Xenolog Classification

Citation for published version:

Digital Object Identifier (DOI):
10.1093/bioinformatics/btw686

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Bioinformatics

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Subject Section

Xenolog Classification

Charlotte Darby\textsuperscript{1,3,†}, Maureen Stolzer\textsuperscript{1,†}, Patrick Ropp\textsuperscript{1}, Daniel Barker\textsuperscript{4,5} and Dannie Durand\textsuperscript{1,2,∗}

\textsuperscript{1}Department of Biological Sciences, \textsuperscript{2}Department of Computer Science, Carnegie Mellon University, Pittsburgh, 15213, USA
\textsuperscript{3}Present address: Department of Computer Science, Johns Hopkins University, Baltimore, 21218, USA.
\textsuperscript{4}School of Biology, University of St. Andrews, St. Andrews, Fife KY16 9TH, UK.
\textsuperscript{5}Present address: Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, UK.

\textsuperscript{†}To whom correspondence should be addressed. \textsuperscript{†}These authors contributed equally to this work.

Abstract

Motivation: Orthology analysis is a fundamental tool in comparative genomics. Sophisticated methods have been developed to distinguish between orthologs and paralogs and to classify paralogs into subtypes depending on the duplication mechanism and timing, relative to speciation. However, no comparable framework exists for xenologs: gene pairs whose history, since their divergence, includes a horizontal transfer. Further, the diversity of gene pairs that meet this broad definition calls for classification of xenologs with similar properties into subtypes.

Results: We present a xenolog classification that uses phylogenetic reconciliation to assign each pair of genes to a class based on the event responsible for their divergence and the historical association between genes and species. Our classes distinguish between genes related through transfer alone and genes related through duplication and transfer. Further, they separate closely-related genes in distantly-related species from distantly-related genes in closely-related species. We present formal rules that assign gene pairs to specific xenolog classes, given a reconciled gene tree with an arbitrary number of duplications and transfers. The xenology classification rules have been implemented in software and tested on a collection of \textasciitilde 13,000 prokaryotic gene families. In addition, we present a case study demonstrating the connection between xenolog classification and gene function prediction.

Availability: The xenolog classification rules have been implemented in Notung 2.8, a freely available phylogenetic reconciliation software package. \url{http://www.cs.cmu.edu/~durand/Notung}. Gene trees are available at \url{http://datashare.is.ed.ac.uk/handle/10283/1981}.

Contact: durand@cmu.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Homology analysis, classifying gene pairs according to the evolutionary process by which they diverged, is a fundamental tool of comparative genomics. Identifying orthologs is integral to the functional annotation of novel genes (Wu et al., 2003) and prediction of gene function by various methods, including phylogenetic profiling (Pellegrini et al., 1999) and gene fusion (Marquette et al., 1999; Enright et al., 1999). Phylostratigraphic investigations linking the age of a gene to its functions, disease associations, or ecological distribution exploit the fact that orthologs from the same species diverged at roughly the same time (Capra et al., 2013). Orthologs are used as markers for homologous chromosomal regions for comparative mapping (Nadeau and Sankoff, 1998; O’Brien et al., 1997), phylogenetic footprinting (Duret and Bucher, 1997; Dickmeis and Muller, 2005) and operon prediction (Chen et al., 2004; Ermolaeva et al., 2001; Price et al., 2005; Westover et al., 2005).

Identification of paralogs is a prerequisite for studying processes of gene duplication, the major source of genetic novelty in eukaryotes. Comparison of paralogous pairs with a pre-duplication ortholog reveals patterns and rates of diversification following duplication (Lynch, 2007, and work cited therein), as well as the functional fates of duplicated genes (Lynch, 2007). Spatial patterns of orthologs and paralogs are used
Homology identification is a highly active research area, comprising methodological approaches ranging from sequence comparison to phylogenetic reconciliation. More recent innovations include the exploitation of shared synteny (Shi et al., 2011) and specialized methods for identifying multidomain homologs (Song et al., 2007, 2008; Ali et al., 2016).

Most work on homology analysis to date has not considered genes related through horizontal transfer. Studies of horizontal transfer commonly use approaches that seek to identify genes of foreign origin in a given genome, rather than homologous gene pairs that are related through horizontal transfer (reviewed in Azad and Lawrence, 2012). A few methods, such as gene tree - species tree reconciliation, do infer gene pairs that correspond to the donor and recipient of a transfer. Reconciliation algorithms that account for transfer events are relatively new (reviewed by Nahliel, 2010, 2013; Huson and Scornavacca, 2011), computationally more complex, and are only recently coming into use for genomic analyses (e.g., David and Alm, 2011; Richards et al., 2014).

