes are thought to evolve from preexisting genes but duplications of entire genes or shuffling of preexisting exons provides only a limited repertoire of new sequences that can be presented to a cell. Only pieces that previously existed can be used in the construction and any further divergence depends on the slow accumulation of mutations. We show here the presence of a small, in-frame intron in a ciliate phosphoglycerate kinase gene and the insertion of an unusually random amino acid sequence at the same position in trypanosome phosphoglycerate kinase. The unusual sequences in trypanosomes were likely to have originally been introns that have been subsequently captured by the protein and have now been incorporated as part of the coding sequence. Via this path a truly unique sequence can be incorporated into an existing protein, leading in time to the evolution of a new, functionally distinct protein.**

Addressing questions about molecular mechanisms for the evolution of new genes has been an active area of research since Ohno (1) demonstrated that duplication is a general process for the generation of new protein sequences. This process can operate at several levels ranging from simple local duplications to polyploidy. The concept that most modern genes have originated from distinct, preexisting genes has been confirmed by many studies, which have identified mechanisms such as gene duplication (1), exon shuffling (2–4), duplication of small repeats (5, 6), and horizontal transfer (7) in the evolution of new genes. Although gene duplication is most common, the duplicated gene regions are generally unstable to the effects of unequal crossing-over and gene conversion. Another problem is that the duplicated gene is not significantly different from the original gene. To change the function of one copy requires the accumulation of many individual mutations.

Exon shuffling, where a new gene contains different segments encoding preexisting functional domains (2), eliminates some of these problems. By combining different domains into a single protein, the creation of more distinctly unusual proteins is possible, while still retaining a high likelihood of functionality. The demonstration that some proteins have exons that encode functional domains (8) supports this hypothesis. The existence of introns and consequently ideas about their origin and evolution have been tied to exon shuffling (2, 9). Data to support the hypotheses that introns are ancient and have been subsequently lost in bacteria (10) or that introns are recent and have been inserted into eukaryotic genes (11) have been presented. Whatever the course of their evolution, there is evidence that introns have permitted some genes to be constructed from combinations of exons of other genes such as has been described for the human low density lipoprotein receptor (11).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Both gene duplication and exon shuffling avoid random sequences in the creation of new proteins and have a high likelihood of generating functional genes. These processes are, however, limited in their ability to generate truly unusual gene sequences. In the case of gene duplication, the duplicated copy is identical to the original gene. With exon shuffling, the new gene will be limited to combinations of those sequences present in the preexisting exons (4). One mechanism that has the potential to generate truly unique proteins depends on the presence of short tandem repeats. Ohno (12) showed that such repeats enable the protein encoding 6-aminohexanoic acid linear-oligomer hydrolase in *Flavobacterium* to be generated from an alternative reading frame of another gene. Although this leads to a fairly random collection of nucleotides, an open reading frame is not very likely without very specific oligomeric repeats. Hence, this is not likely to be a common mechanism for the evolution of new genes. Horizontal transfer can bring in unique genes but this simply removes the problem of its evolution to another organism and, in addition, horizontal transfer does not appear to be a common phenomenon.

Another path for the evolution of new genes makes use of intron sequence. It was shown by Butticè *et al.* (13) that collagen genes in mice have a protein coding sequence that may have originally been an intron sequence. This pathway mixes the advantages of using "random" sequence addition to generate new proteins while still retaining most features of the original protein, leading to a high likelihood of a functional product. If the gene is an essential enzyme, a prior duplication is a prerequisite for evolution of a new function via intron capture. Here we present further evidence for this pathway for the evolution of new genes. We show that the phosphoglycerate kinase (PGK; GTP:3-phospho-D-glycerate 1-phosphotransferase, EC 2.7.2.10) gene has a small, in-frame intron in *Paramecium* and that at this exact location in two species of trypanosomes, there are amino acid insertions. These inserted sequences evolve at a rapid rate, which suggests that they may once have been introns and have been captured and incorporated into a functional protein, but they have not yet contributed to the evolution of new function.

MATERIALS AND METHODS

PGK is a monomeric protein that catalyzes the conversion of 3-phospho-D-glycerate to 1,3-diphospho-D-glycerate. We have determined the nucleotide sequence of the gene encoding PGK from *Paramecium primaurelia* as well as the cDNA sequence from this species.

