Evolution in multigene families

Introduction

We now know a lot about the dynamics of nucleotide substitutions within existing genes, but we've neglected one key component of molecular evolution. We haven't talked about where new genes come from. It's important to understand this phenomenon because, after all, new metabolic functions are likely to arise only when there are new genes that can perform them. It's not likely that an existing gene can adopt a new function while continuing to serve its old one.

Fundamentally the source of new genes is the *duplication* of existing genes and their *divergence* in function. As we'll see in a moment, for example, genes coding for myogblobin and hemoglobin in mammals are descendants of a single common ancestor. That's the duplication. Myoglobin is involved in oxygen metabolism in muscle, while hemoglobin is involved in oxygen transport in blood. That's the divergence. Although there are many interesting things to say about the processes by which duplication and divergence occur, we're going to focus on the pattern of nucleotide sequence evolution that arises as a result.

Globin evolution

I've just pointed out the distinction between myoglobin and hemoglobin. You may also remember that hemoglobin is a multimeric protein consisting of four subunits, 2 α subunits and 2 β subunits. What you may not know is that in humans there are actually two types of α hemoglobin and four types of β hemoglobin, each coded by a different genetic locus (see Table 1). The five α-globin loci (α_1, α_2, ζ, and two non-functional pseudogenes) are found in a cluster on chromosome 16. The six β-globin loci (ϵ, γ_G, γ_A, δ, β, and a pseudogene) are found in a cluster on chromosome 11. The myoglobin locus is on chromosome 22.

Not only do we have all of these different types of globin genes in our bodies, they're all related to one another. Comparative sequence analysis has shown that vertebrate myoglobin and hemoglobins diverged from one another about 450 million years ago. Figure 1 shows a phylogenetic analysis of part of the globin gene family, namely the β globin genes within tetrapods. If you stare at this tree for a while, you'll notice a couple of interesting things:

Developmental stage	α globin	β globin
Embryo	ζ	ϵ
	α	ϵ
Fetus	α	β
	α	γ
Adult	α	β
	α	δ

Table 1: Human hemoglobins arranged in developmental sequence. Adult hemoglobins composed of 2α and 2δ subunits typically account for less than 3% of hemoglobins in adults (http://sickle.bwh.harvard.edu/hbsynthesis.html).

- Eutherian β and δ globins are more closely related to marsupial β globins than they are to eutherian ϵ or γ globins.
- Marsupial β globin is more closely related to eutherian β and δ globins than it is to marsupial ϵ globin.

To put that another way, β globin genes within humans (a eutherian) are more closely related to β globin genes in kangaroos (a marsupial) than to ϵ globin genes in humans. Strange as it seems, this pattern is exactly what we expect as a result of duplication and divergence.

Up to the time that a gene becomes duplicated, its evolutionary history matches the evolutionary history of the organisms containing it. Once there are duplicate copies, each follows an independent evolutionary history. Each traces the history of speciation and divergence. And over long periods duplicate copies of the same gene share more recent common ancestry with copies of the same gene in a different species than they do with duplicate genes in the same genome. You can see that in this example if we redraw the gene tree in Figure reffig:globins as a species tree with the gene tree inside it (Figure 2).

A history of duplication and divergence in multigene families makes it important to distinguish between two classes of related loci: those that represent the same locus in different species and between which divergence is a result of species divergence are *orthologs*. Those that represent different loci and between which divergence occurred after duplication of an ancestral gene are *paralogs*. The β-globin loci of humans and chickens are orthologous. The α- and β-globin loci of any pair of taxa are paralogous.

As multigene families go, the globin family is relatively simple and easy to understand. There are only about a dozen loci involved, one isolated locus (myoglobin) and two clusters of loci (α- and β-globins). You'll find a diagram of the β-globin cluster in Figure 3. As you

