BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i549–i558
doi:10.1093/bioinformatics/btm193

Automatic genome-wide reconstruction of phylogenetic gene trees

Ilan Wapinski[1,2,3], Avi Pfeffer[2], Nir Friedman[5] and Aviv Regev[3,4,*]

[1]Broad Institute of MIT and Harvard, [2]School of Engineering and Applied Sciences, Harvard University, [3]FAS Center for Systems Biology, [4]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA and [5]School of Computer Science & Engineering, Hebrew University, Jerusalem, Israel

ABSTRACT

Gene duplication and divergence is a major evolutionary force. Despite the growing number of fully sequenced genomes, methods for investigating these events on a genome-wide scale are still in their infancy. Here, we present SYNERGY, a novel and scalable algorithm that uses sequence similarity and a given species phylogeny to reconstruct the underlying evolutionary history of all genes in a large group of species. In doing so, SYNERGY resolves homology relations and accurately distinguishes orthologs from paralogs. We applied our approach to a set of nine fully sequenced fungal genomes spanning 150 million years, generating a genome-wide catalog of orthologous groups and corresponding gene trees. Our results are highly accurate when compared to a manually curated gold standard, and are robust to the quality of input according to a novel jackknife confidence scoring. The reconstructed gene trees provide a comprehensive view of gene evolution on a genomic scale. Our approach can be applied to any set of sequenced eukaryotic species with a known phylogeny, and opens the way to systematic studies of the evolution of individual genes, molecular systems and whole genomes.

Contact: aregev@broad.mit.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

With the rapid growth of whole-genome sequencing, comparative genomics is increasingly employed in evolutionary and functional studies of biological systems (Cliften *et al.*, 2003; Kellis *et al.*, 2003). Such studies require that we first reconstruct the evolutionary history of individual genes, and their relation to one another through speciation (orthologs) or duplication (paralogs) events. The concepts of gene orthology and paralogy (Fitch, 1970) have been mostly employed to study the evolution of individual gene families. Recently, these concepts have been applied on a genome-wide scale to functionally characterize and classify genes (Tatusov *et al.*, 1997), and to understand the evolutionary impact of genomic events (Blomme *et al.*, 2006; Dietrich *et al.*, 2004; Kellis *et al.*, 2004; Scannell *et al.*, 2006). Genome-scale mapping of orthologs and paralogs is also the first step when studying the evolution of proteins with shared ancestry, interactions and regulation.

Significant efforts have been invested in developing methods to identify orthologous and paralogous genes. Most can be divided into two broad categories. The first class of methods infer homology relations based on *hit-clustering*, using the results (‘hits’) from sequence similarity searches between all the proteins in different species to output an orthology assignment between the genes. The most widely used variant of this approach is ‘reciprocal (bi-directional) best hits’ (RBH) where two genes in two different species are identified as orthologs if each is the others’ best ‘hit’ in that species (Fitch, 1970; Wall *et al.*, 2003). Related approaches include more inclusive clustering methods (e.g. COGs, Tatusov *et al.*, 1997) and algorithms that distinguish between recent and ancient gene duplications (e.g. INPARANOID, Remm *et al.*, 2001; OrthoMCL, Li *et al.*, 2003). Recent extensions have incorporated information on orthologous chromosomal regions (synteny) to guide orthology assignments (e.g. BUS, Kellis *et al.*, 2004). Synteny-based methods are particularly helpful in handling orthology assignments that are ambiguous based on hit-clustering alone, but they cannot be applied between distantly related species, where gene order is not sufficiently conserved. Hit clustering methods are easy to implement and fast, but they do not explicitly reconstruct the evolutionary history of orthologous genes, as they either ignore paralogs altogether (e.g. RBH) or do not resolve exact orthology and paralogy relations when identifying genes with shared ancestry.

A complementary set of approaches identifies homology relations in light of the *phylogenetic gene tree* of a related group of genes. These allow us to infer lineage-specific duplications and losses by comparison to the corresponding species tree (Goodman *et al.*, 1979; Zmasek and Eddy, 2001); see Figure 1. Recent methods attempt to balance the number of inferred duplications and losses with evidence derived from sequence alignments (Arvestad *et al.*, 2003; Durand *et al.*, 2006). While such approaches result in high-quality reconstruction of gene histories, they are computationally intensive and have therefore been typically restricted to pre-defined families of genes rather than applied on a genomic scale.

Recent efforts to to apply phylogenetic methods towards large scale resolution of orthologies (Goodstadt and Ponting, 2006; Jothi *et al.*, 2006), address the task in a sequential way: first, they use hit-clustering methods to identify coarse gene families and then construct gene trees to refine these assignments. The latter phylogenetic step does not employ the more sophisticated but computationally intensive phylogenetic

*To whom correspondence should be addressed.