Appropriate terminology for describing gene pairs related through horizontal transfer is a fundamental requirement for extending the homology analysis framework to include this evolutionary process. The term “xenolog,” proposed by Gray and Fitch (1983) to describe horizontally transferred genes, is in use, but not widely, and there is no consensus on a precise definition. Further, there has been little discussion of differentiating xenologs to convey distinctions between horizontally transferred genes with different properties (see Koonin et al., 2001, for a notable exception). Such xenolog classes would be analogous to paralog subtypes proposed to convey the relative timing of duplications and speciations (e.g., in-paralogs versus out-paralogs, Sonhammer and Koonin, 2002) or distinguish between different mechanisms of duplication (e.g., ohnologs and tandem duplications, reviewed by Durand and Hoberman, 2006; Ramos and Ferrier, 2012).

Background: Fitch (1970) introduced the terms orthology (“homology [that] is the result of speciation”) and paralogy (“homology [that] is the result of gene duplication”) and proposed that “foreign genes...since they are neither orthologous nor paralogous but are clearly homologous...should be called xenologous” (Gray and Fitch, 1983). These definitions, which are framed in terms of the event that caused the divergence, have been widely adopted. In 2000, Fitch proposed more precise definitions of orthology and xenology. Orthology includes the requirement that the “common ancestor lies in the cenancestor of the taxa from which the two sequences were obtained,” whereas a xenolog is the “most recent common ancestor of the [species] taxa under consideration,” and xenology is the “relationship of any two homologous characters whose history, since their common ancestor, involves an interspecies (horizontal) transfer of the genetic material for at least one of those characters.” In other words, a pair of genes, gX and gY, are xenologs, if there is a transfer on the path connecting g1 and g2 in the gene tree. In this updated definition, orthology is defined not just in terms of a speciation event, but in terms of the association of nodes in the gene and species trees. Under a duplication-loss event model, the earlier, event-based definition of orthology and this definition are equivalent. However, when transfers are included in the event model, the sets of orthologs predicted using the two definitions are not identical. Moreover, the event-based definition leads to predicted orthologs that have properties that are not usually associated with orthologs.

For example, nodes gX and gY in Figure 1 are orthologs according to the event-based definition, because the event at their most recent common ancestor (gPa) is a speciation. Yet gX and gY are genes in the same present-day species X, violating the assumption that genes in the same species cannot be orthologous. Gene pairs in species X and Z also exhibit surprising behavior according to the event-based definition. The most recent common ancestor of gX and gZ is a speciation node, as is the most recent common ancestor of gY and gZ, implying that both pairs are orthologs in the species X and Z. However, these pairs arose at very different times in the species tree, violating the assumption that orthologs drawn from the same pair of species are associated with the same species divergence and are roughly the same age (Goodman et al., 1979; Capra et al., 2013). Neither of these problems arises when the cenancestor-based definition is used, because neither gX nor gY are orthologs of gZ under that definition. In both cases, the most recent common ancestor of the genes does not lie in their cenancestor.

More generally, the additional cenancestor requirement results in a restricted set of orthologs that excludes these problematic cases. However, a consequence of defining orthologs narrowly is that xenologs are defined broadly: the set of gene pairs whose history, since their divergence, includes a transfer is substantially larger than the set of genes that diverged through a transfer event at their most recent common ancestor. Xenologs, when broadly defined, exhibit diverse properties. First, not all xenologs have the same event at their most recent common ancestor in the gene tree. We observe xenologs where this divergence arose via transfer (e.g., g1 and g2), speciation (e.g., g1 and g3), and duplication (e.g., g1 and g2). Second, xenologs can occur in the same species (e.g., g1 and g2). Third, xenologs may vary greatly in how closely they are related, and the divergence of a pair of xenologs may pre- or post-date the divergence of their associated species. For example, genes g1 and g2 diverged more recently than species X and species Z, whereas genes g1 and g3 diverged before species X and species W.

Our Contributions: This broad definition of xenologs does not convey important distinctions between the diverse and complex xenologous relationships that arise due to horizontal gene transfer. To address this, we propose xenolog classes that reflect the events associated with the divergence of a xenologous gene pair, and the relative timing of transfer and speciation events. We present formal definitions of these classes in the context of a reconciled gene tree and rules to assign xenologous gene pairs to classes. Further, we show that these classes form a hierarchy, connecting the relationship of xenologs to their placement in the gene and species trees.