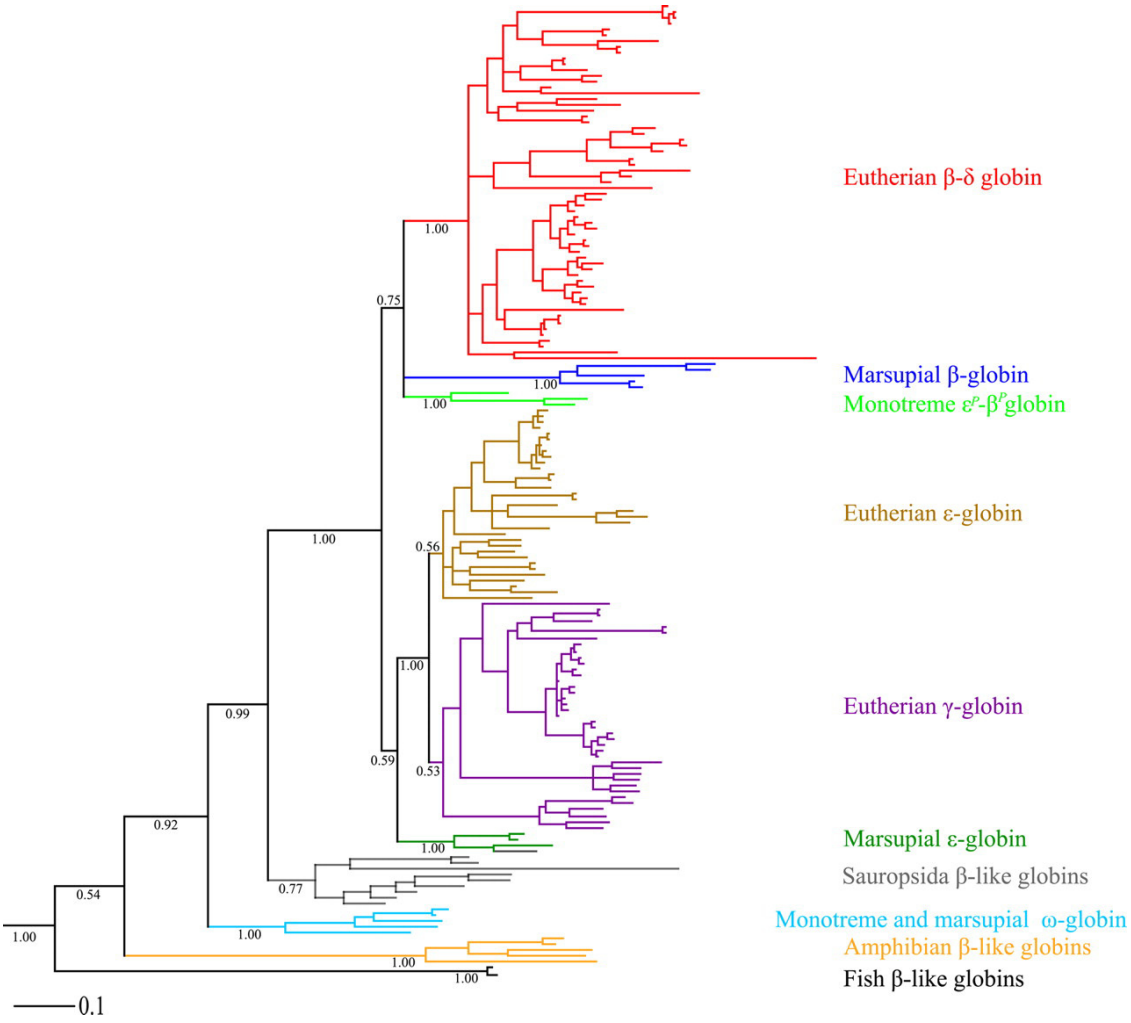

Figure 1: Evolution of β-globin genes in tetrapods drawn as a gene tree (from [7]).