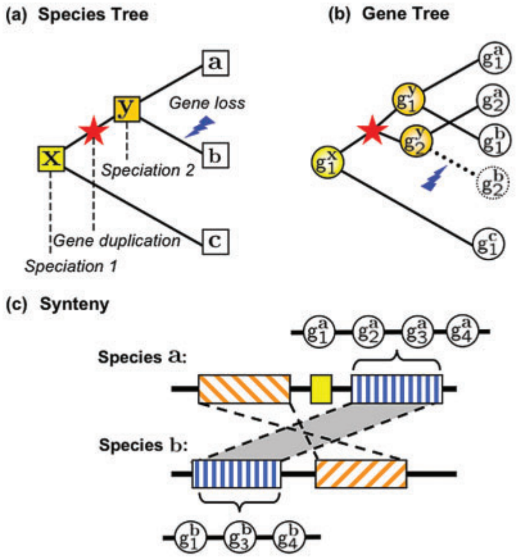


Fig. 1. Homology subtypes—orthology and paralogy—within a group of orthologous genes. (**a**) Species tree. Each node (square) in the tree is a species—either extant (leaf node) or ancestral (internal node). In this toy example, speciation events 1 and 2 have resulted in extant species a, b, and c. (**b**) A gene tree describing the evolutionary events for the genes g_1^a,g_2^a,g_1^b, and g_1^c. Each node in the tree is a gene (circle) or a duplication event (star). The tree shows the evolutionary descent of the ancestral gene g_1^x to paralogs and orthologs following gene duplication in species **y**, and the subsequent speciation yielding species **a** and **b**. Gene g_2^b was lost (blue strike and dashed lines) after the duplication event, but its paralog, g_1^b , was retained. (**c**) Synteny between chromosomal regions in species **a** and **b**. Each chromosome has several similar (syntenic) blocks (hatched boxes) comprised of multiple genes. Some regions in one genome (yellow box) do not have a syntenic counterpart in the other. The synteny similarity score for a pair of genes is the fraction of their neighbors that are orthologous to each other. For example, the score for g_3^a and g_3^b is 2/3.

algorithms. Thus, it does not account for gene tree distortions that induce large numbers of unlikely duplication and loss events (Blomme *et al*., 2006; Dufayard *et al*., 2005; Jothi *et al*., 2006). Such distortions are common as genes within families often evolve at uneven rates, especially following gene duplication events (Kellis *et al*., 2004; Lynch and Katju, 2004). Consequently, laborious manual curation by experts may be required to achieve reasonable results (Li *et al*., 2006) or more complicated families must be ignored a priori (Blomme *et al*.,2006).

Here we present a novel framework for the genome-wide reconstruction of homology relations across multiple eukaryotic genomes and describe a fully automatic and scalable implementation of this framework in the SYNERGY algorithm. Given a set of genomes and the known species phylogeny, our algorithm resolves the orthology and paralogy relations for all the protein coding genes in those genomes, while *simultaneously* reconstructing the phylogenetic gene trees for each group of orthologs. Our approach combines the scalability and automation of hit clustering approaches with the detailed phylogenetic reconstruction of tree-based methods, resulting in a robust resolution of homology relations. Since SYNERGY reconstructs gene trees simultaneously as it identifies orthologous groups, it avoids many of the pitfalls of sequential methods. Our approach is flexible and can incorporate additional types of data whenever available (e.g. synteny). To automatically assess the quality of our assignments, we also develop a jackknife-based method for measuring their robustness to perturbations in the included genes and species.

We applied our method to published fungal genomes (Cliften *et al*., 2003; Dietrich *et al*., 2004; Dujon *et al*., 2004; Kellis *et al*., 2003, 2004), whose phylogeny spans 150 million years, including a whole genome duplication (WGD). We found 5282 (non-singleton) orthology groups that cover 48 265 (92%) of the 52 697 protein-coding genes predicted within these species. Our results markedly improve over the widely used RBH approach, are of comparable quality to a manually curated gold standard (Byrne and Wolfe, 2005), are highly robust to perturbations in input data, and correctly assign more orthologs than previous methods (Remm *et al*., 2001). The reconstructed gene trees provide a detailed history for each group of orthologous genes, pinpointing duplication, loss and divergence events (and resolving orthologies and paralogies) at an unprecendented high resolution for an automatic genomic method, nearing the level of manual expertise. Thus, our algorithm opens the way to a host of comprehensive comparative genomics studies in any group of species with a known phylogeny.

2 METHODS

Given a set of species, their protein-coding genes and their phylogenetic tree, SYNERGY partitions the genes into disjoint subsets, where each subset contains all and only those genes that descended from a single gene in the species' last common ancestor. SYNERGY simultaneously reconstructs the phylogenetic gene tree for each such subset of genes. Briefly, SYNERGY performs this task in a step-wise bottom-up fashion, solving it sequentially for each ancestral node in a species tree from the leaves of the tree to the root. At each stage (i.e. node in the species tree), SYNERGY first clusters together the genes or groups of orthologs from previous stages that share significant sequence similarity. It then reconstructs a phylogenetic gene tree for each of these intermediate groups of orthologs, and uses this tree to partition the clusters such that each contains only genes that are descended from a single hypothetical gene in the ancestral species corresponding to the current stage. Thus, after each stage, SYNERGY has made a complete orthology assignment and gene tree reconstruction for the complement of genes below the corresponding node in the species tree. These are then passed up to the next stage. Once SYNERGY reaches the root of the species tree, a full partition of groups of orthologs that are descended from a single ancestral gene has been made along with a corresponding gene tree for each such group.

2.1 Defining orthogroups

There are two major classes of homology relations between genes (Fitch, 1970). Orthologs are genes that share a common ancestor at a speciation event, while paralogs are related through duplication events (Fig. 1a and b). These are not necessarily simple one-to-one relationships. For example, two paralogous genes that resulted from a duplication after a speciation event, are both orthologous to the same gene in another species (Fig. 1). Conversely, when genes are lost in a particular species or lineage, orthology may be a one-or many-to-none relationship.

