Chapter 9

Inferring Orthology and Paralogy

Adrian M. Altenhoff and Christophe Dessimoz

Abstract

The distinction between orthologs and paralogs, genes that started diverging by speciation versus duplication, is relevant in a wide range of contexts, most notably phylogenetic tree inference and protein function annotation. In this chapter, we provide an overview of the methods used to infer orthology and paralogy. We survey both graph-based approaches (and their various grouping strategies) and tree-based approaches, which solve the more general problem of gene/species tree reconciliation. We discuss conceptual differences among the various orthology inference methods and databases, and examine the difficult issue of verifying and benchmarking orthology predictions. Finally, we review typical applications of orthologous genes, groups, and reconciled trees and conclude with thoughts on future methodological developments.

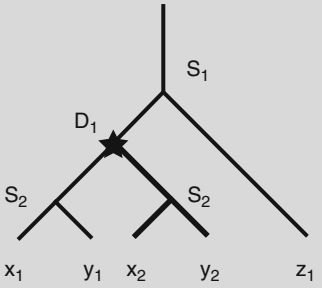
Key words: Orthology, Paralogy, Tree reconciliation, Orthology benchmarking

1. Introduction

The study of genetic material almost always starts with identifying, within or across species, *homologous* regions—regions of common ancestry. As we have seen in previous chapters, this can be done at the level of genome segments (Chapter 8, this volume; ref. 1), genes (Chapter 6, this volume; ref. 2), or even down to single residues, in sequence alignments (Chapter 7, this volume; ref. 3). Here, we focus on genes as evolutionary and functional units. The central premise of this chapter is that it is useful to distinguish between two classes of homologous genes: *orthologs*, which are pairs of genes that started diverging via evolutionary speciation, and *paralogs*, which are pairs of genes that started diverging via gene duplication (4) (Box 1). Originally, the terms and their definition were proposed by Walter M. Fitch in the context of species phylogeny inference, i.e., the reconstruction of the tree of life. He stated, "Phylogenies require orthologous, not paralogous, genes" (4). Indeed, since orthologs arise by speciation, any set of genes in which every pair is orthologous has by definition the same

Maria Anisimova (ed.), *Evolutionary Genomics: Statistical and Computational Methods, Volume 1*,
Methods in Molecular Biology, vol. 855, DOI 10.1007/978-1-61779-582-4_9,

Box 1
Terminology



Homology is a relation between a pair of genes that share a common ancestor. All pairs of genes in the figure above are homologous to each other.

Orthology is a relation defined over a pair of homologous genes, where the two genes have emerged through a speciation event (4). Example pairs of orthologs are (x_1, y_1) or (x_2, z_1). Orthologs can be further subclassified into one-to-one, one-to-many, many-to-one and many-to-many orthologs. The qualifiers *one* and *many* indicate for each of the two involved genes whether they underwent an additional duplication after the speciation between the two genomes. Hence, the gene pair (x_1, y_1) is an example of a one-to-one orthologous pair, whereas (x_2, z_1) is a many-to-one ortholog relation.

Paralogy is a relation defined over a pair of homologous genes that have emerged through a gene duplication, e.g., (x_1, x_2) or (x_1, y_2).

In-paralogy is a relation defined over a triplet. It involves a pair of genes and a speciation event of reference. A gene pair is an in-paralog if they are paralogs and duplicated *after* the speciation event of reference. The pair (x_1, y_2) are in-paralogs with respect to the speciation event S_1.

Out-paralogy is also a relation defined over a pair of genes and a speciation event of reference. This pair are out-paralogs if the duplication event through which they are related to each other *predates* the speciation event of reference. Hence, the pair (x_1, y_2) are out-paralogs with respect to the speciation event S_2.

Co-orthology is a relation defined over three genes, where two of them are in-paralogs with respect to the speciation event associated to the third gene. The two in-paralogous genes are said to be *co-orthologous* to the third (out-group) gene. Thus, x_1 and y_2 are co-orthologs with respect to z_1.

evolutionary history as the underlying species. These days, however, the most frequent motivation for the orthology/paralogy distinction is to study and predict gene function: it is generally believed that orthologs—because they were the same gene in the last common ancestor of the species involved—are likely to have similar biological function. By contrast, paralogs—because they result from duplicated genes that have been retained, at least partly, over the course of evolution—are believed to often differ in

function. Consequently, orthologs are of interest to infer function computationally while paralogs are commonly used to study function innovation.

In this chapter, we first review the main methods used to infer orthology and paralogy. We then discuss the problem of benchmarking orthology inference. In the last main section, we focus on various applications of orthology and paralogy.

2. Inferring Orthology

Most orthology inference methods can be classified into two major types: graph-based methods and tree-based methods (5). Methods of the first type rely on graphs with genes (or proteins) as nodes and evolutionary relationships as edge. They infer whether these edges represent orthology or paralogy, and build clusters of genes on the basis of the graph. Methods of the second type are based on gene/species tree reconciliation, which is the process of annotating all splits of a given gene tree as duplication or speciation, given the phylogeny of the relevant species. From the reconciled tree, it is trivial to derive all pairs of orthologous and paralogous genes. All pairs of genes which coalesce in a speciation node are orthologs, and paralogs if they split at a duplication node. In this section, we present the concepts and methods associated with the two types, and discuss the advantages, limitations, and challenges associated with them.

2.1. Graph-Based Methods

Graph-based approaches were originally motivated by the availability of complete genome sequences and the need for efficient methods to detect orthology. They typically run in two phases: a graph construction phase, in which pairs of orthologous genes are inferred (implicitly or explicitly) and connected by edges, and a clustering phase, in which groups of orthologous genes are constructed based on the structure of the graph.

2.1.1. Graph-Construction Phase: Orthology Inference

In its most basic form, the graph-construction phase identifies orthologous genes by considering pairs of genomes at a time. The main idea is that between any given two genomes, the orthologs tend to be the homologs that diverged least. Why? Because, assuming that speciation and duplication are the only types of branching events, the orthologs branched by definition at the latest possible time point—the speciation between the two genomes in question. Therefore, using sequence similarity score as surrogate measure of closeness, the basic approach consists in identifying the corresponding ortholog of each gene through its genome-wide best hit (BeT), its highest scoring match in the other genome (6).