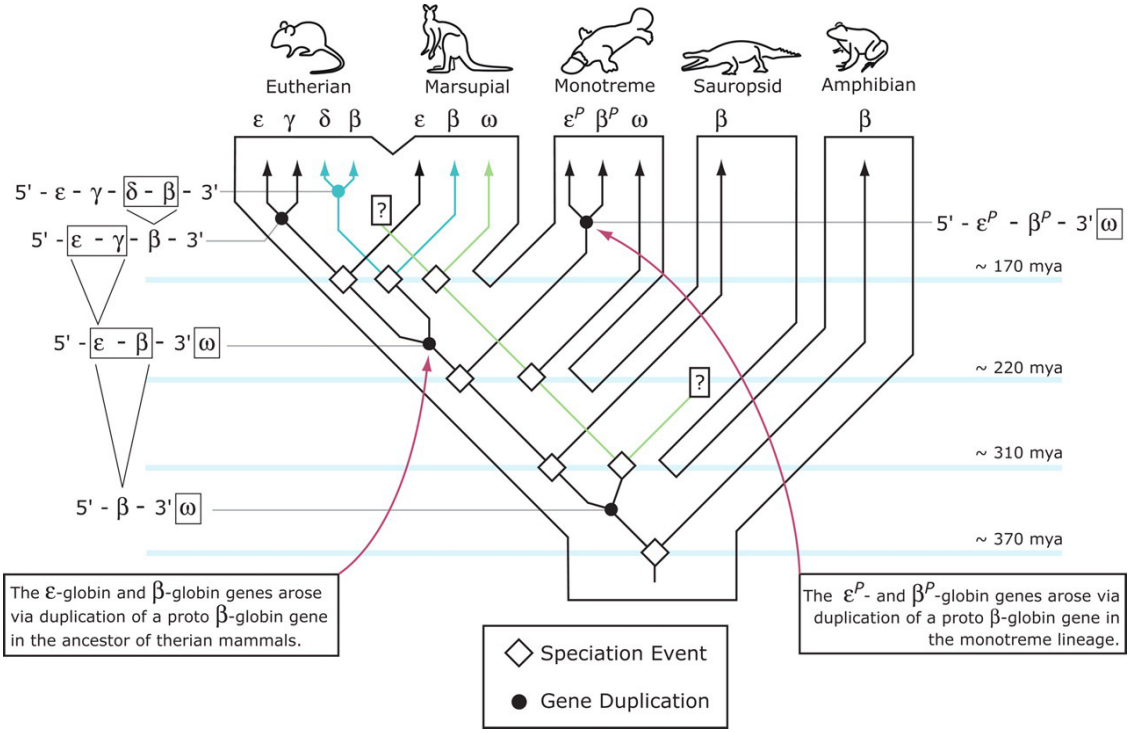

Figure 2: Evolution of β-globin genes in tetrapods drawn as a species tree (from [7]).

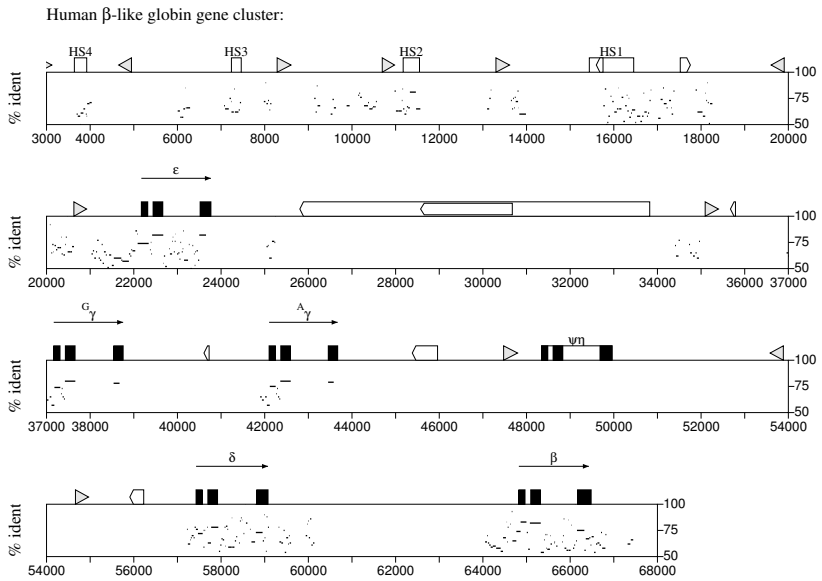

Figure 3: Structure of the human β-globin gene cluster. % identity refers to similarity to the mouse β-globin sequence. From http://globin.cse.psu.edu/html/pip/betaglobin/iplot.ps (retrieved 28 Nov 2006).

can see the β-globins are not only evolutionarily related to one another they occur relatively close to one another on chromosome 11 in humans.

Other families are far more complex. Class I and class II MHC loci, for example are part of the same multigene family. Moreover, immunoglobulins, T-cell receptors, and, and MHC loci are part of a larger superfamily of genes, i.e., all are ultimately derived from a common ancestral gene by duplication and divergence. Table 2 lists a few examples of multigene families and superfamilies in the human genome and the number of proteins produced.