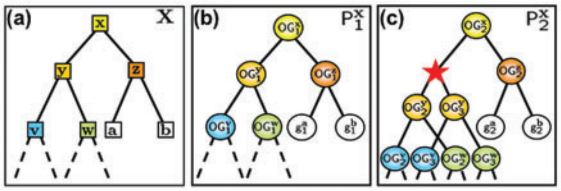


Fig. 2. Orthogroups and their phylogenetic gene trees. (**a**) A species tree X rooted at the ancestral species **x**. Only a fraction of the tree is shown. (**b**, **c**) A gene tree (e.g. $P_1^{\mathbf{x}}$) represents the evolutionary history of all the genes that descended from the gene $g_1^{\mathbf{x}}$ in species **x**. Each internal node in the gene tree defines a corresponding orthogroup, (e.g. $OG_1^{\mathbf{x}}$, $OG_1^{\mathbf{y}}, \ldots$), whose members are the genes below that node in the tree. The gene tree can track duplication events (star, panel c, $OG_2^{\mathbf{y}}$ and $OG_3^{\mathbf{y}}$).

Such relations are captured by phylogenetic trees (Fig. 2a–c). We denote a species tree **T** where internal nodes (**x**, **y**,...) represent ancestral species, and leaf nodes (**a**, **b**,...) represent extant species (Fig. 2a). We denote as g^a a gene g in species a. The exact orthology and paralogy relations between genes are represented in a gene tree P (Fig. 1b).[1] The leaves in P are the genes which descended from a single common ancestor gene at the root of P. Its internal nodes represent the speciation and duplication events that occurred within the course of the genes' evolution (Fig. 1b).

We define an *Orthogroup* as the set of genes that descended from a single common ancestral gene. An orthogroup $OG_i^{\mathbf{x}}$ is defined with respect to an ancestral species **x** in **T** and includes only and all of those genes from the extant species under **x** that descended from a single common ancestral gene $g_i^{\mathbf{x}}$ in **x**.We therefore define:

Definition 1. An orthogroup $OG_i^{\mathbf{x}}$ under the ancestral species $x \in T$ is *sound* if there existed a gene $g_i^{\mathbf{x}}$ in **x** such that for every gene g_i^a in OG_i^x, g_j^a , is a descendant of $g_i^{\mathbf{x}}$.

Definition 2. $OG_i^{\mathbf{x}}$ is *complete* if every gene g_j^a that descended from the ancestral gene $g_i^{\mathbf{x}}$ is in $OG_i^{\mathbf{x}}$.

Each orthogroup $OG_i^{\mathbf{x}}$ has a corresponding gene tree P_i^x. The leaves in P_i^x are the genes $g_j^a \in OG_i^x$ (for every extant species **a** at the leaves of T under **x**), and its internal nodes denote ancestral genes and the duplication events that occurred along $OG_i^{\mathbf{x}}$'s evolution (Fig. 1b). The root of $P_i^{\mathbf{x}}$ represents the ancestral gene $g_i^{\mathbf{x}}$, the last common ancestor of all $g_j^a \in OG_i^x$.

Since an orthogroup $OG_i^{\mathbf{x}}$ represents the ancestral gene $g_i^{\mathbf{x}}$ we will subsequently refer to orthogroups and genes interchangeably.

2.2 Scoring gene similarity

The common ancestry of homologuous proteins implies that they retain some similarity. The estimate of the evolutionary distance between pairs of proteins is the basis for our reconstruction method. Although our method can be applied with any method for computing these pairwise distances, much of the success depends on these choices. Here we use a measure of distance that examines the evolution of both the amino acid sequence of the proteins and the chromosomal organization of genomes.

When comparing amino acid sequences, we use standard models of amino acid evolution. Specifically, our *peptide sequence similarity score* (d^p) between a pair of proteins is based the JTT amino acid substitution rates (Jones *et al*., 1992). To compute d^p we first globally align two proteins, then search for the distance that maximizes the likelihood of substitutions in each aligned position.

To capture the information genome organization conveys about the homology between proteins, our *synteny similarity score* (d^s) quantifies the similarity between the chromosomal neighborhoods of two genes (Fig. 1c). A (preliminary) orthology assignment anchors chromosomal regions in two species to one another. Genes that are highly syntenic to each other will share many such anchors between their chromosomal neighborhoods. Since there is currently no agreed-upon evolutionary model of genome organization, we compute the synteny similarity score between two genes as the fraction of their neighbors that are orthologous to one another (Fig. 1c). The source of the preliminary orthology assignment will be discussed below.

Both d^p and d^s are scaled and treated as distances for assessing protein and chromosomal evolution between pairs of genes. Two genes with high similarity have scores close to 0, while genes sharing no similarity have scores of 2.0. We combine these two measures to identify potentially orthologous genes (Section 2.4).[2]

2.3 Gene similarity graph

SYNERGY relies on the pre-computed distances between genes to make orthology assignments. We could compute the distances between each pair of genes in all the input genomes, but most of these distances will be maximal, as most genes do not share a common ancestor. Instead, we construct a sparse data structure that maintains the relevant distances. This sparse representation also helps guide the algorithm by identifying the potential homologies among the input genes.

These relations are represented by a *gene similarity graph* as a weighted directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ are all the individual genes in the input genomes, and the edges $\mathcal{E}$ represent potential homology relations. To generate $\mathcal{E}$, we first execute all-versus-all FASTA alignments between all genes in our input (Pearson and Lipman, 1988). We next designate gene pairs that are significantly similar, placing an edge between g_i^a in species **a** and g_i^b in species **b** if the FASTA *E*-value of their alignment is below 0.1 and either g_i^b is the best FASTA hit in species **b** to g_i^a or the percent identity between g_i^a and g_i^b is above 50% of that between g_i^a and its best hit in **b**. We weigh each edge by the distance scores defined above.[3] While this distance is symmetric, the edges are directed, and are placed from the query to the target gene based on the direction of the similarity search.

