Comment

An apology for orthologs - or brave new memes

Eugene V Koonin

Address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.
E-mail: koonin@ncbi.nlm.nih.gov

Published: 6 April 2001

Genome ***Biology*** 2001, **2(4)**:comment1005.1–1005.2

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/4/comment/1005

Over the last few months, I have learned to look forward to Gregory Petsko's comments in *Genome Biology*. Beyond enjoying the witticisms, I tend mostly to agree with his message. When I started reading the recent article 'Homologuephobia' (*Genome Biology* 2001, **2:**comment1002), it was no different. Like Petsko, I hate that "ue" at the end of "homologue", and for that matter, all the other obnoxious "e"s'. I never consider having a drink in any establishment that has 'Olde' in its name, or buying as much as a hair comb in a 'Shoppe'. And coming back to the 'homologous' problem, I think I have made a small contribution to leaner, meaner spelling by getting away with 'homolog' in many publications, albeit accepting the forced 'homologue' in others (even as I type this, the impervious little red wave from my spell-checker is, of course, right here, under 'homolog'). So I was very much with Petsko on this momentous issue. As I read on, however, my happiness started to waver, and after reaching the invective against orthologs and paralogs, which Petsko says "add nothing to the subject", I felt that I had to lift my self-imposed ban on writing comments and respond to his article.

Let me put it bluntly: I am confident that orthologs and paralogs not only 'add something to the subject' but are critical for the development of evolutionary genomics (and as soon as two genomes were sequenced, all genomics became evolutionary). These are not fancy words (nor new, by the way: the notion of orthology versus paralogy was introduced by Walter Fitch in a seminal 1970 paper; *Syst Zool* 1970, **19:**99-113), but are essential designations for two distinct types of evolutionary relationships. In a nutshell, orthologs are direct evolutionary counterparts derived from a common ancestor through vertical descent; whenever we speak of 'the same gene in different species', we actually mean orthologs. In contrast, paralogs are genes within the same genome that have evolved by duplication. Distinguishing between ortho and para is critical if we strive to describe evolution with any semblance of accuracy. It is equally important for inferring gene function. Conservation of function is not part of the definition of orthology but rather its consequence.

The distinction is not only logical but also very practical, because it is quite common for the same function in different organisms to be performed by proteins that are not orthologs nor even homologs (defining homologs as any genes that have common ancestry, including both orthologs and paralogs). Hence another neologism of comparative genomics that might induce a cringe even in some stronger souls who put up with orthologs and paralogs: non-orthologous gene displacement, when unrelated - or at least not orthologous - genes perform analogous functions (see Koonin EV, Mushegian AR, Bork P: *Trends Genet* 1996, **12:**334-336). I do maintain, however, that this one also helps us to speak more, rather than less, accurately and comprehensibly about what is really going on during genome evolution.

There is, however, yet another wrinkle that becomes apparent when one tries to think this through. Look at the trivialized schematic in Figure 1. Clearly, genes A1 and A2 are orthologs, and so are B1 and B2; and without hesitation we will call A1 and B1 (or A2 and B2) paralogs, just as A and B were paralogs in the ancestral species. But what about A1 and B2? These are not orthologs - they are not directly connected by vertical descent, not 'the same gene in different species' - but neither are they paralogs, at least not according to the formal definition, because they reside in different genomes. Are we in need of yet another term? Perhaps metalogs? This is not an idle concern. Imagine that B1 and A2 have been lost during evolution and A1 and B2 are all that remain of this gene family. We need to be able adequately to describe the relationships between them, and at present the best way to do so seems to be through the vague statement that they are 'homologs but not orthologs'. Personally, I would prefer a new term.