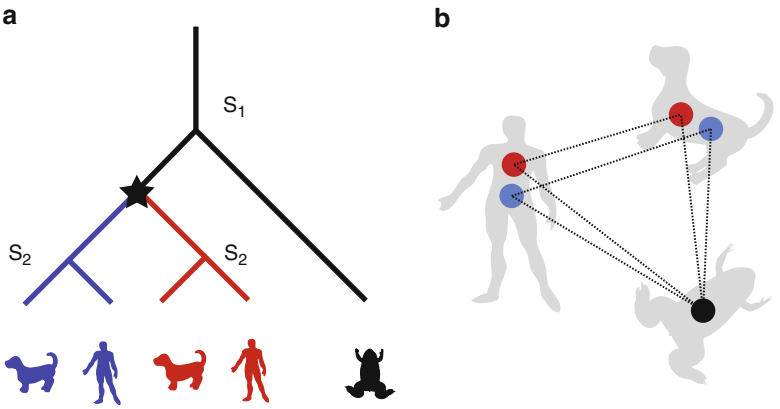


Fig. 1. (**a**) Simple evolutionary scenario of a gene family with two speciation events (S_1 and S_2) and one duplication event (*star*). The type of events completely and unambiguously defines all pairs of orthologs and paralogs: the frog gene is orthologous to all other genes (it coalesces at S_1). The *red* and *blue* genes are orthologs between themselves (they coalesce at S_2), but paralogs between each other (they coalesce at *star*). (**b**) The corresponding orthology graph. The genes are represented here by *vertices*, and orthology relationships by *edges*. The frog gene forms *one-to-many* orthology with both the human and dog genes because it is orthologous to more than one sequence in each of these organisms. In such cases, the *bidirectional best-hit* approach only recovers one of the relations (the highest scoring one). Note that in contrary to BBH the nonsymmetric BeTs approach would, in the situation of a lost *blue* human gene, infer an incorrect orthologous relation between the *blue* dog and *red* human gene.

To make the inference symmetric (as orthology is a symmetric relation), it is usually required that BeTs are reciprocal, i.e., that orthology is inferred for a pair of gene g_1, g_2 if and only if g_2 is the BeT of g_1 and g_1 is the BeT of g_2 (7). This symmetric variant, referred to as bidirectional best hit (BBH), has also the merit of being more robust against a possible gene loss in one of the two lineages (Fig. 1).

Inferring orthology from BBH is computationally efficient because each genome pair can be processed independently, and high-scoring alignments can be computed efficiently using dynamic programming (8) or heuristics, such as BLAST (9). Overall, the time complexity scales quadratically in terms of the total number of genes. Furthermore, the implementation of this kind of algorithm is simple.

However, orthology inference by BBH has several limitations, which motivated the development of various improvements (Table 1).

Allowing for More Than One Ortholog

Some genes can have more than one orthologous counterpart in a given genome. This happens whenever a gene undergoes duplication *after* the speciation of the two genomes in question. Since BBH only picks the BeT, it only captures part of the orthologous relations (Fig. 1). The existence of multiple orthologous counterparts is often referred to as *one-to-many* or *many-to-many*

Table 1
Overview of graph-based orthology inference methods and their main properties

Method	In-paralogs	Based on	Grouping strategy	Database	Extra	Available Algo/DB	Reference
COG	Yes	BLAST scores	Merged adjacent triangles of BeTs	COG/KOG		X/X	(6)
BBH	No	BLAST scores	n.a.	n.a.		−/−	(7)
Inparanoid	Yes	BLAST scores	Only between pairs of species	Inparanoid		X/X	(10, 73)
RSD	No	ML distance estimates	n.a.	RoundUp		X/X	(13, 74)
OMA	Yes	ML distance estimates	Every pair is ortholog	OMA Browser	Detects differential gene loss	−/X	(11, 75)
OrthoMCL	Yes	BLAST scores	MCL clusters	OrthoMCL-DB		X/X	(18, 76)
EggNOG	Yes	BLAST scores	Merged adjacent triangles of BeTs	EggNOG	Computed at several levels of taxonomic tree	−/X	(21, 77)
OrthoDB	Yes	Smith Waterman scores	Merged adjacent triangles of BeTs	OrthoDB	Computed at any level of taxonomic tree	−/X	(22)
COCO-CL	Yes	MSA-induced scores	Hierarchical clusters	n.a.		X/−	(23)
OrthoInspector	Yes	BLAST scores	Only between pairs of species	OrthoInspector		X/X	(78)

n.a. not applicable

orthology, depending whether duplication took place in one or both lineages. To designate the copies resulting from such duplications occurring *after* a speciation of reference, Remm et al. (10) coined the term *in-paralogs* and introduced a method called *Inparanoid* that improves upon BBH by potentially identifying all pairs of many-to-many orthologs. In brief, their algorithm identifies all paralogs within a species that are evolutionarily closer (more similar) to each other than to the BBH gene in the other genome. This results in two sets of in-paralogs—one for each species—whose Cartesian product gives all orthologous relations. Alternatively, it is possible to identify many-to-many orthology by relaxing the notion of "BeT" to "group of BeTs." This can be implemented using a score tolerance threshold or a confidence interval around the BBH (11, 12).

Evolutionary Distances

Instead of using sequence similarity as a surrogate for evolutionary distance to identify the closest gene(s), Wall et al. (13) proposed to use direct and proper maximum likelihood estimates of the evolutionary distance between pairs of sequences. Indeed, previous studies have shown that the highest scoring alignment is often not the nearest phylogenetic neighbor (14). Building upon this work, Roth et al. (15) showed how statistical uncertainties in the distance estimation can be incorporated into the inference strategy.

Differential Gene Losses

As discussed above, one of the advantages of BBH over BeT is that by virtue of the bidirectional requirement the former is more robust to gene losses in one of the two lineages. But if gene losses occurred along both lineages, it can happen that a pair of genes mutually closest to one another are in fact paralogs, simply because both their corresponding orthologs were lost—a situation referred to as "differential gene losses." Dessimoz et al. (16) presented a way to detect some of these cases by looking for a third species in which the corresponding orthologs have not been lost and thus can act as *witnesses of nonorthology*.

2.1.2. Clustering Phase: From Pairs to Groups

The graph-construction phase yields orthologous relationships between pairs of genes. But this is often not sufficient. Conceptually, information obtained from multiple genes or organisms is often more powerful than that obtained from pairwise comparisons only. In particular, as the use of a third genome as potential witness of nonorthology suggests, a more global view can allow identification and correction of inconsistent/spurious predictions. Practically, it is more intuitive and convenient to work with groups of genes than with a list of gene pairs. Therefore, it is often desirable to cluster orthologous genes into groups.

Tatusov et al. (6) introduced the concept of clusters of orthologous groups (COGs). COGs are computed by using triangles (triplets of genes connected to each other) as seeds, and then

merging triangles which share a common face, until no more triangle can be added. This clustering can be computed relatively efficient in time $O(n^3)$, where n is the number of genomes analyzed (17). The stated objective of this clustering procedure is to group genes that have diverged from a single gene in the last common ancestor of the species represented (6). Practically, the COGs have been found to be useful by many, most notably, to categorize prokaryotic genes into broad functional categories.

A different clustering approach was adopted by *OrthoMCL*, another well-established graph-based orthology inference method (18). There, groups of orthologs are identified by Markov Clustering (19). In essence, the method consists in simulating a random walk on the orthology graph, where the edges are weighted according to similarity scores. The Markov Clustering process gives rise to probabilities that two genes belong to the same cluster. The graph is then partitioned according to these probabilities and members of each partition form an orthologous group. These groups contain orthologs and "recent" paralogous genes, where the recency of the paralogs can be somewhat controlled through the parameters of the clustering process.