2.4 Identifying orthogroups

Identifying orthogroups across multiple species amounts to sequentially reconstructing the shared ancestral relationships between genes at each internal position of a phylogenetic tree. To this end, SYNERGY (Fig. 3) recursively traverses the nodes of the given species tree **T** from its leaves to its root, identifying orthogroups with respect to each node. At each recursive Stage, SYNERGY assumes that sound and complete orthogroups and their corresponding gene trees are resolved for the lower nodes in the tree. For each internal node $\mathbf{x} \in \mathbf{T}$, SYNERGY uses the distances between genes (or, equivalently, between orthogroups resolved in previous stages) to determine the orthogroups $\{OG^x\}$ and reconstruct the phylogenetic gene trees $\{P^x\}$ between the member genes of each orthogroup. Once this is completed, the set of newly identified orthogroups and their corresponding gene trees are recorded. At this

[1] We assume that gene fusion and horizontal transfer events are rare and that therefore genes are descended from single genes, allowing us to represent gene phylogenies as trees.

[2] The protein similarity score scales with evolutionary distance. We scale the synteny score to the same range but we do not make any assumptions about its direct evolutionary interpretation.

[3] We rely more heavily on the protein similarity described in 2.2 than on bitscores or E-values because the best 'hit' is often not the nearest phylogenetic neighbor (Koski and Golding, 2001; Wall *et al*., 2003).

```
SYNERGY Algorithm
Input: A species tree node $\mathbf{x}$
Output: A set of orthogroups $\{OG^{\mathbf{x}}\}$

if $\mathbf{x}$ is an extant species
    $\{OG^{\mathbf{x}}\} \leftarrow \{g^{\mathbf{x}}\}$
else
    // Call SYNERGY recursively
    $\{OG^{r}\} \leftarrow$ SYNERGY($\mathbf{x}$.right)
    $\{OG^{l}\} \leftarrow$ SYNERGY($\mathbf{x}$.left)
    // step 1: match orthogroups; make putative orthogroups $\{OG^{\mathbf{x}}\}$
    $\{OG^{\mathbf{x}}\} \leftarrow$ MatchOrthogroups($\mathbf{x}$, $\{OG^{r}\}$, $\{OG^{l}\}$)
    // step 2: make the phylogenetic gene tree $\{P^{\mathbf{x}}\}$ for the orthogroups $OG^{\mathbf{x}}$
    repeat
        Choose an unprocessed $OG_i^{\mathbf{x}} \in \{OG^{\mathbf{x}}\}$
        // construct the unrooted phylogenetic tree topology
        $P_i^{\mathbf{x}} \leftarrow$ MakeTree($OG_i^{\mathbf{x}}$)
        // now use equation 1 to select the root
        RootTree($P_i^{\mathbf{x}}$)
        // break an orthogroup if it does not resolve to a single ancestral gene
        if $P_i^{\mathbf{x}}$.root $\not\subseteq \{g^{\mathbf{x}}\}$
            $(OG_k^{\mathbf{x}}, OG_l^{\mathbf{x}}) \leftarrow$ BreakOrthogroup($OG_i^{\mathbf{x}}$, $P_i^{\mathbf{x}}$)
            // update the set of putative orthogroups
            $\{OG^{\mathbf{x}}\} \leftarrow (\{OG^{\mathbf{x}}\} \setminus OG_i^{\mathbf{x}}) \cup (OG_k^{\mathbf{x}}, OG_l^{\mathbf{x}})$
        else
            Mark $OG_i^{\mathbf{x}}$ as processed
    until all orthogroups are processed
    UpdateSimilarityGraph($\mathbf{x}$, $\{OG^{\mathbf{x}}\}$)
return $\{OG^{\mathbf{x}}\}$
```

Fig. 3. Overview of the SYNERGY algorithm. The algorithm is initially called with the root of the species tree **T**.

point, the procedure updates the gene similarity graph by replacing the genes in species below x by orthogroups in $\{OG^{x}\}$, and the next Stage of the algorithm treats these orthogroups as genes. When the bottom-up recursion reaches the root of **T**, every gene g_i^{a} in each species has been assigned uniquely into an orthogroup and located as a leaf in the corresponding gene tree.

We now expand on the details of each step of the procedure.

2.4.1 Matching orthogroups At each node **x** of the species tree **T**, SYNERGY considers orthology assignments for orthogroups pertaining to the species directly below **x** in the species **T** (denoted **y** and **z**). As noted above, the orthogroups from both **y** and **z** are now vertices in the gene similarity graph. SYNERGY begins by matching orthogroups in both **x** and **y** into *candidate* orthogroups. We assign orthogroups into the same candidate orthogroup if they have reciprocal edges between them and apply transitive closure on these reciprocal relations. More precisely, for a pair of orthogroups OG_i ,$OG_j \in \{OG^{y}\} \cup \{OG^{z}\}$, we have that $OG_i \sim_{x} OG_j$ if either both $OG_i \rightarrow OG_j$ and $OG_j \rightarrow OG_i$ are in ε or if there is a third orthogroup $OG_k \in \{OG^{y}\} \cup \{OG^{z}\}$ such that $OG_i \sim_{x} OG_k$ and $OG_k \sim_{x} OG_j$. This leads to a partitioning of the orthogroups from species **y** and **z** into equivalence classes under $\sim_{x}$. Each such equivalence class is taken to be a single *candidate* orthogroup for x. We find this partitioning in a linear time (in the number of edges) using a standard connected component algorithm.

