Evolutionary History Written in Globin Genes

Genes that duplicate, jump, and diverge: these are the key elements in the evolution of globin gene clusters

The genes that code for globin, the protein component of hemoglobin, have been more thoroughly studied than any other family of genes in higher organisms for two reasons. First, because they appear to offer a very good model in which to investigate the operation of genes that switch on and off at different times in development. And second, because detailed knowledge of the globin family promised some useful insights into the nature of a series of clinically important blood diseases, the thalassemias.

While the results of this considerable worldwide effort have fulfilled many of the primary aims of the work, they have also yielded an unexpected bonus. It is now possible to sketch out in revealing detail the evolutionary history of the globin family over the past 500 million years. The history of the globin gene family contains many lessons about the molecular mechanisms of gene evolution, lessons that should be applicable to all genes of higher organisms.

These days it has become a truism to say that the genetic material of higher organisms is in a greater state of dynamic turnover than could ever have been predicted, and a very recent discovery about globin genes underlines this. "The globin locus is constantly pumping out copies of its genes and is bombarding the rest of the genome with them," says Philip Leder, formerly at the National Institutes of Health, Bethesda, and now at Harvard Medical School. "And what holds for globin must go for other genes too," he suggests.

Paradoxically, a second major feature of globin family history is stability, both in the structure of the individual genes and in the family as a unit. "The gene clusters remain stable over long periods of time," says Alec Jeffreys, of Leicester University, England, "but when change occurs, it seems to go in jumps."

Globin gene evolution, therefore, is apparently governed by a combination of a clear potential for change and a demonstrated property of long-term near constancy. Perhaps the jerkiness of the overall mode of evolutionary change in the globin family is a consequence of the resolution of these two counteracting characteristics.

Hemoglobin is a tetramer of two α-like and two β-like globin molecules. (Each five protein subunit cradles a heme group which is instrumental in combining with oxygen as the hemoglobin passes through the tissues.) Humans have five different β-like globins, ϵ, $^{G}\gamma$, $^{A}\gamma$, δ, and β, which are utilized during early embryonic life, fetal life, very early infancy, and adulthood. There are only three α-like genes, embryonic ζ and two adult α's.

The human α- and β-globin gene clusters are located on separate chromosomes (16 and 11, respectively), and the α-cluster is just half the size of its partner. In both clusters the genes are arranged along the chromosome in the order in which they are expressed in development.

Questions about globin gene evolution fall into two main areas. The first concerns the structure of the genes themselves. The second focuses on the structure of the gene clusters.

> "... what holds for globin must go for other genes too."

The similarities between α- and β-globin genes indicate that they arose from a single ancestral gene, probably by simple duplication. The detailed differences between the genes put the date of their divergence at around 500 million years ago—that is, at the dawn of vertebrate history.

Like most eukaryotic genes so far studied, globin genes are interrupted by noncoding regions, or introns. The α- and β-globin genes have two such intervening sequences, in homologous positions. (This is consistent with the genes' common origin.) Unlike the coding regions of genes, the intervening sequences are notorious for their rapid accumulation of mutations. The α- and β-globin introns conform to this picture, in that their sequences have diverged considerably. What is unusual, however, is that the sizes of the noncoding regions have shifted relatively little.

"The first intervening sequence is about 116 to 130 base pairs long in all mammal α- and β-globin genes so far studied, despite the extreme age of the β-globin gene duplication," says Jeffreys. "The second intron also varies little within the α- and β-globin gene families, although the α sequence is consistently shorter than the β." An obvious inference is that the size of the introns is important for some as yet unidentified function. Such a constraint does not appear to apply to most introns so far examined in other genes.

An intriguing tale, just recently unfolded, carries the clear implication that during their long history vertebrate globin genes lost an intron: there once were three, not two as we now see them.

Mitiko Gō, of Kyushu University, Japan, was interested in the notion that the coding regions of genes, the exons, correspond with structural domains in the resulting protein. This follows the speculations by Walter Gilbert of Harvard University, and Colin Blake of the University of Oxford, England, that genes are assembled Lego-like in evolution, from minigenes that correspond with structurally stable protein domains.

