Correspondence

Orthologs and paralogs - we need to get it right

Roy A Jensen

A response to **Homologuephobia** by Gregory A Petsko, *Genome Biology* 2001, **2**:comment1002.1-1002.2, *to* **An apology for orthologs - or brave new memes** by Eugene V Koonin, *Genome Biology* 2001, **2**:comment1005.1-1005.2, *and to* **Can sequence determine function?** by John A Gerlt and Patricia C Babbitt, *Genome Biology* 2000, **1**:reviews0005.1-0005.10

Address: Department of Microbiology and Cell Science, Gainesville, FL 32611, USA. Department of Chemistry, City College of New York, New York, NY 10031, USA. BioScience Division, Los Alamos National Laboratory, Los Alamos, NM 87544, USA. E-mail: rjensen@ufl.edu

Published: 3 August 2001

Genome ***Biology*** 2001, **2(8)**:interactions1002.1–1002.3

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/8/interactions/1002

Eugene Koonin is absolutely right in his *Genome Biology* article 'An apology for orthologs - or brave new memes' [1] in defending the importance of the terms 'ortholog' and 'paralog' for making significant evolutionary inferences about the relationships between genes. Nevertheless, Gregory Petsko's suggestion in his comment 'Homologuephobia' [2] that the use of ortholog and paralog "adds nothing to the subject" is painfully understandable because of the current rampant misuse of these terms. I believe that Koonin's comment may even add to the confusion. The current widespread confusion about the meaning of these terms has not gone unnoticed, and Walter Fitch, who first used these essential terms [3], was recently asked to address the issue [4]. I cannot hope to improve upon his essay, but maybe this letter can help to push toward a much-needed awareness of what should not really be that complicated.

Fitch [4] showed, in a most illuminating and powerful diagram, four events of evolutionary divergence, two being events of speciation and two being events of gene duplication, yielding six contemporary genes in the three organisms, A, B and C (Figure 1a). Determination of orthology or paralogy in a vertical line of descent is a simple matter of tracking any pair of genes back to where they join, either at an inverted 'Y' (in which case they are orthologs) or at a horizontal line (in which case they are paralogs). Thus, Al has three orthologs in species C, but only C1 is an ortholog of B1. On the other hand, B2 has two orthologs in species C (C2 and C3), whereas B2 and C1 are paralogs. The three genes in species C are paralogous to each other. Notably, every relationship between genes is one of paralogy or orthology, but a given gene in one species may have more than one ortholog in another species (none being any more 'correct' than another), and paralogs are not necessarily restricted to the same species.

In his comment, Koonin [1], posing the simpler hypothetical situation shown in Figure 1b, stated that A1 and B2 are not formally paralogs because they reside in different genomes (see Figure 1b). But, as asserted above, paralogs will often reside in different genomes, and I have illustrated the relationship of orthology and paralogy for the scenario presented in Figure 1b by redrawing it (Figure lc) with the type of diagram suggested by Fitch and exemplified in Figure 1a. The impression that paralogs should always be in the same genome may have arisen because, at the time during evolution when paralogs originate by gene duplication, they will indeed be in the same genome. Multiple homologs in the same genome will always be paralogous, but this does not mean that paralogs will always be restricted to the same genome as evolution progresses. An examination of the evolution of the paralog relationships shown in Figure 1a should help clarify this issue.

In the evolutionary scenario shown in Figure 1b, Koonin considered the situation in which genes B1 and A2 have been lost during evolution and A1 and B2 are all that remain of this gene family; he asked how we can then "adequately describe the relationships between them". They are simply paralogs. The loss of B1 and A2 does not change the paralogous relationship of A1 and B2. The gene relationships given in Figure 1b,c exemplify the fact that a valid gene tree is not necessarily the same as the species tree. On the one hand, the tree relationship between A1 and A2 or B1 and B2 will be the same as the species tree. On the other hand, the tree relationship between A1 and B2 or A2 and B1 will not be the same as the species tree because divergence via gene duplication preceded speciation. The question was raised by Koonin [1] as to whether a new term such as