This step is similar to many hit-based methods (e.g. COGs, Tatusov *et al.*, 1997). Due to our lenient inclusion policy and the promiscuity of edges in the gene similarity graph, candidate orthogroups may contain genes (orthogroups) that are related through duplication events that predate x, and in fact descend from multiple genes in the ancestral species x. Such violations of the orthogroup soundness condition (Definition 1) are handled after each candidate orthogroup is arranged into a phylogenetic tree.

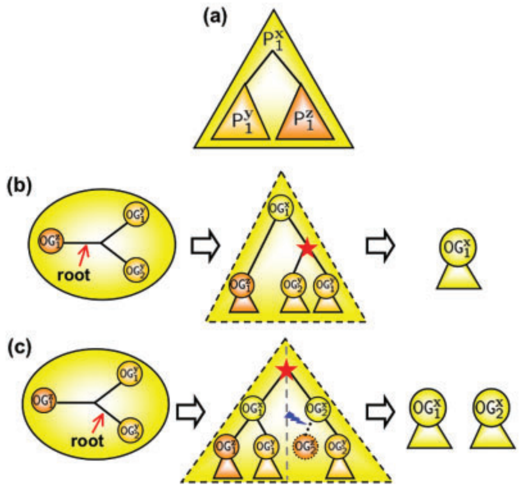


Fig. 4. Construction of phylogenetic gene trees. (**a**) A gene tree P_1^{x} for a candidate orthogroup OG_1^{x} is constructed by joining the trees P_1^{y} and P_1^{z} resolved in Stages **y** and **z**. (b, **c**) When OG_1^{x} consists of more than two members, there are several alternative rootings. In (b), a root is selected that invokes one duplication between species **y** and the root of the tree, such that OG_1^{y} and OG_2^{x} are paralogs. In (c), the rooting suggests a duplication at the root of the gene tree, such that OG_1^{x} and OG_2^{y} are paralagous with respect to a duplication predating **x** and OG_2^{z} is lost after the speciation event. If (c) is selected, the orthogroup must be broken.

2.4.2 Phylogenetic tree reconstruction Given a candidate orthogroup OG_i^{x} , we reconstruct a phylogenetic tree P_i^{x} whose leaves are the orthogroups from **y** and **z** that comprise OG_i^{x} (Fig. 4a). Recall that since the trees $\{P^{y}\}$ and $\{P^{z}\}$ were already resolved in previous Stages, we treat the root of each of these trees an extant gene in the phylogenetic reconstruction.

When only a pair of orthogroups OG_j^{y} and OG_k^{z} are matched into the candidate orthogroup OG_i^{x} , there is a clear one-to-one orthology relation, making this task trivial: the gene tree would appear exactly as the species tree appears at the point **x** (Fig. 4a). When an orthogroup OG_i^{x} contains one-to-many or many-to-many relationships (due to possible duplications and/or losses), we reconstruct the tree using the Neighbor-Joining method (Saitou and Nei, 1987) applied to the distance matrix between the orthogroups that comprise OG_i^{x} . The result is an unrooted phylogentic tree whose leaves are the orthogroups that have been matched together. (Note that we could replace Neighbor-Joining by other phylogentic reconstruction procedures; our choice was based on the efficiency and effectiveness of the Neighbor-Joining procedure.)

2.4.3 Tree rooting The resulting unrooted tree contains all of the orthogroup components that were matched into the candidate orthogroup. To obtain the exact phylogenetic relationships between these components, the tree must first be rooted. Correct rooting is important since the selected root position may determine whether all of an orthogroup's members descended from a single gene in species **x** or from multiple genes (Fig. 4b and c).

Assuming equal rates of evolution amongst all the leaves in a tree, a tree's root should be approximately equidistant to all the leaves. Given an unrooted tree, we compute the leaf-to-root variances for every possible rooting *r* at an internal branch in it, and assign a score to each rooting that is proportional to the variance in both amino acid and synteny scores, termed π_r and σ_r, respectively.

Following a gene duplication, one or both of the paralogs are often under relaxed selection, and can evolve at an accelerated rate (Lynch and Katju, 2004; Ohno, 1970). This conflicts with the assumption above that all branches of the tree evolve at an equal rate, and complicates tree rooting. We therefore introduce a preference for root locations that are more likely in terms of the number of duplication and loses it invokes. For each root position r, we compute the number of duplications and losses it implies for each branch s below **x** in the species tree (i.e. either y or z). We denote these as $\#\text{dups}_r^{\mathbf{s}}$ and $\#\text{loss}_r^{\mathbf{s}}$, respectively. To estimate the probabilities of such events, we assume that they are governed by a Poisson distribution.[4] We define

$$\begin{aligned}\omega_r &= \prod_{\mathbf{s}\in\{\mathbf{y},\mathbf{z}\}} P(\#\text{dups}_r^{\mathbf{s}} = d^{\mathbf{s}},\ \#\text{loss}_r^{\mathbf{s}} = l^{\mathbf{s}}) \\ &= \prod_{\mathbf{s}\in\{\mathbf{y},\mathbf{z}\}} \left(\frac{e^{-\delta_{\mathbf{s}}}\delta_{\mathbf{s}}^{d^{\mathbf{s}}}}{d^{\mathbf{s}}!}\right)\left(\frac{e^{-\lambda_{\mathbf{s}}}\lambda_{\mathbf{s}}^{l^{\mathbf{s}}}}{l^{\mathbf{s}}!}\right)\end{aligned}$$

where, δ_{s} and λ_{s} are the rates of duplication and loss at the branch s, respectively. These rates may either be learned by the algorithm through repeated iterations or be based on prior knowledge of the studied lineages (see Results section).