Sequences encoding the genomic PGK from *P. primaurelia* were amplified in a PCR with total DNA (kind gift of F. Caron, Ecole Normale Superieure, Paris) and degenerate primers designed to anneal to highly conserved sequences near the N and C termini of PGK. The upstream primer was 5′-GAATTCGTNGAYTTYAAYGTNCCN-3′ and the

Abbreviation: PGK, phosphoglycerate kinase.

```
A  Trypanosoma brucei^A      ...LGRPRGVSMVAGSALPQMSVSAGYEEQYSLRPVAVR...
   Trypanosoma brucei^B      ...LGRPKGVSMAEGKELGSAGGIPGFEQKATLKPVAKA...
   Trypanosoma brucei^C      ...LGRPKGIPMAQAGKIRSTGGVPGFQQKATLKPVAKA...
   Crithidia fasciculata^A   ...LGRPTGVDYDTA-ELFPQLGSAGYEKEFSLEPVAVK...
   Crithidia fasciculata^B   ...LGRPKGAKMSDPK---PAKSVRGYEEAATLRPVAAR...
   Crithidia fasciculata^C   ...LGRPKGAKMSDPK---PAKSVRGYEEAATLRPVAAR...
   Tetrahymena thermophila   ...LGRPDGNR----------------VEKHSMKPVVPK...
   Paramecium primaurelia    ...MGRPDGKR----------------VEKDSLKIVVPK...
                                                  ^

B  Paramecium intron -       GATGGAAAAAGgtaatataagatattaatttatagAGTTGAA
                              D  G  K  R                        V  E
```

FIG. 1. (*A*) Amino acid sequence alignment of a portion of the PGK proteins from *T. brucei* A, B, and C isozymes, *C. fasciculata*, *T. thermophila*, and *P. primaurelia*. Alignment is essentially as presented by Vohra *et al.* (19) with the addition of data for *P. primaurelia* obtained in this study. Position of 24-nt intron in the genomic DNA of *P. primaurelia* is indicated with an arrowhead. (*B*) Nucleotide sequence of the intron in *P. primaurelia* with some sequence 5′ and 3′ of the predicted splice junctions. There are conserved GT and AG dinucleotides in the intron at 5′ and 3′ splice junctions, respectively. The intron splits the arginine codon with 2 nt of this codon in exon I and 1 nt in exon II.

downstream primer was 5′-GGATCCNGCNCCNCCNC-CNGT-3′. The amplified product was cloned into pEMBL18 and -19 (14) and sequenced by the dideoxynucleotide chain-termination method (15). The nucleotide sequence was translated by the program ANALYSEQ (16).

To check for the presence of a potential intron, a region of the nucleotide sequence was checked from messenger RNA. Total RNA from *P. primaurelia* was reverse transcribed with a *Paramecium*-specific primer (5′-CGACCTTAAGGTC-CATTCC-3′) downstream of the putative intron. Reverse transcription was carried out according to Sambrook *et al.* (17). The DNA·RNA hybrid was then PCR amplified with the downstream primer and the *Paramecium*-specific primer (5′-TGTCTCATATGGGTAGACC-3′). The primers were chosen to flank a putative 8-amino acid insertion in genomic DNA. PCR amplification was done by the "touchdown" procedure (18) at 94°C for 4 min with 1 unit of AmpliTaq DNA polymerase (Perkin–Elmer/Cetus) added at 2 min; 94°C for 30 sec, 64°C for 2 min, 72°C for 1.5 min for 1 cycle; 94°C for 30 sec, 62°C for 2 min, 72°C for 1.5 min for 2 cycles; 94°C for 30 sec, 60°C for 2 min, 72°C for 1.5 min for 2 cycles; 94°C for 30 sec, 58°C for 2 min, 72°C for 1.5 min for 30 cycles; and 72°C for 10 min for 1 cycle in a Coy thermal cycler. The region covering the putative intron was sequenced directly from the double-stranded PCR product without cloning.

The gene encoding PGK has been sequenced from >26 species covering a broad range of organisms from humans to archaebacteria (19). The nucleotide sequence from another ciliate *Tetrahymena thermophila* is included in the above data and sequences of five additional PGK genes are now available. All of the amino acid sequences were collected from published sources and amino acid alignments were prepared as described by Vohra *et al.* (19).