Although globin has no distinct protein domains, its amino acids are assembled in discrete compact regions, four in all, according to Gō. He showed that the first region corresponds with the first coding sequence, as does the last region with the third exon. But the middle two regions are coded for by a single exon, the middle one of the three. Perhaps "selection pressure eliminated the intervening sequence that was present in an ancestral gene," speculated Gō in his paper earlier this year [*Nature (London)* **291**, 90 (1981)].

The following month a second *Nature* paper offered some corroborating data determined from an unexpected source: the root nodules of leguminous plants. Nitrogen fixation is the prime business of root nodules, and the presence of a protein that readily mops up oxygen permits the process to proceed unhindered. That protein is leghemoglobin, a molecule that has for some time been recognized as

0036-8075/81/1023-0426$01.00/0 Copyright © 1981 AAAS

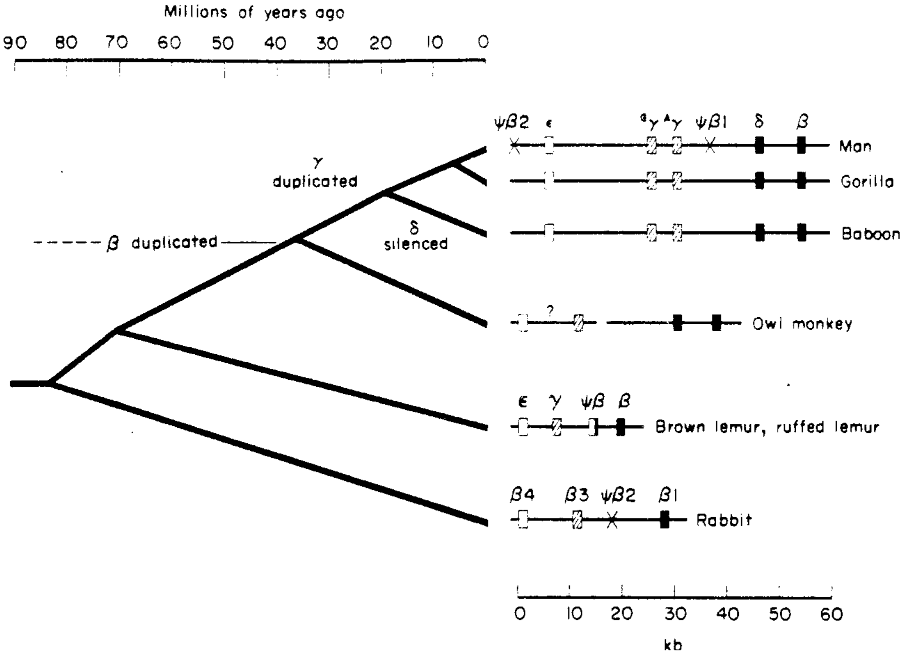
being very similar in structure to vertebrate hemoglobin. When K. A. Marcker and his colleagues at the University of Aarhus, Denmark, published their analysis of the leghemoglobin gene in June [*Nature (London)* **291**, 677 (1981)], it was clear that the similarity was not the product of convergent evolution, as had often been supposed. The vertebrate and plant genes were closely homologous, except that the leghemoglobin gene has three introns (and therefore four exons), not two.

Marcker concluded that their results did not support the idea that exons correspond with functional domains in proteins. What they did not know when they wrote this was that leghemoglobin's "extra" intron falls precisely at the division of the vertebrate globin's central exon into two compact regions, as revealed by Gō's analysis. A simple loss of leghemoglobin's extra intron would make the gene equivalent to vertebrate globin.

The section of the globin molecule that is coded for by the central exon is responsible for binding with the heme component of hemoglobin, with major contacts being distributed equally between what Gō has identified as the second and third compact regions of the protein structure. Perhaps way back in early evolutionary history there was a small heme-binding protein that was coded for by a gene with two exons. Later, further exons were added to give the primordial globin gene, which then lost the intron that divides the heme-binding region. The similarity of part of the structure of certain cytochrome proteins (which also bind with heme) to globin's central exon supports the notion of an ancient heme-binding gene, which then traveled along several evolutionary paths to give a family of different heme-binding proteins.