So what's the issue with all these new terms (or "exapted" ones, to use a favorite term of Stephen Jay Gould's to indicate something pre-existing that has been recruited for a new function)? Or, for a good measure, with all the mushrooming '-omes' - transcriptome, proteome, metabolome,

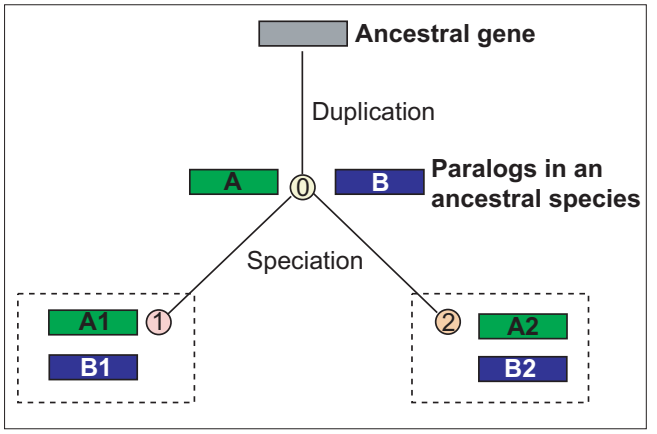

Figure 1
Relationships between orthologs and paralogs, illustrated by the evolution of an ancestral gene, via duplication in species 0, to two genes in each of species 1 and 2. The relationships among all these genes are discussed in the text.

and even phylome, a really fresh one that has been introduced to designate the complete set of phylogenetic trees for the genes from a given genome (Sicheritz-Ponten T, Andersson SG: *Nucleic Acids Res* 2001**, 29:**545-552)? Is the world a better place because of them? Actually, for what it's worth, I think it is. These are not just words, after all: they are new memes for the science of a new age. The meme, of course, is itself a quintessential neologism, a brilliant (in my opinion) invention Richard Dawkins first introduced in his 1976 book *The Selfish Gene*. A meme is a resilient unit of cultural inheritance, a definition that includes the more conventional notions of 'concept', 'idea' or 'theory' but is much broader. The 'ortholog meme', for example, encapsulates a whole panoply of diverse concepts beyond the strict definition: the existence of discrete gene histories that can be traced back to the last universal common ancestor, at least in principle; a broad distribution of evolutionary rates, as a result of which some orthologs are highly conserved whereas the similarity between others is barely detectable; as a synthesis of the previous two notions, the primacy of the actual historical relationship, once revealed, over the quantitative criterion of similarity, in establishing orthology and predicting function; the possibility of one-to-many and many-to-many orthologous relationships as a result of terminal duplications; the genuine difficulties encountered by the concept of orthology because of the fluidity of protein-domain architectures, particularly in multicellular eukaryotes... and more. Expounding it all would require a dissertation, some parts of which would be vague or controversial - but the simple term ortholog conveys all these notions, and that is why it seems to be a (relatively) successful meme.

Now, by talking about a 'successful' meme, we have hit on something important. Memes, like genes, evolve through an interplay of mutation and selection (as described, in fascinating detail, by Daniel Dennett in his 1995 book *Darwin's Dangerous Idea: Evolution and the Meanings of Life*). This is why it doesn't make sense to worry too much about the proliferation of terms: those that correspond to good, fit memes will survive and prosper, while the rest will die out or will lead a marginal existence. This is how new scientific paradigms are born - as collections of new memes, at first haphazard, and then, after selection does its job, coherent. Those old enough to remember the early, heroic days of molecular biology (see also Horace Judson's fascinating 1996 book *The Eighth Day of Creation: Makers of the Revolution in Biology*) will recall the wild bloom of various '-ons'. Some, like codon and operon, designate truly important memes and are now all over the place; others have respectable but modest lives, like replicon or regulon; yet others have been more or less marginalized, like cistron (which remains respectable in poly- and mono-cistronic forms); and some became extinct, for example, recon (used to mean a unit of recombination). Similar fates will, of course, befall the '-omes' that propagate in these early days of genome science. Personally, I have little doubt that proteome and transcriptome are here to stay; I am a little less confident about metabolome and phylome, although I quite like the latter. My main point, however, is a tribute to meme selection: the fittest will survive!

These are just some of the thoughts triggered by Petsko's comment about orthologs and paralogs. I thank him for this and now return to my daily occupation - the identification of orthologs by analysis of proteomes.