RESULTS AND DISCUSSION

We have used PGK as a phylogenetic marker in our ongoing studies on the evolution of ciliated protozoa (19). Determination of the nucleotide sequence of the genomic copy of PGK from *P. primaurelia* and alignment of the derived amino acid sequence with that of other ciliate PGKs revealed a unique, 8-amino acid insertion (Fig. 1) at approximately position 90 according to the alignment presented previously (19). Analysis of the sequence encoding this insertion revealed the presence of 5′ GT and 3′ AG dinucleotides, suggesting the possibility that this was in fact an in-frame intron. The sequence TTAAT has been found in two other small in-frame introns in *Paramecium tetraurelia* (20) and is also present in this insertion. This TTAAT sequence has been implicated in the processing of these short introns (20). The cDNA sequence and its alignment with the genomic sequence confirmed the existence of a 24-bp in-frame intron (Fig. 1).

Most organisms have a single functional PGK gene but PGKs in *Trypanosoma brucei* and *Crithidia fasciculata* are part of a three-member multigene family. The A and C forms are localized in the glycosome and the B form is localized in the cytosol (21). Trypanosome PGK sequences have an unusual amino acid insertion at a position homologous to that of the *Paramecium* intron. In the PGK A forms there is a 94-amino acid insertion at this position (22), and the B/C forms have an identically positioned 13- to 16-amino acid insertion (23). These amino acid insertions have not been found in any other PGK sequences. Two methods can be used to show that this region has unusually rapid rates of substitution. Fig. 2*A* presents the total number of substitutions along each branch of the trypanosome phylogeny (the tree length) for a sliding window of 20 amino acids. In addition, sequences within each window were randomly

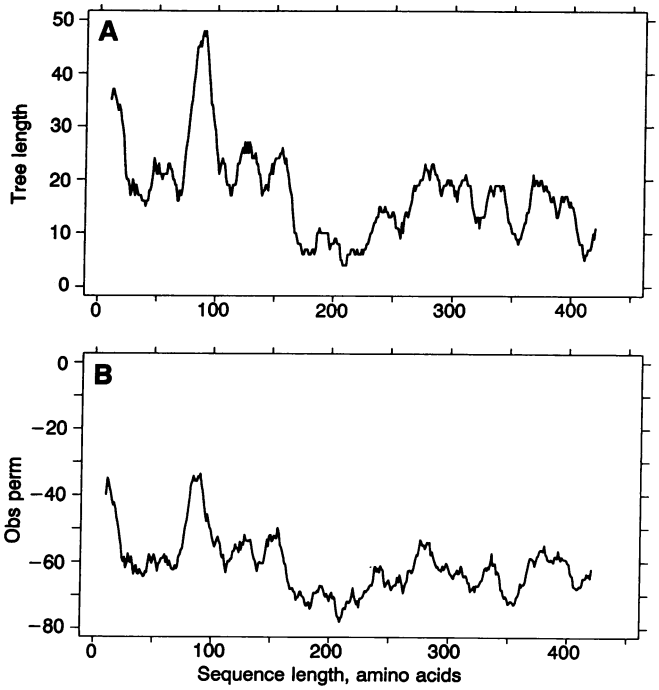

FIG. 2. Tree length (*A*) and average difference between observed and permuted tree lengths (Obs perm) (*B*) for the aligned PGK proteins from six isozymes of flagellates. As in Fig. 1, the large insert in the A isozymes has been deleted for this analysis. The gene tree was inferred by using the neighbor joining algorithm of Saitou and Nei (24). Given this tree, a window of 20 residues is moved along the sequence and for each window the number of substitutions required (the tree length) is calculated. The maximum parsimony algorithm of Fitch (25) is used to infer the number of substitutions along each branch of the phylogeny.

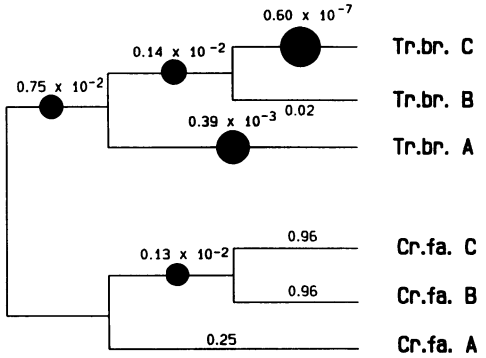

FIG. 3. The gene tree inferred for the flagellate PGKs with the probabilities that the number of substitutions observed within a 20-amino acid window centered on position 90 could be due to chance (given an expectation based on the number of substitutions in the remainder of the gene). Circles illustrate relative magnitude of the probabilities. Cr.fa., *C. fasciculata*; Tr.br., *T. brucei*.

permuted and the tree length was recalculated (Fig. 2*B*). This provides a measure of how different the sequence is from a random sequence relative to its phylogenetic history. The combination of these two methods permits a rapid and interactive exploration of multiple sequence alignments.