An algorithm implementing these rules has been integrated into the Notung 2.8 software package. An analysis of ~13,000 prokaryotic gene families demonstrates that all of the proposed classes arise in real gene tree data. We further present a case study that illustrates the potential functional implications of xenolog classification. Finally, we discuss how this framework could be used in future research to explore the evolutionary and functional fates of transferred genes.
2 Methods

Our classification takes as input a gene tree, \( T_G = (V_G, E_G) \), that has been reconciled with a species tree, \( T_S = (V_S, E_S) \), using a duplication-transfer model. The model may also include losses; losses have no impact on xenolog classification and we do not discuss them further. Reconciliation infers a mapping, \( \sigma (\cdot) \), between genes and species, where \( \sigma (g) = s \) indicates that gene \( g \) in \( V_G \) was present in the genome of species \( s \) in \( V_S \). Each internal node, \( r \), is annotated with \( \sigma (r) \), the event that caused the divergence of \( g \), where \( g \) may be a duplication (\( D \)), a transfer (\( T \)) or a speciation (\( S \)). Transfer edges are denoted by \( r \rightarrow \sigma (r) \), where \( g_r \) is the recipient gene given \( \sigma (r) \), and \( \sigma (g_r) \) gives \( r \). We say transfer \( r \) is on the path from \( g_i \) to \( g_j \) if the path from \( g_i \) to \( g_j \) passes through both \( g_r \) and \( g_s \).

The output of our classification scheme is a homology table \( H = \{ g_{ij} \} \), where \( g_{ij} \in \{ V_S \} \). In this classification, which is based on the definitions introduced by Fitch (2000), genes \( g_i \) and \( g_j \) are

- **orthologs** if \( g_i \in \text{MRCA}(\{g_{ij}\}) \) or \( g_j \in \text{MRCA}(\{g_{ij}\}) \); and there is no transfer on the path from \( g_i \) to \( g_j \).
- **paralogs** if \( g_i \in \text{MRCA}(\{g_{ij}\}) \) and there is no transfer on the path from \( g_i \) to \( g_j \).
- **xenologs** if there is at least one transfer on the path from \( g_i \) to \( g_j \).

Note that by explicitly defining orthologs to be genes that are not connected by a transfer, this definition of ortholog ensures that the ancestor of orthologous genes lie in the same sub-tree; i.e., \( \text{MRCA}(\{g_{ij}\}) = \text{MRCA}(\{g_i\}, \{g_j\}) \).

If \( g_i \) and \( g_j \) are orthologs, then \( H(g_i, g_j) = H(g_j, g_i) = 0 \). If they are paralogs, \( H(g_i, g_j) = H(g_j, g_i) = P \). If \( g_i \) and \( g_j \) are xenologs, then \( H(g_i, g_j) = \Delta(g_i, g_j) \), where \( \Delta(g_i, g_j) \) is the xenolog class of genes \( g_i \) and \( g_j \). In contrast to orthology and paralogy, xenology is not symmetric, due to the directional nature of horizontal transfer.

In the remainder of this section, we define new xenolog classes and give formal rules for determining the xenolog class, \( \Delta(g_i, g_j) \), for a given gene pair \( g_i \) and \( g_j \).

### 2.1 Xenolog classification with a single transfer

Consider a gene tree with a single transfer \( T = (V, E) \) from donor species \( Y_D = \{ g_D \} \) to recipient species \( Y_R = \{ g_R \} \). Let \( g_{ij} = \text{MRCA}(g_D, g_R) \) be the common ancestor of \( g_D \) and \( g_R \) and \( n \) be the set of nodes in the subtree of \( T \) rooted at \( g_{ij} \). Transfer \( T \) defines three, non-overlapping sets of species tree nodes:

\[
D = \{ s \in V | \text{MRCA}(s, g_D) < g_{ij} \}, \text{i.e. the species that are more closely related to the donor than the recipient;}
\]

\[
R = \{ s \in V | \text{MRCA}(s, g_R) < g_{ij} \}, \text{i.e. the species that are more closely related to the recipient than the donor;}
\]

\[
O = V \setminus (D \cup R), \text{i.e. the nodes in the species tree equally related to the donor and recipient.}
\]

We define four, mutually exclusive xenolog classes based on these sets. Xenolog classes are defined with respect to a reference gene \( g \in L(T_G) \) that is a descendant of the recipient of the transfer, i.e. \( g \in D(g_R) \). For every \( g \in \{ L(V_S) \} \), \( g \in O \) is on the path from \( g \) to \( g_R \) and \( g \) is a