A third grouping strategy consists in building groups by identifying fully connected subgraphs (called "cliques" in graph theory) (11). This approach has the merits of straightforward interpretation (groups of genes which are all orthologous to one another) and high confidence in terms of orthology within the resulting groups due to the high consistency required to form a fully connected subgraph. But it has the drawbacks of being hard to compute (Clique finding belongs to the NP-complete class of problems, for which no polynomial-time algorithm is known) and being excessively conservative for many applications.

As emerges from these various strategies, there is more than one way orthologous groups can be defined, each with different implications in terms of group properties and applications (20). In fact, there is an inherent trade-off in partitioning the orthology graph into clusters of genes because orthology is a nontransitive relation: if genes A and B are orthologs and genes B and C are orthologs, genes A and C are not necessarily orthologs, e.g., consider in Fig. 1 the blue human gene, the frog gene, and the red dog gene. Therefore, if groups are defined as sets of genes in which all pairs of genes are orthologs (as with OMA groups), it is not possible to partition A, B, and C into groups capturing all orthologous relations while leaving out all paralogous relations.

More inclusive grouping strategies necessarily lead to orthologs and paralogs within the same group. Nevertheless, it can be possible to control the nature of the paralogs included. For instance, as seen above, OrthoMCL attempts at including only "recent" paralogs in its groups. This idea can be specified more precisely by defining groups with respect to a particular speciation event of

interest, e.g., the base of the mammals. Such *hierarchical groups* are expected to include orthologs and in-paralogs with respect to the reference speciation—in our example, all copies that have descended from a single common ancestor gene in the last mammalian common ancestor. EggNOG (21) and OrthoDB (22), for example, both implement this concept by applying the COG clustering method for various taxonomic ranges. Another method, COCO-CL, identifies hierarchical orthologous groups recursively using correlations of similarity scores among homologous genes (23) and, interestingly, without relying on a species tree. By capturing part of the gene tree structure in the group hierarchies, these methods try in some way to bridge the gap between graph-based and tree-based orthology inference approaches. We now turn our attention to the latter.

2.2. Tree-Based Methods

At their core, tree-based orthology inference methods seek to reconcile gene and species trees. Reconciliation is needed because in most cases gene and species trees have different topologies due to evolutionary events acting specifically on genes, such as duplications, losses, lateral transfers, or incomplete lineage sorting (24). Goodman et al. (25) pioneered research to resolve these incongruences. They showed how the incongruences can be explained in terms of speciation, duplication, and loss events on the gene tree (Fig. 2), and provided an algorithm to infer such events. Once all branchings of the gene tree have been inferred as speciation or duplication event, it is trivial to establish whether a pair of genes is orthologous or paralogous, based on the type of the branching where they coalesce. Therefore, orthology/paralogy inference can be reduced to tree reconciliation.

Most tree reconciliation methods rely on a parsimony criterion: the most likely reconciliation is the one which requires the least number of gene duplications and losses. This makes it possible to compute reconciliation efficiently, and is tenable as long as duplication and loss events are rare compared to speciation events. In their seminal article, Goodman et al. (25) had already devised their reconciliation algorithm under a parsimony strategy. In the subsequent years, the problem was formalized in terms of a map function between the gene and species trees (26), whose cost was conjectured (27), and later proved (28, 29) to coincide with the number of gene duplication and losses. With the proofs came highly efficient algorithms, either in terms of asymptotic time complexity with an $O(n)$ algorithm (28) or in terms of actual runtime on typical problem sizes (30). With these near-optimal solutions, one could think that the tree reconciliation problem has long been solved. As we shall see in the rest of this section, however, the original formulation of the tree reconciliation problem has several limitations in practice, which have stimulated the development of various refinements to overcome them (Table 2).

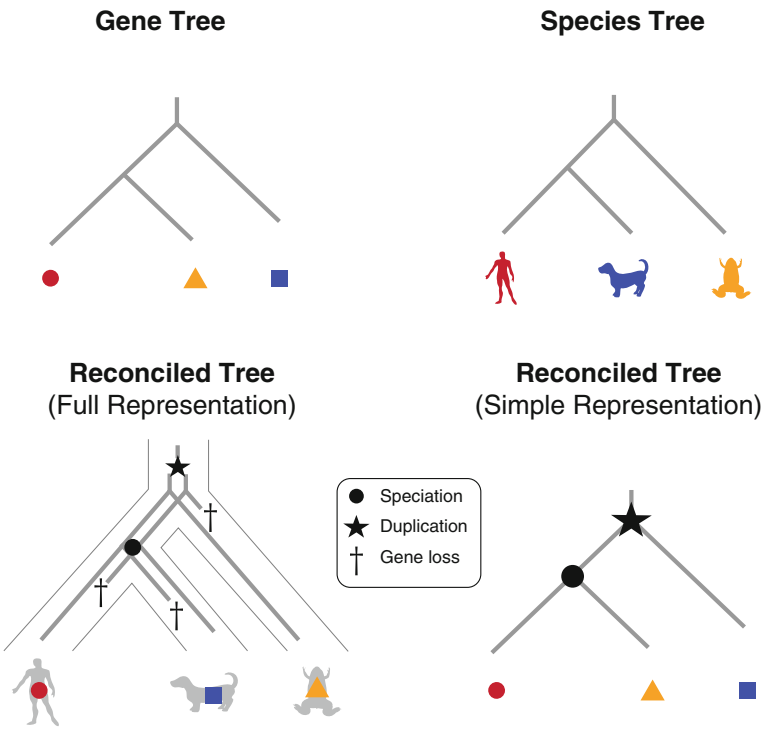


Fig. 2. Schematic example of the gene/species tree reconciliation. The gene tree and species tree are not compatible. Reconciliation methods resolve the incongruence between the two by inferring speciation, duplication, and losses events on the gene tree. The reconciled tree indicates the most parsimonious history of this gene, constrained to the species tree. The simple representation (*bottom right*) suggests that the human and frog genes are orthologs, and that they are both paralogous to the dog gene.

2.2.1. Unresolved Species Tree

A first problem ignored by most early reconciliation algorithms lies in the uncertainty often associated with the species tree, which these methods assume as correct and heavily rely upon.

One way of dealing with the uncertainties is to treat unresolved parts of the species tree as multifurcating nodes (also known as *soft polytomies*). By doing so, the reconciliation algorithm is not forced to choose for a specific type of evolutionary event in ambiguous regions of the tree. This approach is, for instance, implemented in *TreeBeST* (31) and used in the *Ensembl Compara* project (32).

Alternatively, van der Heijden et al. (33) demonstrated that it is often possible to infer speciation and duplication events on a gene tree without knowledge of the species tree. Their approach, which they call *species-overlap*, identifies for a given split the species represented in the two subtrees induced by the split. If at least one species has genes in both subtrees, a duplication event is inferred; else, a speciation event is inferred. In fact, this approach is a special case of soft polytomies, where all internal nodes have been collapsed.