The tantalizing question is, how did legumes get their globin-like gene? "A common ancestry going back to the split between plants and animals seems out of the question," says Jeffreys. "Maybe there has been a horizontal transfer in recent times"; he suggests, "perhaps the gene has been carried on an insect-borne plant pathogenic virus." This intriguing possibility would imply that insect hemoglobin also has three introns and that the evolutionary line that led to vertebrates lost the "extra" intron before the primordial gene duplicated to establish the α and β groups. Analysis of insect globin is awaited with considerable interest.

The heroic devotion to the physical analysis of globin loci during recent years, particularly in the laboratories of Leder, Tom Maniatis, at the California Institute of Technology (now at Harvard), Oliver Smithies, University of Wisconsin, Jeffreys, and others, has opened the way to considerable armchair molecular archeology. "It is now possible to trace the evolution of the globin clusters, particularly the β-globin cluster, in some detail since the beginning of the mammalian era," says Jeffreys.



Phylogeny of the β-globin clusters

The maps of β-globin clusters reveal the sequence in which certain genes changed through evolutionary time. The striking stability of the cluster through the Old World monkeys, apes, and humans is indicated by the constant size and organization of the cluster. The positions of pseudogenes are shown by crosses.

The human β-globin cluster is the most complex yet analyzed. In addition to the five functional globin genes, the 60-kilobase locus also contains two so-called pseudogenes. Pseudogenes are relatively recent newcomers to molecular biology and are still somewhat enigmatic. Mostly they bear a close resemblance to a known gene, yet they are clearly disabled through additions or deletions in their structure that would prevent normal transcription and translation. Some people have suggested that pseudogenes might be important in regulating the activity of neighboring genes. Others consider them to be the diverged products of gene duplications, not necessarily relics of evolutionary change but, instead, potential new genes.

Leder points out that pseudogenes would be very fertile units for evolutionary change. "Once a duplicated gene is released from selection pressure it is free to diverge wildly in sequence," he says. "Pseudogenes have the great advantage of carrying with them all the necessary processing signals for normal generation, such as transcription initiation sites and splice junctions."

Pseudogenes have an even richer potential for evolutionary change than this description implies, as will become apparent later.

Even though the human β-globin complex contains a relatively large number of active genes, 95 percent of the locus is made up of DNA that does not code for proteins. What is the role of this extra DNA, if any? The pseudogenes constitute just a small proportion of the region, although more pseudogenes might exist. Some of the DNA is made up of representatives of well-known families of repetitive sequences. And the remainder is DNA of no known function or comparable sequence.

"We wanted to test the hypothesis that this extra DNA is 'junk DNA,' " says Jeffreys, "so we compared the β loci in humans, gorillas, and baboons." Jeffreys and his colleagues reasoned that if it were junk DNA, then over the 20 to 40 million years of evolution represented by humans, apes, and Old World monkeys both the sequence and the overall quantity of intergenic DNA could be expected to vary. "It turned out that the cluster is remarkably stable," reports Jeffreys. "The overall pattern and size of the cluster is the same, and the rate of nucleotide substitutions is one-quarter to one-fifth of what would be expected in functionless DNA." The noncoding DNA therefore appears not to be junk, but what function it might perform is still a mystery.

A step further back in evolutionary time, to 40 million years, when the New and Old World monkeys separated, suddenly presents a different picture. The ϵ and γ genes are much closer together; the γ gene has not yet duplicated; and the whole complex is substantially smaller.

Yet another step back into the past, to 70 million years ago, the beginning of the primate order and close to the explosive mammalian diversification, brings one to a very simple β cluster. The brown lemur (a prosimian) has just one each of the ϵ, γ, and β genes contained within a very small cluster. It also has one pseudogene. The similarity between the lemur cluster and that of the rabbit indicates that this arrangement is likely to reflect very closely the basic mammalian β-globin cluster, established perhaps 80 to 90 million years ago.

The increase in complexity of the β-globin cluster follows the conceptually simple pattern of gene duplication followed by sequence diversification of the new gene, eventually to produce a relative that acts at a different point in the developmental program. Hence, in humans, the original β gene is now accompanied by four closely related neighbors (plus the pseudogenes).

The story is not all simple and straightforward, however. For instance, although the γ gene duplicated some 20 to 40 million years ago, the sequences of the two genes in humans is so similar as to suggest a very recent separation. The cause of this anomaly, apparently, is the phenomenon of concerted evolution. Genes repeated in tandem, as the two γ genes are, exchange sequences by the action of a number of mechanisms, with the result that their structures remain very similar.