- **Primary xenolog** if \( g \in \Delta(g_D) \); \( H(g_D, g) = PX \)
- **Sibling Donor xenolog** if \( m(g) \in D \) and \( g \notin \Delta(g_D) \); \( H(g_D, g) = SDX \)
- **Sibling Recipient xenolog** if \( \sigma (g) \in R \); \( H(g_R, g) = SRX \)
- **Outgroup xenolog** if \( g \in O \); \( H(g_R, g) = OX \)

Xenologs are classified relative to a reference gene; therefore, xenolog class assignments are not symmetrical. In the homology table, when \( H(g_D, g) = \Delta(g_D, g), H(g_R, g) = 0 \) is used to indicate that \( g \) is the xenolog of the reference gene, \( g \), and that its class is given by \( H(g_D, g) \).

In Figure 1, all genes are xenologous to \( g \). Both \( g_D \) and \( g_R \) are in set \( D \); \( g_D \) is a Primary xenolog (\( X(g_D, g) = PX \)) and \( g_R \) is a Sibling Donor xenolog (\( X(g_D, g) = SDX \)), because \( g_D \) is a descendant of the donor \( g \in D \) and \( g_R \) is not. Genes \( g_D \) and \( g_R \) are in set \( R \) and are Sibling Recipient xenologs (\( X(g_R, g) = SRX \)), because \( g_R \) is in set \( O \) and \( g_R \) is a Sibling Recipient xenolog (\( X(g_R, g) = SRX \)).

### Xenolog class hierarchies

The xenolog classes form a hierarchy that can elucidate how xenologs are related in both the gene and species trees. Primary xenologs are closest in the xenolog hierarchy and Outgroup xenologs are most distant. We denote this hierarchy by

\[
PX < SDX < SRX < OX
\]

where \( X(g_D, g) \leq X(g_R, g) \), if \( g_D \) and \( g_R \) are closer in the hierarchy than \( g_D \) and \( g_R \).

Genes that are more closely related in the hierarchy are also more closely related in the gene tree. Let genes \( g_D \) and \( g_R \) in \( V_S \setminus \Delta(g_D) \) be xenologs such that there is no transfer ancestral to either \( g_D \) or \( g_R \). Then, \( \text{MRCA}(g_D) < \Delta(g_D) \), if \( g_D \) and \( g_R \) are in set \( O \). This hierarchy, which is illustrated in Figure 2, is stated formally as follows:

**Theorem 2.1** (Xenolog class hierarchy in the gene tree) Given \( g \in \Delta(g) \), for any Primary xenolog, \( g_D \), Sibling Donor xenolog, \( g_D \), Sibling Recipient xenolog, \( g_R \), and Outgroup xenolog, \( g_D, g_R, \text{if } g \)

\[
\text{MRCA}(g_D, g) < \Delta(g_D, g) < \Delta(g_D, g_R) < \Delta(g_D, g_R), \text{ if } g
\]

Proof. See Section 2.1.

We sketch the basis of this theorem informally, here. For every xenolog \( g \in V_S \setminus \Delta(g) \), the common ancestor of \( g \) and \( g_D \) is a node on the path from \( g_D \) to the root of \( T_G \); i.e., there exists \( g \in V_S \), such that \( g_D = \text{MRCA}(g_D, g) \) and \( g = \Delta(g_D, g) \). If \( g = g_D \), then \( g \notin \Delta(g_D) \); \( \Delta(g_D) \) and is therefore a Primary xenolog.

For \( g_D \), the descendants of \( g_D \) that is incomparable to the transfer, must satisfy two requirements. First, since all xenologs in \( \Delta(g_D) \) are equally related to \( g_D \), all xenologs in \( \Delta(g_D) \) must be assigned to
the same xenolog class. This will be true if all ascendants of $c_i$ are in the same species set, $D, R$ or $O$. Second, for any $g_j \triangleright g_i$, the xenologs in $\Delta(g_j)$ are more distantly related to $\hat{g}$ than the xenologs in $\Delta(g_i)$; therefore, consistency requires that the class of xenologs in $\Delta(g_i)$ not be closer in the hierarchy than the class of xenologs in $\Delta(g_j)$. Both of these conditions are satisfied when there is no transfer that is ancestral to either $g_1$ or $g_2$. This is always true in a reconciled tree with a single transfer and no duplications. We will reexamine the hierarchical properties of xenolog classes in trees with more complex event histories in the following sections.