We select the root for each orthogroup OG_i by combining the three scores into a single rooting score ρ_r, reflecting the relative importance of each score. We select the rooting that maximizes:

$$\rho_r(\text{OG}_i) = -\alpha\pi_r + -\beta\sigma_r + \gamma\omega_r \tag{1}$$

where, α, β and γ are constants specifying the relative contribution to the rooting score of peptide similarity, synteny similarity and the likelihood of the invoked duplications and losses.

2.4.4 Breaking orthogroups Once a rooting r for an orthogroup tree $\text{P}_i^{\mathbf{x}}$ is chosen, we may find that the root of $\text{P}_i^{\mathbf{x}}$ no longer represents a single gene as the last common ancestor of all the genes present, but rather an earlier duplication event from which two ancestral genes were derived (Fig. 4c). This violates Definition 1, and we must therefore split the orthogroup's components at the root of its current tree $\text{P}_i^{\mathbf{x}}$. This situation frequently occurs when orthogroups are paralogous with respect to a duplication event that predates **x**.

This step allows us to be very permissive with the edges we include between genes in the gene similarity graph and in how we match candidate orthogroups. By admitting more edges, we include many spurious ones, but we also include edges that capture the many-to-many relations that may arise from duplications, ensuring that our orthogroups satisfy Definition 2 of orthogroup completeness. If the spurious edges cause non-orthologous orthogroups to be matched, an accurate rooting will subsequently lead the procedure to partition the candidate orthogroup into separate orthogroups. SYNERGY iterates this until each orthogroup represents a single ancestral gene and no orthogroups need to be partitioned.

2.4.5 Updating the Gene Similarity Graph Once we constructed orthogroups at the ancestral node **x**, we no longer need to consider the orthogroups in the species below **x** individually. We avoid doing so by removing vertices in the gene similarity graph that correspond to orthogroups in $\{\text{OG}^{\text{y}}\}$ and $\{\text{OG}^{\text{z}}\}$ and introducing new vertices that correspond to the newly created orthogroups in $\{\text{OG}^{\text{x}}\}$. The edges incident to the new vertices are acquired by taking the union of the edges that were incident to its constituent orthogroups (or genes).

To weight these new edges, we recall that each new orthogroup represents the root of a tree. Thus, we can use the standard distance update procedure used by distance-based algorithms such as Neighbor-Joining. Specifically, when vertices i and j are merged to form a new vertex k in the tree, the distances between k and and any other vertex m is calculated as $d_{km} = \frac{1}{2}(d_{im} + d_{jm} - d_{ij})$. This formula is applied in the order specified by the topology of orthogroups' corresponding gene trees. When one of the distances in question is not defined in the original similarity graph, we use the maximal distance value.

The edges in this updated similarity graph can always be traced to one (or more) edges between extant genes in the original similarity graph. However, reciprocal edges between two orthogroups (that might lead them to be merged into the same orthogroup in subsequent iterations) may originate from two different pairs of extant genes that are assigned to the two orthogroups. Thus, our matching criteria is able to capture non-trivial paths in relating the extant genes.

3 MEASURING ORTHOGROUP CONFIDENCE

To empirically measure SYNERGY's robustness to the specifics of a given dataset and to evaluate our confidence in each orthogroup's assignments, we developed a jackknife-based approach. By systematically and repeatedly excluding different portions of the data, we measure orthogroup robustness to (1) the choice of species included and (2) the accuracy of gene predictions within each species.

We test the soundness and completeness of the identified orthogroups. A complete orthogroup (Definition 2) contains *all* the genes that descended from a single common ancestor and thus its genes should not 'migrate out' of it in the holdout experiments. To test this, we count the number of orthologous gene pairs (g_j, g_k) in an orthogroup OG_i that remained orthologous in a holdout experiment.[5] We compute η_i^c for each orthogroup OG_i by counting the fraction of orthology assignments that remained constant across each holdout experiment h:

$$\eta_i^c = \frac{|\{(\text{g}_j, \text{g}_k) \in \text{OG}_i \mid \text{h}(\text{g}_j, \text{g}_k) = \text{OG}_i(\text{g}_j, \text{g}_k)\}|}{N} \tag{2}$$

where, $\text{h}(g_j , g_k)$ and $\text{OG}_i\ (g_j , g_k)$ specify the last species in the tree in which g_j and g_k share a common ancestor in the holdout experiment h and the original orthogroup, respectively (this is equal to -1 if g_j and g_k are not members of the same orthogroup), and N is the number comparisons made across all holdout experiments.

A sound orthogroup (Definition 2) contains *only* the genes that descended from a single common ancestor, and thus new genes should not 'migrate into' the orthogroup in the holdout experiments. We use a similar formula to obtain η_i^s , except we count the number of pairs of non-orthologous genes (g_j, g_k), $g_j \in \text{OG}_i$, $g_k \notin \text{OG}_i$ that became orthologous each the holdout condition h:

$$\eta_i^s = 1 - \frac{|\{(\text{g}_j, \text{g}_k) \notin \text{OG}_i \mid \text{h}(\text{g}_j, \text{g}_k) \neq -1\}|}{N} \tag{3}$$

Since pairs of genes that share no protein sequence similarity are highly unlikely to be considered orthologous in h, we

[4] The Poisson model assumes that these events occur as a memoryless process. This is likely true for most duplications and losses, a notable exception being loci with tandemly duplicated genes, where subsequent duplications and losses may occur at higher rates.