Another oddity is the absence of the δ-globin protein in Old World monkeys, even though a gene is correctly in place in the cluster. The gene has, it seems, recently been switched off. In other words, it is now a pseudogene.

One of the greatest curiosities of all is the lemur's pseudogene. The front, 5′, section of the gene resembles the ϵ gene, while the back, 3′, section is β-like. Jeffreys and his colleagues speculate that the correction mechanism has been in action twice, once with the β gene as the "model" and once with the ϵ gene as model, the result being a mosaic pseudogene.

The earliest events in globin gene evolution are more difficult to discern. The assumption that the primordial gene duplicated to give a tandem pair is supported by the simple arrangement of a single α and a single β gene as neighbors on one chromosome in the amphibian *Xenopus tropicalis*.

The next question is, how did the α and β clusters become separated on different chromosomes? Was a copy of one of the genes transported to another chromosome, where it was inserted, the original copy gradually falling into disuse? Or was one of the genes plucked out of its original position and relocated elsewhere? Or perhaps the whole chromosome set doubled up (tetraploidization) in an ancient ancestor, followed by the switching off of the α gene in one location and the β gene in the other? Support for this last mechanism comes from *X. laevis*, an evolutionary relative of *X. tropicalis*. The evolutionary event here did involve tetraploidization, so that *X. laevis* has two globin clusters, both of which contain a single α and a single β gene. Sequence divergence in this species appears to be silencing α and β genes in the same cluster. Nevertheless, tetraploidization is known to have been important in vertebrate evolution, and the differential gene silencing speculated upon is a firm possibility.

Gene evolution, as exemplified by the globin family, is a dynamic affair, with major changes occurring in stepwise manner.

Gene evolution, as exemplified by the globin family, is therefore seen to be a dynamic affair, with major changes manifesting themselves in stepwise manner. The source of the changes is principally gene duplication followed by sequence divergence to yield related, although distinctly different, members of the globin family. In addition to evolution within the locus, the globin cluster can be the source of new genes elsewhere.

Work at Leder's and Smithies' laboratories has in recent years revealed the existence of two α pseudogenes in the mouse. One of them was a conventional pseudogene, in that it has base changes that would prevent it from being expressed. The second one, however, is the most intriguing pseudogene yet discovered: its introns have been perfectly spliced out.

The removal of the introns in one of this pair of pseudogenes was the first surprise. The second emerged when Aya Leder attempted to discover exactly where in the globin cluster the pseudogenes were located. They weren't. They turned out to be on entirely different chromosomes, not in the α cluster as expected.

How did the pseudogenes move from their original location to different chromosomes? For a number of reasons, including the precise excision of introns, Philip Leder is attracted by Stephen Goff and David Baltimore's suggestion that the gene lacking intervening sequences was transported by the action of a retrovirus. There are sequences typical of retroviruses flanking the gene, but their arrangement is somewhat disrupted. It has to be possible that the insertion of the pseudogene and the passage of the retrovirus in that region were unconnected events, in which case some other explanation would have to be sought to explain the gene's movement with the loss of its introns.

The transposition of the second pseudogene could have occurred by a more conventional mechanism, possibly involving unequal crossing over, a process that for mechanical reasons is not uncommon in tandemly repeated genes.

In their search for the mouse α pseudogenes, the NIH workers increased the sensitivity of their detection system; instead of identifying just two, they now have evidence for as many as ten, although the location of these has not yet been determined. Leder likens the constant pumping out of pseudogenes by the globin locus to a volcano: "It's a Vesuvian model," he quipps. The implication is that all gene families are potential sources of new genes, through transposition, subsequent divergence of gene structure, and eventual recruitment to a new function.

It has already been demonstrated that histone gene families throw out gene copies that become integrated away from the main cluster (*Science*, 1 May, p. 530). This locus therefore appears to conform with the Vesuvian model. What is required now is a survey of gene structure that will reveal the echoes of evolutionary relatedness that must exist between many genes, if the model is correct in its evolutionary implications.

The globin genes have indeed proved to be a fertile source of data and a sharp stimulus for ideas in this rapidly growing field of molecular aspects of evolutionary change.—ROGER LEWIN