The proposed xenolog classes also convey information about the relationship of a xenolog pair in the gene tree relative to their relationship in the species tree. If $g_1, g_2$, and $g_3$ are orthologs, then $g_1$ and $g_2$ are more closely related than $g_3$, iff the associated species $\hat{t}(g_1)$ and $\hat{t}(g_2)$ are more closely related than $\hat{t}(g_3)$. As a result, knowing the species associated with a pair of orthologs provides a quick estimate of the time since their divergence (Capra et al., 2013). This is not true of xenologs, where the xenolog of $\hat{g}$ and $\hat{c}$ can predate or postdate the species containing MRCA($\hat{g}, \hat{c}$). Our xenolog classes distinguish between these three cases: Primary and Sibling Donor xenologs are more closely related in the gene tree than in the species tree, whereas Sibling Recipient xenologs are more closely related in the species tree than in the gene tree. Outgroup xenologs are equally related in both trees. These relationships are summarized in Table S1.

2.2 Xenolog classification with transfers and duplications

We next consider the classification of genes $g_i$ and $g_j$ when there is a single transfer on the path from $g_i$ to $g_j$ and they diverge by duplication (i.e., $\hat{t}(\text{MRCA}(\hat{g}_i, \hat{g}_j)) = \Delta$). Such gene pairs satisfy both the paradox and the xenolog criteria proposed by Fitch (2000), leading to potential terminological confusion. To avoid this confusion, we introduce the explicit designation \textit{paraxenolog}¹, for xenologs that diverged via a duplication at their common ancestor.

¹ Patterson (1988) used “paraxenolog” to refer to a different phenomenon.
Incomparable transfers: We first consider the special case where \( k = 2 \) and the transfers are incomparable. Given a pair of genes, \( g_1 \) and \( g_2 \), connected by two incomparable transfers, \( t^1 \) and \( t^2 \) (Figure 5), one gene is a descendant of one transfer recipient (\( g_1 \in \Delta(g^1_1) \)), and the other gene is a descendant of the other transfer recipient (\( g_2 \in \Delta(g^2_1) \)). Since \( g_1 \) and \( g_2 \) are both descendants of a transfer recipient, xenolog \( g_2 \) can be classified with respect to \( g_1 \), \( g_1 \), and \( \text{vice versa} \).

With incomparable transfers, the xenolog classes do not satisfy the hierarchical properties of Theorem 2.1. As before, let \( g_1 = \text{MRCA}(\{\hat{g}_c, g_2\}) \) and let \( c_i \) be the child of \( g_i \) that is ancestral to \( t^1 \) but not \( t^2 \) (i.e., \( c_i \preceq g_2 \) and \( c_i \nsubseteq g_1 \)). Recall that the first condition for preservation of the hierarchy is that all xenologs in \( \Delta(c_i) \) must be in the same species set. Satisfaction of this condition is not guaranteed for incomparable xenologs because \( \Delta(c_i) \) contains a transfer, \( t^1 \), that can move \( g^2_2 \) to a species that is not in the same set as \( \Delta(g^1_1) \). Suppose, for example, the donor of \( t^1 \) is in a species in \( O \), but its recipient is in a species in \( D \). Since both \( g^2_1 \) and \( g^2_2 \) are in \( \Delta(c_i) \), more than one species set is represented in \( \Delta(c_i) \), violating the first condition. Primary xenologs are the one exception to this problem. Primary xenologs are defined in terms of \( g_2 \) and not in terms of \( D, R \), and \( O \), and are therefore unaffected by incomparable transfers. Primary xenologs are always more closely related to \( \hat{g} \) than are xenologs of any other class, in both the comparable and incomparable cases.

To avoid a classification that violates the hierarchy, we do not assign xenologs separated by incomparable transfers to specific subclasses. Given two genes separated by incomparable transfers, \( t^1 \) and \( t^2 \), without loss of generality, let \( \hat{g}_c \in \Delta(g^1_1) \) be the reference gene, \( g_2 \in \Delta(g^2_1) \) be the xenolog under classification, and \( g_\text{MRCA}(\{g_1, g_2\}) = \text{their common ancestor} \). Then \( g_1 \) is a

- Primary xenolog iff \( g_2 \in \Delta(g^1_1) \):
  \[ X(\hat{g}_c; g_2) = \bot \]
- Incomparable xenolog iff \( g_2 \notin \Delta(g^1_1) \) and \( \tau(g_\alpha) = \tau \):
  \[ X(\hat{g}_c; g_2) = \text{IX} \]
- Incomparable paraxenolog iff \( g_2 \notin \Delta(g^1_1) \) and \( \tau(g_\alpha) = \delta \):
  \[ X(\hat{g}_c; g_2) = \text{IX}^2 \]

In the incomparable case, \( H[\hat{g}_c; g_2] = X(\hat{g}_c; g_2) \) is the classification of \( g_2 \) with respect to \( \hat{g}_c \) and \( F(\hat{g}_c; g_2) = X(\hat{g}_c; g_2) \) is the classification of \( g_2 \) with respect to \( \hat{g}_c \). Either \( X(\hat{g}_c; g_1) = \bot \) and \( X(\hat{g}_c; g_2) = \text{IX} \) (or vice versa), or \( X(\hat{g}_c; g_1) = X(\hat{g}_c; g_2) = \text{IX}^2 \).