[5] We must account for the fact that some assignments are expected to change when genes within an orthogroup are among those hidden.

restrict our tests to gene pairs that can be loosely regarded as similar ($E < 0.1$), rendering this task computationally feasible.

The confidence measures, η_i^c and η_i^s , can be computed for both species- and gene-holdout experiments, giving us four measures of robustness for each orthogroup.

4 RESULTS

4.1 Test case: *Ascomycota* fungi

We applied SYNERGY to resolve the homology relations in a set of nine *Ascomycota* fungal species with a total of 52 092 protein coding genes (Fig. 5). This group of species includes the extensively studied model organism, *Saccharomyces cerevisiae*, and offers much ground for studies of genome evolution and function (Kellis *et al*., 2003). Recent studies show that a WGD event has occurred within this phylogeny, followed by widespread loss of paralogous genes (Byrne and Wolfe, 2005; Dietrich *et al*., 2004; Kellis *et al*., 2004). Proper resolution of paralogy and orthology relations is essential for studying such events.

To run SYNERGY, we first set the parameters α, β and γ, which weigh protein similarity, synteny similarity and probability of duplication and losses when rooting a gene tree (Section 2.4).We set $\alpha = 0.01$ and $\beta = 1.0$ since we found in initial experiments that once orthogroups were matched in Step 1, the synteny similarity score was more informative than protein similarity score for rooting the trees in Step 2. While synteny plays an important role in determining orthology relations amongst these species, SYNERGY is also applicable to lineages where chromosomal structure is not conserved (data not shown).

Since duplication and loss events are relatively rare, we set their rates $\delta = \lambda = 0.05$ per branch across all branches of the species tree, and $\gamma = 5.0$. The notable exceptions to these uniform rates are the branch containing the WGD and those immediately following it, where we expect a higher rate of duplications and losses, respectively. We can empirically estimate the specific duplication and loss rates at these branches by running SYNERGY iteratively, using rates derived from previous iterations during each iteration. In practice, we chose rates that were consistent both with previously reported observations for these branches (Byrne and Wolfe, 2005), as well as with our own iterative estimates. We set 0.5 as the duplication rate for the WGD branch and as the loss rate for the branches below it.