Table 2
Overview of gene/species tree reconciliation methods and their main properties

Method	Species tree[a]	Rooting[b]	Gene tree uncertainty[c]	Framework[d]	Available Algo/DB	Reference
SDI	Fully resolved	n.a.	None	MP	X/–	(30)
RIO	Fully resolved	min dupl	Bootstrap	MP	–/X[5]	(37)
OrthoStrapper	Fully resolved	min dupl	Bootstrap	MP	X/–	(39)
GSR	Fully resolved	n.a.	n.a.	Probabilistic	X/–	(54, 57)
HOGENOM	Partially resolved	Min dupl	Multifurcate	MP	X/X	(50, 79)
Softparsmap	Partially resolved	Min dupl + min loss	Multifurcate	MP	X/–	(38)
Ensembl/ TreeBeST	Partially resolved	Min dupl + min loss	None	MP	–/X	(31, 32)
LOFT	Species overlap	Min dupl	None	MP	X/–	(33)
PhylomeDB	Species overlap	Outgroup	None	MP	–/X	(34)
BranchClust	Species overlap	Min number of clusters	None	n.a.	–/X	(35)

[a]Required species tree: Fully resolved, multifurcations allowed, computed from species overlap
[b]Approach to root gene tree (n.a. indicates that the initial rooting is assumed to be correct)
[c]Approach taken to handle reconstruction uncertainties of the gene tree (bootstrap: reconcile every bootstrap sample; multifurcate: splits in the gene tree with low support are collapsed)
[d]Used optimization framework (MP, maximum parsimony)
[e]No longer maintained

Thus, the only information needed for this approach is a rooted gene tree. Since then, this approach has been adopted in other projects, such as PhylomeDB (34). A different, but conceptually related, idea was proposed by Poptsova and Gogarten (35): their *BranchClust* method delineates COGs-like clusters in gene trees by identifying subtrees consisting of sequences represented in most species.

2.2.2. Rooting

The classical reconciliation formulation requires both gene and species trees to be rooted. But most models of sequence evolution are time reversible and thus do not allow to infer the rooting of the reconstructed gene tree. Hallett and Lagergren (36) proposed to root a gene tree so that it minimizes the number of duplication events. Thus, this method uses the parsimony principle for both rooting and reconciliation. For cases of multiple optimal rootings, Zmasek and Eddy (37) suggested in the software package *RIO* to break ties by selecting the tree that minimizes the tree height. As an alternative, Berglund-Sonnhammer et al. (38) suggested to use the rooting which minimizes the number of gene losses.

Another approach, proposed by Storm and Sonnhammer (39) and implemented in *Orthostrapper*, is to place the root at the "center of the tree." The idea of this method goes back to Farris (40) and is motivated by the concept of a molecular clock. But for most gene families, assuming a constant rate of evolution is inappropriate (41, 42), and thus this approach is not used widely.

For species tree, the most common and reliable way of rooting trees is by identifying an outgroup species. Recently, Huerta-Cepas et al. (34) have used genes from outgroup species to root gene trees. One main potential problem with this approach is that in many situations it can be difficult to identify a suitable outgroup. For example, in analysis covering all kingdoms of life, an outgroup species may not be available or the relevant genes might have been lost (43). A suitable outgroup needs to be close enough to allow for reliable sequence alignment, yet it must have speciated clearly before any other species separated. Furthermore, ancient duplications can cause outgroup species to carry *ingroup* genes. These difficulties make this approach more challenging for automated large-scale analysis (44).

2.2.3. Gene Tree Uncertainty

Another assumption made in the original tree reconciliation problem is the (topological) correctness of the gene tree. But it has been shown that this assumption is commonly violated often due to finite sequence lengths, taxon sampling (45, 46), or gene evolution model violations (47). On the other hand, techniques of expressing uncertainties in gene tree reconstruction via support measures, e.g., bootstrap values, have become well established. Storm and Sonnhammer (39) and Zmasek and Eddy (37) independently suggested to extend the bootstrap procedure to reconciliation, thereby reducing the dependency of the reconciliation procedure on any one gene tree while providing a measure of support of the inferred speciation/duplication events. The downsides of using the bootstrap are the high computational costs and interpretation difficulties associated with it (see, e.g., 48, 49, for discussions).

Similarly to how unresolved species tree can be handled, unresolved parts of the gene tree can also be collapsed into multifurcating nodes. For instance, Dufayard et al. (50) (*HOGENOM*) and Berglund-Sonnhammer et al. (38) (*Softparsmap*) collapse branches with low bootstrap support values.

A third way of tackling this problem consists in simultaneously solving both the gene tree reconstruction and reconciliation problems (51). They use the parsimony criterion of minimizing the number of duplication events to improve on the gene tree itself. This is achieved by rearranging the local gene tree topology of regions with low bootstrap support such that the number of duplications and losses is further reduced.

2.2.4. Parsimony Versus Likelihood

All the approaches mentioned so far try to minimize the number of gene duplication events. This is generally justified by a parsimony argument, which assumes that gene duplications and losses are rare events. But what if this assumption is frequently violated? Little is known about duplication and loss rates in general (52), but there is strong evidence for historical periods with high gene duplication occurrence rates (53) or gene families specifically prone to massive duplications (e.g., olfactory receptor, opsins, serine/threonine kinases, etc.).

Motivated by this reasoning, Arvestad et al. (54) introduced the idea of a probabilistic model for tree reconciliation. They used a Bayesian approach to estimate the posterior probabilities of a reconciliation between a given gene and species tree using Markov Chain Monte Carlo (MCMC) techniques. Arvestad et al. (55) modeled gene duplication and loss events through a *birth–death process*. In the subsequent years, they refined their method to also model sequence evolution and substitution rates in a unified framework called Gene Sequence evolution model with iid Rates (GSR) (56, 57).

Perhaps, the biggest problem with the probabilistic approach is that it is not clear how well the assumptions of their model (the *birth–death process* with fixed parameters) relate to the true process of gene duplication and gene loss. In a recent study, Doyon et al. (58) have compared the maximum parsimony reconciliation trees from 1,278 fungi gene families to the probabilistically reconciled trees using gene birth/death rates fitted from the data. They found that in all but two cases the maximum parsimony scenario corresponds to the most probable one. This remarkably high level of consistency indicates that in terms of the accuracy of the "best" reconciliation there is little to gain from using a likelihood approach over the parsimony criterion of minimizing the number of duplication events. But how this result generalizes to other datasets has yet to be investigated.

2.3. Graph-Based Versus Tree-Based: Which Is Better?

Given the two fundamentally different paradigms in orthology inference that we reviewed in this section, one can wonder which is better. Conceptually, tree reconciliation methods have several advantages. In terms of inference, by considering all sequences jointly, it can also be expected that they can extract more information from the sequences, which should translate into higher statistical power. In terms of their output, reconciled gene trees provide the user more information than pairs or groups of orthologs. For example, the trees display the order of duplication and speciation events, as well as evolutionary distances between these events. In practice, however, these methods have the disadvantage of having much higher computational complexity than their graph-based counterparts. Furthermore, the two approaches are in practice often not that strictly separated. Tree-based methods often

start with a graph-based clustering step to identify families of homologous genes. Conversely, several hierarchical grouping algorithms also rely on species trees in their inference.