We now address the case where \( k > 2 \) by reducing the problem to one involving two incomparable super-transfers and applying the protocol just described. Let \( t^1 \cdot t^2 \) be the transfers, in descending order, on the path from \( \text{MRCA}(\{g_1, g_2\}) \) to \( g_1 \) and \( \text{MRCA}(\{g_1, g_2\}) \) to \( g_2 \). Since \( t^1 \cdot t^2 \) must be mutually comparable, they can be replaced with super-transfer \( t^{1+} = (g^1_{1'}, g^2_{1'}) \), where \( g^1_{1'} = g^1_{1} \) and \( g^2_{1'} = g^2_{1} \). Similarly, we replace \( t^{2+} \cdot t^3 \) with super-transfer \( t^{2+} = (g^2_{2'}, g^2_{2'}) \), where \( g^2_{2'} = g^2_{2} \) and \( g^2_{2'} = g^2_{2} \). Gene tree hierarchy for multiple transfers: With multiple comparable transfers, the hierarchical properties in Theorem 2.1 hold for xenologs that share the same super-transfer from \( \text{MRCA}(\hat{g}_c; \hat{g}) \) to \( \hat{g} \). For example, in Figure 4, the xenolog class hierarchy is preserved for nodes \( g_3 \) and \( g_4 \), which are xenologs of \( \hat{g} \) with respect to \( t^2 \) only. Similarly, xenologs \( g_5, g_6, g_7, \) and \( g_8 \), which are all defined with respect to the super-transfer \( t^1 \), also obey the hierarchy. However, \( g_7 \) and \( g_8 \) do not share the super-transfer and thus, do not obey the hierarchy, \( \text{MRCA}(\hat{g}_c; g_7) \), yet \( X(\hat{g}_c; g_7) = \text{SDX} \cdot \text{X}(\hat{g}_c; g_8) = \text{PX} \).
Incomparable xenologs that are not Primary may fall anywhere in the tree is consistent with the hierarchy. Classes, but guarantees a classification in which relatedness in the gene tree is consistent with the ancestor containing MRCA in incomparable relationship, is possible between the commonest and the ascendant of MRCA(g, g). Since g, g, or g are separated by both transfers. Since g, g, g, g, then H(g, g) gives the xenolog class of g with respect to g, a reference gene that is the recipient of at least one transfer on the path from MRCA(g, g) to g. If there is also a transfer on the path from MRCA(g, g) to g, then H(g, g) = x(g, g) gives the xenolog class of g with respect to reference g. Otherwise, H(g, g) = ∗.

Primary xenologs, including those connected by incomparable transfers, are more closely related than any other class of xenologs. Incomparable xenologs that are not Primary may fall anywhere in the hierarchy; that is, a given pair of Incomparable xenologs may be more closely related, or more distantly related, than a given pair of Sibling or Outgroup xenologs. Thus, the Incomparable xenolog class provides less information about relatedness than the specific Sibling and Outgroup classes, but guarantees a classification in which relatedness in the gene tree is consistent with the hierarchy.

The species tree hierarchy for single transfers (Table S1) also holds for multiple comparable transfers summarized by a super-transfer, with one exception. When the recipient species of the super-transfer is a descendant of the donor species (as in Figure S2), Primary xenologs, with respect to this super-transfer, are more or equally related in the species tree than in the gene tree.

The species tree hierarchy is not guaranteed for multiple, incomparable transfers, even when the pair are classified as Primary xenologs. The reason for this is that the recipient of t can be in any of the sets, D, D, or D, defined by t. Therefore the cenanccestor of g and g can be in any species in V3. Any relationship, even an incomparable relationship, is possible between the cenanccestor and the ancestor containing MRCA(g, g).

3 Algorithms and Implementation

The classification procedure for the xenolog classes described in Section 2 is shown as pseudocode in Section S.4. We have implemented this procedure and integrated it in Notung 2.8, a freely available software package that implements gene tree-species tree reconciliation with transfers in a parsimony framework (Stolzer et al., 2012).