Concerted evolution

Although the patterns of gene relationships produced through duplication and divergence can be quite complex, the processes are relatively easy to understand. In some multigene families, however, something quite different seems to be going on. In many plants and animals, genes encoding ribosomal RNAs are present in hundreds of copies and arranged end to end in long tandem arrays in one or a few places in the genome (Figure 4). Brown et al. [1] compared the ribosomal RNA of *Xenopus laevis* and *X. mulleri* and found a surprising pattern. There

Protein family domain	Number of proteins
Actin	61
Immunoglobulin	381
Fibronectin type I	5
Fibronectin type II	11
Fibronectin type III	106
Histone	
H2A/H2B/H3/H4	75
Homeobox	160
Immunoglobulin	381
MHC Class I	18
MHC Class IIα	5
MHC Class IIβ	7
T-cell receptor α	16
T-cell receptor β	15
T-cell receptor γ	1
T-cell receptor δ	1
Zinc finger, C2H2	564
Zinc finger, C3HC4	135

Table 2: A few gene families from the human genome (adapted from [6, 2]).

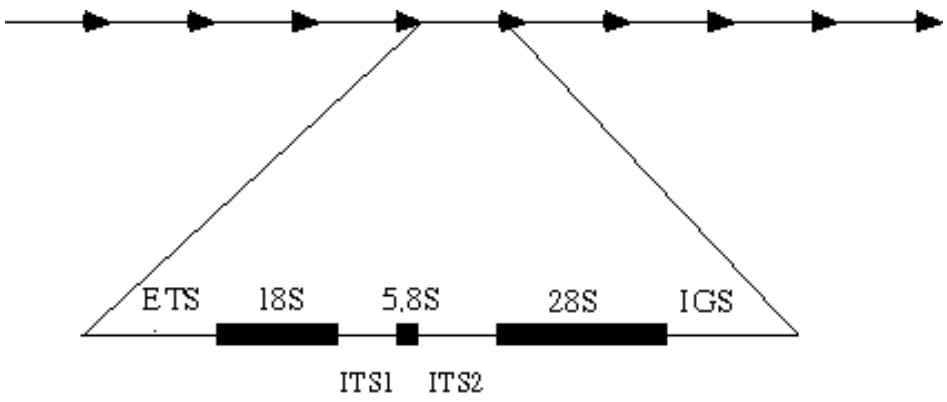

Figure 4: Diagrammatic representation of ribosomal DNA in vascular plant genomes (from Muir & Schlötterer, 1999 http://webdoc.sub.gwdg.de/ebook/y/1999/whichmarker/m11/Chap11.htm).

was little or no detectable variation among copies of the repeat units within either species, in spite of substantial divergence between them. This pattern can't be explained by purifying selection. Members of the gene family presumably diverged before *X. laevis* and *X. mulleri* diverged. Thus, we would expect more divergence among copies *within* species than *between* species, i.e., the pattern we see in the globin family. Explaining this pattern requires some mechanism that causes different copies of the repeat to be homogenized within each species while allowing the repeats to diverge between species. The phenomenon is referred to as concerted evolution.

Two mechanisms that can result in concerted evolution have been widely studied: unequal crossing over and gene conversion. Both depend on misalignments during meiotic prophase. These misalignments allow a mutation that occurs in one copy of a tandemly repeated gene array to "spread" to other copies of the gene array. Tomoko Ohta and Thomas Nagylaki have provided exhaustive mathematical treatments of the process [3, 5]. We'll follow Ohta's treatment, but keep it fairly simple and straightforward. First some notation:[1]

$$
\begin{aligned}
f &= \text{P(two alleles at same locus are ibd)} \\
c_1 &= \text{P(two alleles at different loci in same chromosome are ibd)} \\
c_2 &= \text{P(two alleles at different loci in different chromosomes are ibd)} \\
\mu &= \text{mutation rate} \\
n &= \text{no. of loci in family}
\end{aligned}
$$