Thus, it is difficult to make general statements about the relative performance of the two classes of inference methods. Instead, we need to evaluate methods on an individual basis, based on empirical tests. As we shall see in the next section, this is an entire topic of its own.

3. Benchmarking Orthology

Assessing the quality of orthology predictions is important, but difficult. The main challenge is that the precise evolutionary history of entire genomes is largely unknown and, thus, predictions can only be validated indirectly, using surrogate measures. To be informative, such measures need to strongly correlate with orthology/paralogy. At the same time, they should be independent from the methods used in the orthology inference process. (To be precise, inferred orthology/paralogy and the surrogate measure should be *conditionally independent* with respect to true orthology/paralogy.) Concretely, this means that the orthology inference is not based on the surrogate measure, and the surrogate measure is not derived from orthology/paralogy.

The first surrogate measures proposed revolved around conservation of function (59). This was motivated by the common belief that orthologs tend to have conserved function while paralogs tend to have different functions. Thus, Hulsen et al. (59) assessed the quality of ortholog predictions in terms of conservation of co-expression levels, domain annotation, and protein–protein interaction (PPI) partners. In addition, they also proposed using conservation of gene neighborhood as surrogate measure: the fraction of orthologs that have neighboring genes themselves orthologs is an indicator of consistency, and therefore to some extent also of quality of orthology predictions. The main limitation of these measures is that it is not so clear how much they correlate with orthology/paralogy. Indeed, it has been argued that the difference in function conservation trends between orthologs and paralogs might be much smaller than commonly assumed, and indeed many examples are known of orthologs that have dramatically different functions (60). Similarly, gene neighborhood can be conserved among paralogs, such as those resulting from whole-genome duplications. Furthermore, some methods use gene neighborhood conservation to help in their inference process, which can bias the assessment done on such measures (principle of independence stated above).

The quality of ortholog predictions can also be assessed based on phylogeny. By definition, the tree relating a set of genes all orthologous to one another only contains speciation splits, and has the same topology as the underlying species. We introduced a benchmarking protocol that quantify how well the predictions from various orthology inference methods agree with undisputed species tree topologies (61). The advantage of this measure is that by virtue of directly ensuing from the definition of orthology it correlates strongly with it, and thus satisfies the first principle. However, the second principle, independence from the inference process, is not satisfied with methods relying on the species tree—typically, all reconciliation methods, but also most graph-based methods producing hierarchical groups. In such cases, interpretation of the results must be done carefully.

For inference methods based on reconciliation between gene and species trees, Vilella et al. (32) proposed a different phylogeny-based assessment scheme. For any duplication node of the labeled gene tree, a consistency score is computed, which captures the balance of the species found in the two subtrees. Unbalanced nodes correspond to an evolutionary scenario involving extensive gene losses and therefore, under the principle of parsimony, are less likely to be correct. Given that studies to date tend to support the adequacy of the parsimony criterion in the context of gene family dynamics (Subheading 2.2.4), it can be expected that this metric correlates highly with correct orthology/paralogy assignments. However, since virtually all tree-based methods themselves incorporate this very criterion in their objective function (i.e., minimizing the number of gene duplications and losses), the principle of independence is violated, and thus the adequacy of this measure is questionable.

Finally, Chen et al. (62) proposed a purely statistical benchmark based on latent class analysis (LCA). Given the absence of definitive answer on whether two given genes are orthologs, the authors argue that by looking at the agreement and disagreement of predictions made by several inference methods on a common dataset one can estimate the reliability of individual predictors. More precisely, LCA is a statistical technique that computes maximum likelihood estimates of sensitivity and specificity rates for each orthology inference methods, given their predictions and an error model. This is attractive because it does not depend on any surrogate measure. However, the results depend on the error model assumed. Thus, we are of the opinion that LCA merely shifts the problem of assessing orthology to the problem of assessing an error model of various orthology inference methods.

Overall, it becomes apparent that there is no “magic bullet” strategy for orthology benchmarking, as each approach discussed

here has its limitations (though some limitations are more serious than others). Nevertheless, comparative studies based on these various benchmarking measures have reported surprisingly consistent findings (20, 59, 61, 62): these assessment generally observe that there is a trade-off between accuracy and coverage, and most common databases are situated on a Pareto frontier. The various assessments concur that the "best" orthology approach is highly dependent on the various possible applications of orthology.

4. Applications

As we have seen so far, there is a large diversity in the methods for orthology inference. The main reason is that, although the methods discussed here all infer orthology as part of their process, many of them have been developed for different reasons and have different ultimate goals. Unfortunately, this is often not mentioned explicitly, and tend to be a source of confusion. In this section, we review some of these ultimate goals, and discuss which methods and representation of orthology are better suited to address them and why.

As mentioned in the introduction, most interest for orthology is in the context of function prediction, and is largely based on the belief that orthologs tend to have conserved function. A conservative approach consists in propagating function between one-to-one orthologs, i.e., pairs of orthologous genes that have not undergone gene duplication since they diverged from one another. Several orthology databases directly provide one-to-one orthology predictions. But even with those that do not, it might still be possible to obtain such predictions, for instance by selecting hierarchical groups containing at most one sequence in each species or extracting from reconciled trees subtrees with no duplication. A more sophisticated approach consists in propagating gene function annotations across genomes on the basis of the full reconciled gene tree. Thomas et al. (63), for instance, proposed a way to assign gene function to uncharacterized proteins using a gene tree and a Hidden Markov Model (HMM) among gene families. Engelhardt et al. (64) developed a Bayesian model of function change along reconciled gene trees, and showed that their approach significantly improves upon several methods based on pairwise gene function propagation. Ensembl Compara (32) or Panther (63) are two major databases providing reconciled gene trees.

Since Darwin, one traditional question in biology has always been how species are related to each other. As we recall in the introduction of this chapter, Fitch's original motivation for defining orthology was phylogenetic inference. Indeed, the gene tree

reconstructed from a set of genes which are all orthologous to each other should by definition be congruent to the species tree. OMA Groups (11) have this characteristic and, crucially, are constructed without the help of a species tree.

Yet another application associated with orthology are general alignments between genomes, e.g., PPI network alignments or whole-genome alignments. Finding an optimal PPI network alignment between two genomes on the basis of the network topology alone is a computationally hard problem (i.e., it is an instance of the subgraph isomorphism problem which is NP complete (65)). Orthology is often used as heuristic to constrain the mapping of the corresponding genes between the two networks, and thus to reduce the problem of complexity of aligning networks (66). For whole-genome alignments, people most often use homologous regions and use orthologs as anchor points (67). These types of applications typically rely on ortholog predictions between pairs of genomes, as provided, e.g., by Inparanoid (10) or OMA (11).