Upon reconciling a gene tree with a species tree, Notung 2.8 generates a homology table, H, for all pairs of leaves in the gene tree. There may be more than one minimum-cost event history that reconciles the gene and species trees. A homology table is generated for each optimal, temporally feasible reconciliation reported. Transfers imply temporal constraints because the donor and recipient of a transfer must have co-existed; a reconciliation is temporally feasible if all temporal constraints imposed by the inferred transfers are mutually compatible. Notung 2.8 reports all optimal reconciliations that are temporally feasible, up to a user-specified limit (Stolzer, 2012).

Homology tables can be viewed in the graphical user interface or exported from the command line in a tab-delimited, CSV, or HTML format. Row H[g, g] contains the homology relationships between reference gene, g, and all other genes in V3. For orthologs and paralogs, H[g, g] = H[g, g]. For xenologs, H[g, g] = x(g, g) gives the xenolog class of g with respect to g, a reference gene that is the recipient of at least one transfer on the path from MRCA(g, g) to g. If there is also a transfer on the path from MRCA(g, g) to g, then H[g, g] = x(g, g) gives the xenolog class of g with respect to reference g. Otherwise, H[g, g] = ∗.

The classification procedure is generally applicable to reconciled gene trees and can be implemented in any reconciliation software package that enforces temporal consistency. When temporal consistency is not enforced, reconciliations with transfers between ancestor and descendant species can arise. Since this scenario is similar to super-transfers that form a loop, the classification procedure proposed here could easily be adapted for programs that do not enforce consistency.

4 Empirical Results

Genomic Study: As a proof of principle, we analyzed 13,623 gene families from a dataset of 65 genomes of Proteobacteria and Cyanobacteria (Latysheva et al., 2012). To control for spurious inference of transfers due to phylogenetic error, weakly supported branches were rearranged using a species-tree aware method as described in Section S.5.1. The resulting rooted, rearranged trees were then reconciled with the species tree with default costs (C = 3, C = 1.5, C = 1). These costs are consistent with costs used in other recent phylogenomic analyses (David and Alm, 2011; Richards et al., 2014), which were selected to minimize the total net change in genome content. The time required to reconcile the 13,623 trees, including generating all optimal reconciliations and testing them for temporal feasibility, was 7.25 minutes on an Intel Xeon 2.3GHz processor (128GB RAM). The computational complexity of calculating the homology table, once the gene tree has been reconciled, is negligible.

Homology tables were computed for the 13,194 trees possessing at least one temporally feasible solution. From these, homologs of all categories were tabulated. For families with more than one optimal reconciliation, the number of pairs in each category was averaged over all reported, optimal event histories.
Orthologs, paralogs, and xenologs are all represented in this dataset, and every xenolog class is also observed (Figure 6 and Tables S2 – S6). More than a quarter of homologous gene pairs were xenologs. Of these pairs, 85.7% are xenologs with only one reference gene, where all transfers on the path from the reference to its xenolog are mutually comparable. Of these xenologs, 60.2% are either Primary or Sibling Donor (para)xenologs; thus, the majority of the inferred xenologs are closer to the donor than the recipient.

Gene pairs separated by incomparable transfers are fairly rare compared with all types of xenologs separated by any number of transfers. Such pairs have two xenologs, one for each reference gene; at most one member of each pair can be classified as a Primary xenolog (PX), otherwise they are untyped (IX). The fraction of incomparable xenologs for which the hierarchy provides no information is quite small: 72.0% of incomparable (para)xenologs are (PX, IX) pairs; the rest are (IX, IX) or (IX+, IX+) pairs.

Less than 1% of all xenologous pairs are autonologs, which could be due to preferential transfer of novel genes or a high incidence of xenologous gene transfer. Thus, the majority of the inferred xenologs are closer to the donor than the recipient. These xenologs, 60.2% are either Primary or Sibling Donor (para)xenologs; the rest are untyped (IX) or (IX+, IX+) pairs. Incomparable xenologs transfer cost resulted in a moderate increase in the number of non-paralogous xenologs inferred. The change in the relative frequencies of the other various classes was generally small (less than 15%) with one exception: the proportion of Outgroup xenologs decreased by more than 50%. The increase in para(xenologs and decrease in Outgroup xenologs, taken together, suggests that more duplications may be inferred near the roots of gene trees with a higher transfer cost. Thus, in this analysis, the trade-off between duplications and transfers does not affect all xenolog classes equally.