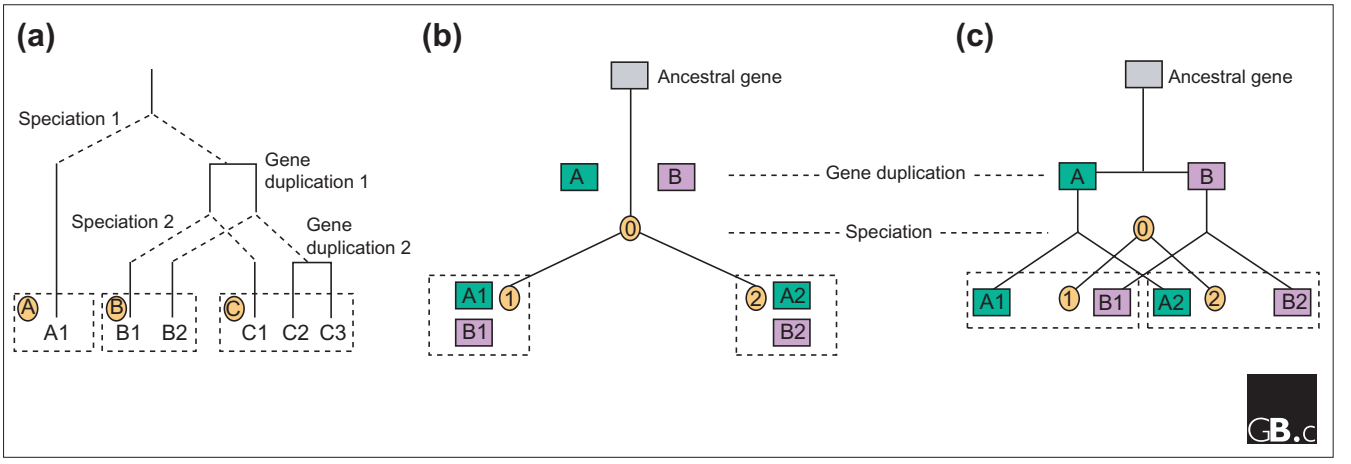

Figure 1
(a) Simplified diagram of homology subtypes (showing orthologs and paralogs, but not xenologs); adapted from [4]. Speciation events produce the species A, B and C. The genes A1, B1, B2, C1, C2, and C3 have descended from the ancestral gene following evolutionary events of speciation and gene duplication. **(b,c)** Evolutionary descent of an ancestral gene to paralogs and orthologs following gene duplication in species 0, and then speciation to yield species 1 and 2. Diagram (b) shows the resulting relationship between paralogs and orthologs as illustrated by Koonin in his comment [1]. Diagram (c) is my version of Koonin's diagram using a Fitch diagram for visualization. Note that the two evolutionary events depicted are a subset of the four shown in (a) (gene duplication 1 and speciation 2), and that the use of capital letters for genes and numbers for species is the opposite of that used in (a).

'metalog' might be coined to describe evolutionary situations in which genes corresponding to a certain function in different species are paralogs (for example, A1 and B2), rather than orthologs (for example, A1 and A2). This would seem ill advised because we are dealing with a particular relationship between paralogs, yet the term implies equal status of 'metalogy' with the subtypes of homology - orthology, paralogy, and xenology (the relationship of any two homologs whose history, since their common ancestor, involves horizontal transfer of at least one of them). If any new terminology is coined, it perhaps could define different classes of paralogs.

Yet another misuse of the terms 'ortholog' and 'paralog' is quite common in the literature as seen, for example, in a review in *Genome Biology* by Gerlt and Babbitt [5]. Here, orthologs are defined as homologs in different species that catalyze the same reaction, and paralogs are defined as homologs in the same species that do not catalyze the same reaction. Although plenty of examples exist for which this evolutionary scenario has indeed played out, it is quite possible for orthologs to acquire different catalytic (or regulatory) properties and for paralogs to retain the same function. Orthology and paralogy differ in that one proceeds from speciation and the other from gene duplication, but either evolutionary course of divergence has the same potential for acquisition of new properties. Biochemists may find it useful to classify isofunctional homologs and heterofunctional homologs and to find acceptable words to distinguish between these, but to distort the meaning of the classic terms ortholog and paralog risks causing chaos in the evolutionary context.