5. Conclusions and Outlook

The distinction between orthologs and paralogs is at the heart of many comparative genomics studies and applications. The original and generally accepted definition of orthology is based on the evolutionary history of pairs of genes. By contrast, there is considerable diversity in how groups of orthologs are defined. These differences largely stem from the fact that orthology is a nontransitive relation, and therefore dividing genes into orthologous groups either misses or wrongly includes orthologous relations. This makes it important and worthwhile to identify the type of orthologous group best suited for a given application.

Regarding inference methods, we observe that while most approaches can be ordered into two fundamental paradigms—graph based and tree based—the difference between the two is shrinking, with graph-based methods increasingly striving to capture more of the evolutionary history. On the other hand, the rapid pace at which new genomes are sequenced limits the applicability of tree-based methods, computationally more demanding.

Benchmarking this large variety of methods remains a hard problem—not only from a conceptual point as described above, but also because of very practical challenges, such as heterogeneous data formats, genome versions, or gene identifiers. This has been recognized by the research community and there is now a joint initiative to overcome at least these practical hurdles (68).

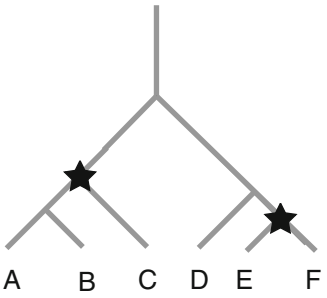
Looking forward, we see potential in extending the current model of gene evolution, which is limited to speciation,

duplication, and loss events. Indeed, nature is often much more complicated. For instance, lateral gene transfer (LGT) is believed to be a major mode of evolution in prokaryotes. While there has been several attempts at extending tree reconciliation algorithms to detecting LGT (69, 70), none of these have been adopted by orthology databases. Another relevant evolutionary process omitted by most methods are whole-genome duplications (WGDs). Even though WGD events act jointly on all gene families, with few exceptions (71, 72), most methods consider each gene family independently.

Overall, the orthology/paralogy dichotomy has proved to be useful, but also inherently limited. Reducing the whole evolutionary history of homologous genes into binary pairwise relations is bound to be a simplification—and at times an oversimplification. Thus, the trend toward capturing more features of the evolutionary history of genes can be expected to continue for a long time, as we are nowhere close to grasp the formidable complexity of nature.

6. Exercises

Assume the following evolutionary scenario



where duplications are depicted as *, and all other splits are speciations.

Problem #1: Draw the corresponding orthology graph, where the vertices correspond to the observed genes and the edges indicate orthologous relations between them.

Problem #2: Apply the following two clustering methods on your orthology graph. First, reconstruct all the maximal fully connected subgraphs (cliques) that can be found. Second, reconstruct the COGs. COGs are built by merging triangles of orthologs whenever they share a common face. Remember that in both methods a gene can only belong to one cluster.

Acknowledgments

We thank Stefan Zoller for helpful feedback on the manuscript. Part of this chapter started as assignment for the graduate course "Reviews in Computational Biology" (263-5151-00L) at ETH Zurich.

References

1. Dewey C (2012) Whole-genome alignment. In Anisimova, M., (ed.), Evolutionary genomics: statistical and computational methods (volume 1). Methods in Molecular Biology, Springer Science+Business media, LLC.
2. Alioto T (2012) Gene prediction. In Anisimova, M., (ed.), Evolutionary genomics: statistical and computational methods (volume 1). Methods in Molecular Biology, Springer Science+Business media, LLC.
3. Loytynoja A (2012) Alignment methods: strategies, challenges, benchmarking, and comparative overview. In Anisimova, M., (ed.), Evolutionary genomics: statistical and computational methods (volume 1). Methods in Molecular Biology, Springer Science+Business media, LLC.
4. Walter M Fitch. Distinguishing homologous from analogous proteins. *Syst Zool*, 19 (2):99–113, 1970.
5. Arnold Kuzniar, Roeland C H J van Ham, Sándor Pongor, and Jack A M Leunissen. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet*, 24 (11):539–51, Nov 2008.
6. Roman L. Tatusov, Eugene V. Koonin, and David J. Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–7, 1997.
7. Ross Overbeek, Michael Fonstein, Mark D. Souza, Gordon D. Pusch, and Natalia Maltsev. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U.S.A.*, 96:2896–2901, 1999.
8. Temple F. Smith and Michael S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197, 1981.
9. Altschul S F, Madden T L, Schaffer A A, Zhang J, Zhang Z, Miller W, and Lipman D J. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402, Sep 1997.
10. Remm M, Storm CE, and Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, 314(5):1041–52, 2001.
11. Christophe Dessimoz, Gina Cannarozzi, Manuel Gil, Daniel Margadant, Alexander Roth, Adrian Schneider, and Gaston Gonnet. OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: Introduction and first achievements. In Aoife McLysath and Daniel H. Huson, editors, *RECOMB 2005 Workshop on Comparative Genomics*, volume LNBI 3678 of *Lecture Notes in Bioinformatics*, pages 61–72. Springer-Verlag, 2005.
12. Fulton DL, Li YY, Laird MR, Horsman BG, Roche FM, and Brinkman FS. Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics*, 28(7): 270, 2006.
13. Wall D P, Fraser H B, and Hirsh A E. Detecting putative orthologs. *Bioinformatics*, 19(13): 1710–1711, 2003.
14. Liisa B. Koski and G. Brian Golding. The closest BLAST hit is often not the nearest neighbor. *J Mol Evol*, 52(6):540–542, 2001.
15. Alexander C Roth, Gaston H Gonnet, and Christophe Dessimoz. The algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics*, 9:518, 2008. doi: 10.1186/1471-2105-9-518.
16. Christophe Dessimoz, Brigitte Boeckmann, Alexander C J Roth, and Gaston H Gonnet. Detecting non-orthology in the cogs database and other approaches grouping orthologs using genome-specific best hits. *Nucleic Acids Res*, 34(11):3309–3316, 2006. doi: 10.1093/nar/gkl433. URL http://dx.doi.org/10.1093/nar/gkl433.
17. David M Kristensen, Lavanya Kannan, Michael K Coleman, Yuri I Wolf, Alexander Sorokin, Eugene V Koonin, and Arcady Mushegian. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics*, 26(12):1481–1487, Jun 2010. doi: 10.1093/bioinformatics/btq229. URL http://dx.doi.org/10.1093/bioinformatics/btq229.