**BIO4 Case Study**: To explore the connection between xenolog classes and protein function, we applied our approach to the BIO4 gene family; several BIO4 genes have been horizontally transferred and have been characterized experimentally (Hall and Dietrich, 2007). The BIO4 protein is an enzyme in the biotin (vitamin B7) biosynthesis pathway (Figure S11). Plants and some fungi possess a BIO4 homolog that encodes a bi-functional enzyme, which acts as both a 7,8-Diaminopelargonic acid synthase (DAPAS) and a dethiobiotin synthetase (DTBS), steps 3 and 4 in the pathway, respectively. In bacteria, the BIO4 homolog only performs the DTBS function; the 3rd step is carried out by an unrelated protein. Unlike other fungi, however, the BIO4 homolog in yeast (Saccharomyces cerevisiae, and its close relatives) also encodes a DTBS-only protein. Phylogenetic analysis shows that a horizontal transfer from bacteria to yeast replaced the ancestral bi-functional homolog (Hall and Dietrich, 2007). Using Notung 2.8, we reconciled the gene and species trees (Figures S12 and S13) constructed by Hall and Dietrich (2007) and inferred xenolog classes (Figures 7 and S14).

The hierarchical nature of the xenolog classification aids in the interpretation of the functional evolution of the family in this case study. The molecular function of yeast BIO4 is closer to that of Sibling Donor xenologs, which encode the DTBS-only enzyme, than its Sibling Recipient xenologs, which encode bi-functional enzymes. In contrast, the Sibling Recipient xenologs provide information about genomic context. The fact that the Sibling Recipient xenologs encode a bi-functional enzyme raises a red flag: the replacement of a bi-functional enzyme with a DTBS-only enzyme in yeast suggests loss of the DAPAS function. Either a different enzyme must be carrying out the DAPAS function or yeast no longer has a functional biotin synthesis pathway. In fact, the former is true; the DAPAS function is performed by an unrelated gene, that was also acquired horizontally (Hall and Dietrich, 2007).

In this example, a closely related gene (a DTBS-only enzyme) in a distantly related (α-proteobacterial) species is a better predictor of BIO4 enzymatic function than a distantly related gene (the dual function homolog) in a closely related species (Kerovia lipolytica). The distantly related homolog in a closely related species provides information about...
the genetic background; i.e., the genome could be lacking a gene encoding the DAPAS function. These insights are linked to the hierarchical structure of the xenolog classes and may represent general trends, suggesting hypotheses for future investigation. If it proves generally true, for example, that Sibling Donors are better predictors of molecular function and Sibling Recipients are better predictors of cellular context, then this system of xenolog classification could support large scale, automated analyses in comparative, evolutionary genomics.

5 Discussion

Distinguishing orthologs from paralogs, as well as the division of paralogs into subclasses based on the timing and nature of the events by which they arise, has proved to be a valuable analytical approach in molecular evolution, systematics, comparative genomics, and homology-based function prediction.

When temporal consistency is enforced, the sets of genes that share a history with at least one transfer encompasses a very broad set of relationships.

In this work, we propose subtypes that provide a more nuanced classification of xenologs. We provide formal rules for classification, given a reconciled gene tree with an arbitrary number of transfers and duplications. These rules have been implemented in Notung 2.8, a freely available phylogenetic reconciliation software package.

Consistent with the framework Fitch first introduced in the 1970s, phylogenetic reconciliation captures information about the historical association between genes and species, as well as the divergence events that characterize the xenologs in each class. A potential limitation of this approach is that it requires that species evolution be modeled as a tree. While some have argued against tree-like models, given the prevalence of horizontal gene transfer in bacteria, a tree can provide a useful heuristic, despite the reticulate nature of prokaryotic evolution (Mindell, 2013, and work cited therein).

As with most theoretical work on reconciliation, our classification assumes that the gene tree and the inferred events are correct. In practice, this assumption is not valid. For example, incomplete lineage sorting could lead to downstream errors in xenolog classification. Just as information about the spatial organization of relationships.

4 Acknowledgments

We thank Minli Xu for his work preprocessing the bacterial genomic dataset, Rosie Alderson for help with biochemical information, Vivien Junker for assistance with phylogeny reconstruction, Annette McLeod for graphic design, and EuSCHEM for computational support via the EuSCHEM Research Computing Facility.

Funding

This material is based upon work supported by the National Science Foundation under Grant No. DBI-1262593. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References


Chen, X. and Zhang, J. (2012). The ortholog conjecture is unstable by the current gene ontology and is supported by RNA sequencing data. PLOS Comput Biol, 8(11).