[1]See Figure 5 for a diagram that you may find helpful

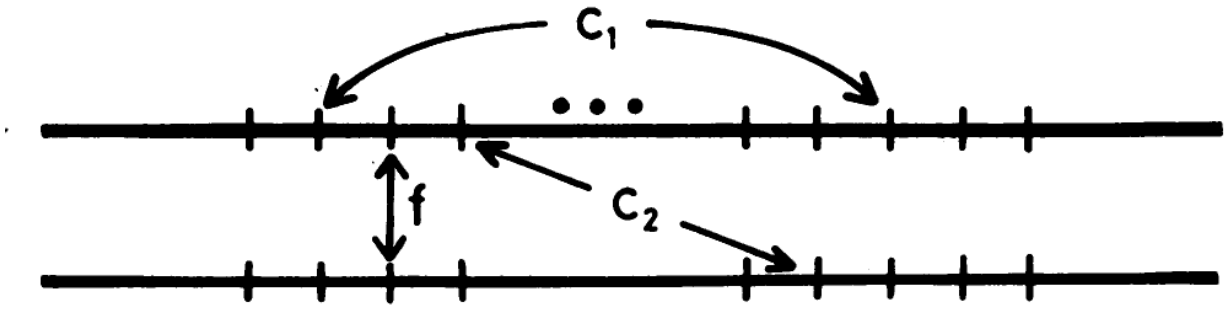

Figure 5: Types of identity by descent within a tandem repeat (from [4]).

λ = rate of gene conversion

Now remember that for the infinite alleles model

$$f = \frac{1}{4N_e\mu + 1} \quad ,$$

and f is the probability that neither allele has undergone mutation. By analogy

$$g = \frac{1}{4N_e\lambda + 1} \quad ,$$

where g is the probability that two alleles at a homologous position are ibd in the sense that neither has ever moved from that position in the array. Thus, for our model

$$\begin{aligned} f &= P(\text{neither has moved})\text{P(ibd)} \\ &\quad +\text{P(one has moved)P(ibd anyway)} \\ &= \left(\frac{1}{4N_e\lambda + 1}\right)\left(\frac{1}{4N_e\mu + 1}\right) + \left(\frac{4N_e\lambda}{4N_e\lambda + 1}\right) c_2 \\ &\approx \frac{4N_e\lambda c_2 + 1}{4N_e\lambda + 4N_e\mu + 1} \\ c_1 = c_2 &= \frac{\lambda}{\lambda + (n-1)\mu} \quad . \end{aligned}$$

Notice that $(n-1)\mu$ is approximately the number of mutations that occur in a single array every generation. Consider two possibilities:

- *Gene conversion occurs much more often than mutation*: $\lambda \gg (n-1)\mu$.

  Under these conditions $c_2 \approx 1$ and $f \approx 1$. In short, all copies of alleles at every locus in the array are virtually identical — concerted evolution.

- *Gene conversion occurs much less often than mutation*: $\lambda \ll (n-1)\mu$.

  Under these conditions $c_2 \approx 0$ and $f \approx \frac{1}{4N_e\mu+1}$. In short, copies of alleles at different loci are almost certain to be different from one another, and the diversity at any single locus matches neutral expectations — non-concerted evolution.

References

[1] D D Brown, P C Wensink, and E Jordan. Comparison of the ribosomal DNA's of Xenopus laevis and Xenopus mulleri: the evolution of tandem genes. *J. Mol. Biol.*, 63:57–73, 1972.

[2] J C et al. Venter. The sequence of the human genome. *Science*, 291:1304–1351, 2001.

[3] T Nagylaki. Evolution of multigene families under interchromosomal gene conversion. *Proceedings of the National Academy of Sciences USA*, 81:3796–3800, 1984.

[4] T. Ohta. Allelic and nonallelic homology of a supergene family. *Proceedings of the National Academy of Sciences, USA*, 79:3251–3254, 1982.

[5] T Ohta. Some models of gene conversion for treating the evolution of multigene families. *Genetics*, 106:517–528, 1984.

[6] T Ohta. Gene families: multigene families and superfamilies. In *Encyclopedia of the Human Genome*. Macmillan Publishers Ltd., London, 2003.

[7] J. C. Opazo, F. G. Hoffman, and J. F. Storz. Genomic evidence for independent origins of β-like globin genes in monotremes and therian mammals. *Proceedings of the National Academy of Sciences, USA*, 105:1590–1595, 2008.

Creative Commons License