The results in Fig. 2 are for the six trypanosome PGK genes shown in Fig. 1 (the large unique insertion in the A form has been omitted from the figure). A large peak in tree length is very noticeable centered around position 90. This is the location of the insertion that has no counterpart in a PGK of any other organism yet examined (19). The size of the peak demonstrates rates of substitutions that are significantly greater than those for the rest of the protein. The apparent peak at the N terminus is due to the poor nature of the alignment at the beginning of the gene.

The large number of substitutions in this region has nearly eliminated all traces of phylogenetic ancestry. Fig. 2*B* shows the average difference between the observed tree length and the tree length for 50 random sequence permutations. Again, there is a clear peak around position 90. Indeed most of the phylogenetic structure to the sequences is due to the B and C forms in *Crithidia*, which are identical in this region. If one of these sequences is excluded from the analysis then there is little difference between randomized sequences and the actual observed PGK sequences in this region.

The excessive number of substitutions is not limited to one branch of phylogeny but rather pervades most branches. The probabilities of observing such high rates of amino acid replacement are shown in Fig. 3. Here the rate of change in a 20-amino acid window centered on position 90 is compared to the rate of evolution found in the remainder of the protein. Assuming the global rate of substitution within this small region, the probability of observing as many substitutions can be calculated. It is shown in Fig. 3 as a circle on the branch where the radius of the circle is inversely proportional to the probability. In most cases, the probability of such an excess of substitutions is very small for each of the branches as indicated by the large width of the circles.

The probabilities in Fig. 3 are based on the assumption that the number of differences within each window will be binomially distributed. This may not be the case. To ensure that the region centered on this insert is evolving at a rapid rate we have compared the rate of evolution within this region to the rate of evolution elsewhere in the protein. The frequency histograms in Fig. 4 show the actual distribution of changes. The changes observed for the insert are shown by a solid box. Changes observed from other regions of the protein are shown as stippled boxes. It will again be observed that an excess number of substitutions occur within this region compared to other windows. This is the case along most lineages of descent (excepting those that lead to the near identical B/C forms in *Crithidia* and the A form). The large number of changes along each branch suggests that this region does not differ from random permutations of the sequence. There are also an excess number of substitutions between the A forms of the two species, but this sequence has not yet been completely randomized. The insertions are not introns since Alexander and Parsons (21) have shown that the A form of *T. brucei* PGK can be distinguished from other *Trypanosoma* PGK proteins on the basis of its size on immunoblots.

Small (27 bp) in-frame introns have also been described in the β-tubulin gene of *P. tetraurelia* (20). These would also be candidates for eventual intron captures. While the majority of introns are quite large and serendipitously carry in-frame stop codons, a large deletion in the interior of an intron is easy to

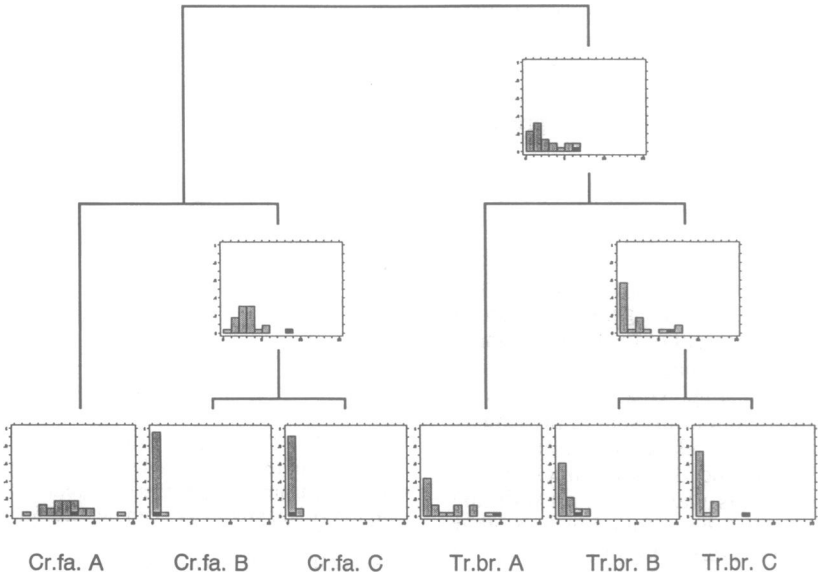

FIG. 4. Frequency histogram of the number of substitutions within sequential 20-amino acid windows along the length of the PGK gene. The numbers of substitutions for the window centered on residue 90 are shown as solid squares. (The *x* axis shows the number of substitutions within a window scaled from 0 to 15. The *y* axis shows the frequency of these windows scaled from 0.0 to 1.0.) Cr.fa., *C. fasciculata*; Tr.br., *T. brucei*.