18. Li Li, Christian J Jr Stoeckert, and David S Roos. Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome Res*, 13(9):2178–2189, Sep 2003.
19. Stijn van Dongen. *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, May 2000.
20. Brigitte Boeckmann, Marc Robinson-Rechavi, Ioannis Xenarios, and Christophe Dessimoz. Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. *Brief Bioinform*, 12(5):423–435, Sep 2011.
21. Lars Juhl Jensen, Philippe Julien, Michael Kuhn, Christian von Mering, Jean Muller, Tobias Doerks, and Peer Bork. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucl. Acids Res.*, 36 (Database issue):D250–D254, 2008. doi: 10.1093/nar/gkm796.
22. Evgenia V Kriventseva, Nazim Rahman, Octavio Espinosa, and Evgeny M Zdobnov. Orthodb: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res*, 36 (Database issue):D271–D275, Jan 2008. doi: 10.1093/nar/gkm845. URL http://dx.doi.org/10.1093/nar/gkm845.
23. Raja Jothi, Elena Zotenko, Asba Tasneem, and Teresa M Przytycka. Coco-cl: hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics*, 22 (7):779–788, Apr 2006. doi: 10.1093/bioinformatics/btl009. URL http://dx.doi.org/10.1093/bioinformatics/btl009.
24. Masatoshi Nei. *Molecular Evolutionary Genetics*. Columbia University Press, New York, 1987.
25. Morris Goodman, John Czelusniak, G W Moore, and A E Romero-Herrara. Fitting the gene lineage into its species lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst Zool*, 28 (2):132–168, 1979.
26. Roderic Page. Maps between trees and cladistic – analysis of historical associations among genes, organisms, and areas. *Syst Biol*, 43 (1):58–77, Jan 1994.
27. Mirkin B, Muchnik I, and Smith T F. A biologically consistent model for comparing molecular phylogenies. *J Comput Biol*, 2(4):493–507, Jan 1995.
28. Zhang L. On a mirkin-muchnik-smith conjecture for comparing molecular phylogenies. *J Comput Biol*, 4(2):177–87, Jul 1997.
29. Oliver Eulenstein. A linear time algorithm for tree mapping. *Arbeitspapiere der GMD No. 1046, St Augustine, Germany*, page 1046, 1997.
30. Zmasek C M and Eddy S R. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics*, 17(9):821–8, Sep 2001.
31. Heng Li, Avril Coghlan, Jue Ruan, Lachlan James Coin, Jean-Karim Hrich, Lara Osmotherly, Ruiqiang Li, Tao Liu, Zhang Zhang, Lars Bolund, Gane Ka-Shu Wong, Weimou Zheng, Paramvir Dehal, Jun Wang, and Richard Durbin. Treefam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res*, 34 (Database issue):D572–D580, Jan 2006. doi: 10.1093/nar/gkj118. URL http://dx.doi.org/10.1093/nar/gkj118.
32. Albert J J. Vilella, Jessica Severin, Abel Ureta-Vidal, Richard Durbin, Li Heng, and Ewan Birney. Ensemblcompara genetrees: Analysis of complete, duplication aware phylogenetic trees in vertebrates. *Genome research*, 19 (2):327–335, 2009. doi: http://dx.doi.org/10.1101/gr.073585.107.
33. Rene TJM van der Heijden, Berend Snel, Vera van Noort, and Martijn A Huynen. Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics*, 8 (1):83, 2007.
34. Jaime Huerta-Cepas, Hernán Dopazo, Joaquín Dopazo, and Toni Gabaldón. The human phylome. *Genome Biol*, 8(6):R109, Jan 2007. doi: 10.1186/gb-2007-8-6-r109. URL http://genomebiology.com/2007/8/6/R109.
35. Maria Poptsova and J Peter Gogarten. Branchclust: a phylogenetic algorithm for selecting gene families. *BMC Bioinformatics*, 8(1):120, 2007. doi: 10.1186/1471-2105-8-120. URL http://www.biomedcentral.com/1471-2105/8/120.
36. Hallett M and Lagergren J. New algorithms for the duplication-loss model. *RECOMB '00*: Apr 2000. URL http://portal.acm.org/citation.cfm?id=332306.332359.
37. Zmasek C M and Eddy S R. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, 3(14), May 2002. doi: 10.1186/1471-2105-3-14.
38. Ann-Charlotte Berglund-Sonnhammer, Pär Steffansson, Matthew J Betts, and David A Liberles. Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. *J Mol Evol*, 63(2):240–50, Aug 2006. doi: 10.1007/s00239-005-0096-1.
39. CE Storm and EL Sonnhammer. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, 18(1):92–9, Jan 2002.

40. James S. Farris. Estimating phylogenetic trees from distance matrices. *The American Naturalist*, 106(951):645–668, 1972. ISSN 00030147. URL http://www.jstor.org/stable/2459725.
41. Avise J C, Bowen B W, Lamb T, Meylan A B, and Bermingham E. Mitochondrial dna evolution at a turtle's pace: evidence for low genetic variability and reduced microevolutionary rate in the testudines. *Mol Biol Evol*, 9(3):457–473, May 1992.
42. Ayala F J. Molecular clock mirages. *Bioessays*, 21(1):71–75, Jan 1999. URL http://dx.doi.org/3.0.C0;2-B.
43. John P Huelsenbeck, Jonathan P Bollback, and Amy M Levine. Inferring the root of a phylogenetic tree. *Syst Biol*, 51(1):32–43, Feb 2002. doi: 10.1080/106351502753475862. URL http://dx.doi.org/10.1080/106351502753475862.
44. R. Tarrío, F. Rodríguez-Trelles, and F. J. Ayala. Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: the drosophila saltans and willistoni groups, a case study. *Mol Phylogenet Evol*, 16 (3):344–349, Sep 2000. doi: 10.1006/mpev.2000.0813. URL http://dx.doi.org/10.1006/mpev.2000.0813.
45. Anna Graybeal. Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst Biol*, 47:9–17, 1998.
46. Antonis Rokas, Barry L Williams, Nicole King, and Sean B Carroll. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425(6960):798–804, Oct 2003. doi: 10.1038/nature02053. URL http://dx.doi.org/10.1038/nature02053.
47. Z. Yang, N. Goldman, and A. Friday. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol Biol Evol*, 11(2):316–324, Mar 1994.
48. Holmes. *Statistics in Genetics*, chapter Phylogenies: An Overview, pages 81–118. Springer, NY, 1999.
49. Anisimova M and Gascuel O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol*, 55 (4):539–52, 2006.
50. Jean-François Dufayard, Laurent Duret, Simon Penel, Manolo Gouy, François Rechenmann, and Guy Perriere. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, 21(11):2596–603, Jun 2005. doi: 10.1093/bioinformatics/bti325. URL http://bioinformatics.oxfordjournals.org/cgi/content/full/21/11/2596.
51. Dannie Durand, Bjarni V Halldórsson, and Benjamin Vernot. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol*, 13(2):320–35, Mar 2006. doi: 10.1089/cmb.2006.13.320.
52. Lynch M and Conery J S. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155, Nov 2000. Comment.
53. Robinson-Rechavi M, Marchand O, Escriva H, Bardet P L, Zelus D, Hughes S, and Laudet V. Euteleost fish genomes are characterized by expansion of gene families. *Genome Res*, 11 (5):781–788, May 2001. doi: 10.1101/gr.165601. URL http://dx.doi.org/10.1101/gr.165601.
54. Lars Arvestad, Ann-Charlotte Berglund, Jens Lagergren, and Bengt Sennblad. Bayesian gene/species tree reconciliation and orthology analysisusing mcmc. *Bioinformatics*, 19(suppl 1):i7–15, 2003. doi: 10.1093/bioinformatics/btg1000.
55. David G. Kendall. On the generalized "birth-and-death" process. *Ann of Math Stat*, 19 (1):1–15, 1948. ISSN 00034851. URL http://www.jstor.org/stable/2236051.
56. Lars Arvestad, Ann-Charlotte Berglund, Jens Lagergren, and Bengt Sennblad. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. *RECOMB '04*. URL http://portal.acm.org/citation.cfm?id=974614.974657.
57. Orjan Åkerborg, Bengt Sennblad, Lars Arvestad, and Jens Lagergren. Simultaneous bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci USA*, 106(14):5714–9, Apr 2009. doi: 10.1073/pnas.0806251106.
58. Jean-Philippe Doyon, Vincent Ranwez, Vincent Daubin and Vincent Berry. Models, algorithms and programs for Phylogeny reconciliation. *Brief Bioinform*, 12(5):392–400, Sep 2011. doi: 10.1093/bib/bbr045. URL http://dx.doi.org/10.1093/bib/bbr045.
59. Tim Hulsen, Martijn A Huynen, Jacob de Vlieg, and Peter MA Groenen. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol*, 7 (4):R31, April 2006. doi: 10.1186/gb-2006-7-4-r31.
60. Romain A Studer and Marc Robinson-Rechavi. How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet*, 25 (5):210–216, May 2009. doi: 10.1016/j.tig.2009.03.004. URL http://dx.doi.org/10.1016/j.tig.2009.03.004.