References

1. Koonin EV: **An apology for orthologs - or brave new memes.** *Genome Biol* 2001, **2**:comment1005.1-1005.2.
2. Petsko GA: **Homologuephobia.** *Genome Biol* 2001, **2**:comment1002.1-1002.2.
3. Fitch WM: **Distinguishing homologous from analogous proteins.** *Syst Zool* 1970, **19**:99-113.
4. Fitch WM: **Homology a personal view on some of the problems.** *Trends Genet* 2000, **16**:227-31.
5. Gerlt JA, Babbitt PC: **Can sequence determine function?** *Genome Biol* 2000, **1**:reviews0005.1-0005.10

John Gerlt and Patricia Babbitt respond:

We agree with Jensen that communication between genomic and evolutionary biologists can be frustrated by the imprecise use of terms that were coined in the simpler, more abstract period of 'pre-genomic' biology. In part, this problem is associated with the realization that the functional distinctions associated with divergence of sequence are far more complex than could have been imagined when the terms 'ortholog' and 'paralog' were originally proposed by Fitch [3,4].

We have used 'ortholog' and 'paralog' to describe relationships between gene products, at least in part, because we prefer to adapt the definitions of existing words to a new intellectual environment rather than to invent new words. But, as we are reminded by Jensen, the terms originated within the evolutionary biology community and strictly refer to sequence divergence associated with either speciation or gene duplication, respectively, and do not have either implicit or explicit functional implications.

Setting function aside, correct usage of 'ortholog' and 'paralog' requires knowledge of the details of the evolutionary pathways that produced the divergence of biological functions that we and others are attempting to describe in the context of both sequence and three-dimensional structure. Jensen states that "determination of orthology or paralogy is a simple matter of tracking any pair of genes back to where they join" (speciation or gene duplication). But we believe that insufficient information is available to accurately determine the timing of many of the speciation and gene duplication events that gave rise to the contemporary slate of genomes. In particular, analysis of the interesting structure-function relationships among highly divergent proteins must usually proceed without benefit of this information. So, whether two contemporary proteins are orthologs or paralogs cannot be determined with certainty.

Genomic biology needs to get beyond semantic issues. It needs to focus on defining those sequence-structure-function relationships that are necessary for understanding both the structural origins of biological function and the molecular bases for the divergence of biological function. So, those of us who study the relationships among sequence, structure, and function should discontinue the use of 'ortholog' and 'paralog', unless we want to focus on the speciation and gene duplication events that produced functional diversity in homologs.

But, unlike Petsko [2], we believe that genomic biologists need to describe, compare, and contrast sequence-structure-function relationships not only for a complete group of homologs but also for subsets of homologs that share particular attributes. Based on our experiences, genomic biologists need words to describe 'homologs encoded by different genomes' and 'homologs that have different functions'.

To accomplish these needs, we suggest the following adjectives to describe homologs: 'Isofunctional' homologs exhibit the same function(s); 'heterofunctional' homologs exhibit different functions; 'isospecic' homologs are found in the same species; and 'heterospecic' homologs are in different species.

Let us take an example from our review in *Genome Biology* [5]. The *Escherichia coli* genome encodes eight homologs of enoyl-CoA hydratase; the *Bacillus subtilis* genome encodes seven homologs. The 1,4-dihydroxynaphthoyl-CoA synthases in *E. coli* and *B. subtilis* are heterospecic, isofunctional homologs; and the 1,4-dihydroxynaphthoyl-CoA synthase and methylmalonyl-CoA decarboxylase in *E. coli* are isospecic, heterofunctional homologs; whereas the methylmalonyl-CoA decarboxylase in *E. coli* and the 1,4-dihydroxynaphthoyl-CoA synthase in *B. subtilis* are heterospecic, heterofunctional homologs. Neither genome encodes isospecic, isofunctional homologs of enoyl-CoA hydratase. Although the enoyl-CoA hydratase domains of FadB and YcfX in *E. coli* both catalyze the enoyl-CoA hydratase reaction in fatty-acid oxidation, the reaction catalyzed by the former occurs under aerobic conditions whereas the reaction catalyzed by the latter occurs under anaerobic conditions.

Hopefully, with these words for clarifying the specific and functional relationships of homologs, genomic biologists can focus on deciphering the information contained in genomes and communicating that information to all segments of the biology community.

John A Gerlt[1] and Patricia C Babbitt[2],
[1]Departments of Biochemistry and Chemistry, University of Illinois, Urbana, IL 61801, USA.
[2]Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, University of California, San Francisco, CA 94143, USA.

Correspondence: John A Gerlt.
E-mail: j-gerlt@uiuc.edu