engineer and may not be deleterious to the cell. The removal of in-frame stop codons opens the door for the intron to be captured as part of a larger protein. If the in-frame stop codons are present, the intron could not be incorporated and mutations to remove them must be awaited. Thus, smaller introns are more likely to be readily incorporated.

We have demonstrated that a small in-frame intron in the PGK gene of one organism is at the exact location of an amino acid insertion in this protein in at least two other organisms. These insertions are so saturated with amino acid replacements that there is little difference between them and a random collection of amino acids. This suggests that the inserts may be the results of intron capture. Such a rapid rate of amino acid replacement is not typical of functional sequence and seems to preclude a functional aspect to these inserts at present. Therefore, although the intron has been captured, it has not yet evolved a function as part of the complete PGK molecule. Potentially, new functions could be added to trypanosome PGKs, or if the evolutionary experiment proves fruitless, the sequence might be lost again.

The authors thank Dr. F. Caron (Ecole Normale Superieure, Paris) for his kind gift of *P. primaurelia* DNA; Drs. P. Moens, M. Dobson, and R. Morton for discussion; and Dr. P. Borst for his helpful suggestions. G.B.G. is a Fellow of the Evolution Program of the Canadian Institute for Advanced Research and is supported by a Natural Sciences and Engineering Research Council of Canada grant. Work from R.E.P.'s laboratory is supported by the Medical Research Council of Canada.

1. Ohno, S. (1970) *Evolution by Gene Duplication* (Springer, New York).
2. Gilbert, W. (1978) *Nature (London)* **271**, 501.
3. Gilbert, W. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52**, 901–905.
4. Dorit, R. L., Schoenbach, L. & Gilbert, W. (1990) *Science* **250**, 1377–1382.
5. Ohno, S. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 7657–7661.
6. Blake, C. (1983) *Nature (London)* **306**, 535–537.
7. Smith, M. W., Feng, D.-F. & Doolittle, R. F. (1992) *Trends Biochem. Sci.* **17**, 489–493.
8. Go, M. & Nosaka, M. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52**, 915–924.
9. Doolittle, W. F. (1978) *Nature (London)* **272**, 581–582.
10. Cavalier Smith, T. (1985) *Nature (London)* **315**, 283–284.
11. Sudhof, T. C., Goldstein, J. L., Brown, M. S. & Russell, D. W. (1985) *Science* **228**, 815–822.
12. Ohno, S. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 2421–2425.
13. Butticè, G., Kaytes, P., D'Armiento, J., Vogeli, G. & Kurkinen, M. (1990) *J. Mol. Evol.* **30**, 479–488.
14. Dente, L., Cesarini, G. & Cortese, R. (1983) *Nucleic Acids Res.* **11**, 1645–1655.
15. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
16. Staden, R. (1990) *Methods Enzymol.* **183**, 193–211.
17. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY), 2nd Ed.
18. Don, R. H., Cox, P. T., Wainwright, B. J., Baker, K. & Mattick, J. S. (1991) *Nucleic Acids Res.* **19**, 4008.
19. Vohra, G. B., Golding, G. B., Tsao, N. & Pearlman, R. E. (1992) *J. Mol. Evol.* **34**, 383–395.
20. Dupuis, P. (1992) *EMBO J.* **11**, 3713–3719.
21. Alexander, K. & Parsons, M. (1991) *Mol. Biochem. Parasitol.* **46**, 1–10.
22. Le Blancq, S. M., Swinkels, B. W., Gibson, W. C. & Borst, P. (1988) *J. Mol. Biol.* **200**, 439–447.
23. Swinkels, B. W., Evers, R. & Borst, P. (1988) *EMBO J.* **7**, 1159–1165.
24. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4**, 406–425.
25. Fitch, W. M. (1971) *Syst. Zool.* **20**, 406–416.