61. Adrian M. Altenhoff and Christophe Dessimoz. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol*, 5(1):e1000262, 2009. doi: 10.1371/journal.pcbi.1000262.
62. Chen F, Mackey A J, Vermunt J K, and Roos D S. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, 2(4):e383, 2007. doi: 10.1371/journal.pone.0000383.
63. Paul D Thomas, Michael J Campbell, Anish Kejariwal, Huaiyu Mi, Brian Karlak, Robin Daverman, Karen Diemer, Anushya Muruganujan, and Apurva Narechania. Panther: a library of protein families and subfamilies indexed by function. *Genome Res*, 13(9):2129–2141, Sep 2003. doi: 10.1101/gr.772403. URL http://dx.doi.org/10.1101/gr.772403.
64. Barbara E Engelhardt, Michael I Jordan, Kathryn E Muratore, and Steven E Brenner. Protein molecular function prediction by bayesian phylogenomics. *PLOS Comp Biol*, 1 (5):432–445, 2005.
65. Stephen A. Cook. The complexity of theorem-proving procedures. In *STOC '71: Proceedings of the third annual ACM symposium on Theory of computing*, pages 151–158, New York, NY, USA, 1971. ACM. doi: http://doi.acm.org/10.1145/800157.805047.
66. Roded Sharan and Trey Ideker. Modeling cellular machinery through biological network comparison. *Nat Biotechnol*, 24(4):427–433, Apr 2006. doi: 10.1038/nbt1196. URL http://dx.doi.org/10.1038/nbt1196.
67. Colin N Dewey and Lior Pachter. Evolution at the nucleotide level: the problem of multiple whole-genome alignment. *Hum Mol Genet*, 15 Spec No 1:R51–R56, Apr 2006. doi: 10.1093/hmg/ddl056. URL http://dx.doi.org/10.1093/hmg/ddl056.
68. Toni Gabaldón, Christophe Dessimoz, Julie Huxley-Jones, Albert J Vilella, Erik Ll Sonnhammer, and Suzanna Lewis. Joining forces in the quest for orthologs. *Genome Biol*, 10 (9):403, 2009. doi: 10.1186/gb-2009-10-9-403. URL http://dx.doi.org/10.1186/gb-2009-10-9-403.
69. Pawel Górecki. Reconciliation problems for duplication, loss and horizontal gene transfer. *RECOMB '04*. URL http://portal.acm.org/citation.cfm?id=974614.974656.
70. Mike Hallett, Jens Lagergren, and Ali Tofigh. Simultaneous identification of duplications and lateral transfers. *RECOMB '04*. URL http://portal.acm.org/citation.cfm?id=974614.974660.
71. Guigó R, Muchnik I, and Smith T F. Reconstruction of ancient molecular phylogeny. *Mol Phylogen Evol*, 6(2):189–213, Oct 1996. doi: 10.1006/mpev.1996.0071.
72. Mukul S Bansal and Oliver Eulenstein. The multiple gene duplication problem revisited. *Bioinformatics*, 24(13):i132–8, Jul 2008. doi: 10.1093/bioinformatics/btn150.
73. Gabriel Ostlund, Thomas Schmitt, Kristoffer Forslund, Tina Köstler, David N Messina, Sanjit Roopra, Oliver Frings, and Erik L L Sonnhammer. Inparanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res*, 38(Database issue):D196–D203, Jan 2010. doi: 10.1093/nar/gkp931. URL http://dx.doi.org/10.1093/nar/gkp931.
74. Todd F. DeLuca, I-Hisen Wu, Jian Pu, Thomas Monaghan, Leonid Peshkin, Saurav Singh, and Dennis P. Wall. Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics*, 22(16):2044–2046, Jun 2006.
75. Adrian M Altenhoff, Adrian Schneider, Gaston H Gonnet, and Christophe Dessimoz. OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res*, 39(Database issue):D289–D294, Jan 2011. doi: 10.1093/nar/gkq1238. URL http://dx.doi.org/10.1093/nar/gkq1238.
76. Feng Chen, Aaron J Mackey, Christian J Stoeckert, and David S Roos. Orthomcldb: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res*, 34 (Database issue):D363–D368, Jan 2006. doi: 10.1093/nar/gkj123. URL http://dx.doi.org/10.1093/nar/gkj123.
77. Muller J, Szklarczyk D, Julien P, Letunic I, Roth A, Kuhn M, Powell S, von Mering C, Doerks T, Jensen L J, and Bork P. eggnog v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res*, 38(Database issue):D190–D195, Jan 2010. doi: 10.1093/nar/gkp951. URL http://dx.doi.org/10.1093/nar/gkp951.
78. Benjamin Linard, Julie Thompson, Olivier Poch, and Odile Lecompte. Orthoinspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics*, 12(1):11, 2011. doi: 10.1186/1471-2105-12-11. URL http://www.biomedcentral.com/1471–2105/12/11.
79. Simon Penel, Anne-Muriel Arigon, Jean-Franois Dufayard, Anne-Sophie Sertier, Vincent Daubin, Laurent Duret, Manolo Gouy, and Guy Perrire. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, 10 Suppl 6:S3, 2009. doi: 10.1186/1471-2105-10-S6-S3. URL http://dx.doi.org/10.1186/1471-2105-10-S6-S3.