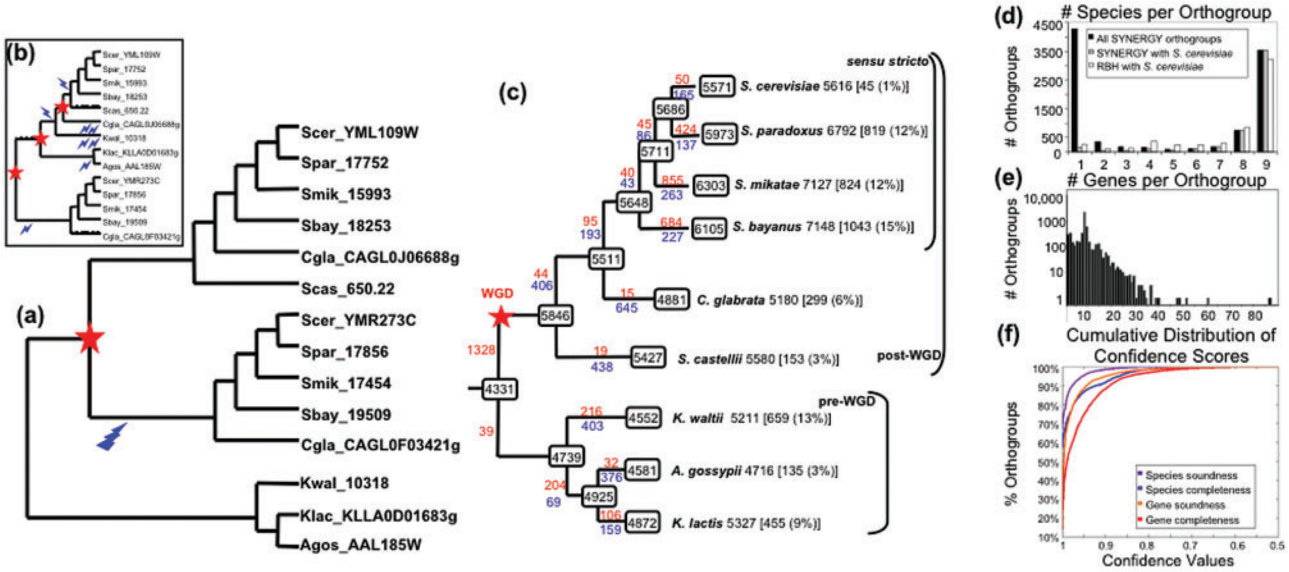


Fig. 5. Reconstructed gene trees identify duplication and loss events. (**a**) The gene tree reconstructed by SYNERGY for orthogroup OG#3184. The identified duplication and loss events are indicated by a star and a lightning bolt, respectively. The orthogroup contains the known *S.cerevisiae* paralogs, Zds1 and Zds2, which are the result of the WGD. The insert in (**b**) shows a gene tree constructed for the same set of genes using CLUSTALW's Neighbor-Joining algorithm (Thompson *et al*., 1994). A much larger number of duplication and loss events must be invoked to reconcile this tree with the known species phylogeny. (**c**) Orthogroup reconstruction in nine yeast species. A phylogenetic species tree of nine *Ascomycota* fungi (Scannell *et al*., 2006), six of which speciated after a WGD event (red star). The number of predicted protein coding genes in each extant species is noted next to the species' name with the number and percent of singletons in SYNERGY's predictions shown in brackets. The number of ancestral genes inferred from SYNERGY's gene trees are marked within each node of the species tree, and the numbers of duplication and loss events that occurred along each branch as derived from these trees are denoted one above the other in red and blue, respectively, next to each branch. By definition, no losses can be identified immediately below the root. (**d**) Species distribution of orthogroups is shown for the full SYNERGY results (black) and for the subset of orthogroups containing at least one gene from *S.cerevisiae* (gray). A similar distribution is also shown for the orthologous gene sets obtained with an RBH approach with *S.cerevisiae* as a reference species (white). (**e**) Gene distribution of orthogroups. Many orthogroups have more than nine genes, a result of gene duplication events. (**f**) Cumulative distributions of confidence scores for orthogroup soundness (purple,orange) and completeness (blue,red) under species (purple,blue) and gene (orange,red) holdout experiments. Most orthogroups are robustly sound and complete to both types of perturbations.

SYNERGY identified 5282 orthogroups containing more than a single gene, accounting for 48 265 (91.6%) of the protein coding genes. 4432 (8.4%) genes were assigned to singleton orthogroups, most (61%) of which we attribute to faulty open reading frame predictions amongst three *sensu stricto* species (data not shown).

By analyzing the phylogenetic gene trees corresponding to the identified orthogroups, we can automatically trace gene counts and ancestral duplication and loss events along the branches of the species tree used (Fig. 5a, b). We found 4331 orthogroups ancestral to all of the species studied (Fig. 5c). In addition, we found 187 orthogroups specific to the post-WGD *Saccharomyces* clade and 369 orthogroups specific to the pre-WGD clade. While some such cases might indicate SYNERGY's failure to resolve orthologies between these orthogroups, they may also suggest the appearance of novel genes in these lineages or complete loss from one of the clades. Including an additional outgroup to this clade of species might help determine which is the most likely case for each orthogroup. SYNERGY also detected a large number of duplication events along the branches leading to the *sensu stricto* species *S.paradoxus* (424 duplications), *S.mikatae* (855), and *S.bayanus* (684) (Fig. 5c). These are mostly due the to the same faulty gene predictions as those causing a multitude of singletons in these species. Our results can therefore be used for improving genome annotations.

Two-thirds of the orthogroups SYNERGY identified contain at least one member from each of the nine species and 80% were represented by at least eight of these species (Fig. 5d). Traditional approaches such as RBH fail to identify many of the orthologous gene sets that span many species. We compared our results with those attained by RBH anchored by *S.cerevisiae* and noticed a marked improvement in performance. (The large number of singleton orthogroups reported by our method is due to the inclusion of singletons from all species. When considering only singletons in *S.cerevisiae*, SYNERGY outperforms RBH.) We identified orthologs for 106 more genes in *S.cerevisiae* than RBH and identified 298 (200) more orthogroups spanning all nine (eight) species than RBH. This relative shortcoming of the RBH approach is compounded as the number of genomes included grows, rendering a significant disadvantage for comparative genomics studies across numerous species. The greatest weakness of the RBH approach is its assumption that orthology relations must be one-to-one. Of the orthogroups we identified, 52% contained at least one duplication or loss event. Many (55%) of these resulted in orthogroups containing more than nine member genes in them (Fig. 5e). By permitting multiple hits between genes in the gene similarity graph, SYNERGY was able to detect a large number of many-to-many orthologous relations while obtaining orthologies that were not overly coarse. A comparison to a more-inclusive hit clustering method is discussed below (Section 4.3).

4.2 Fungal orthogroup robustness

To obtain an objective measure of orthogroup robustness, we applied the species and gene jackknife procedures described in Section 3. For the gene-holdout experiments, we set the probability of hiding each gene at 0.1, and performed 50 holdout experiments. We performed the branch-holdout experiments by removing each branch in the tree separately once, resulting in 31 holdout experiments.

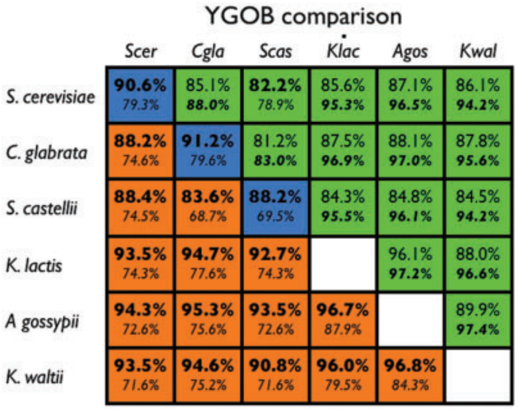
Of the non-singleton orthogroups identified, 79% had all four confidence values above 0.9 and 99% obtained a confidence value above 0.9 in at least one class of experiments (Fig. 5f). Perturbations to gene content were more disruptive than to species, and soundness was more robust than completeness (i.e. perturbations caused introduction of new 'incorrect' orthologies rather than loss of 'correct' ones).

As expected, orthogroups exhibiting higher frequencies of duplication and loss events tended to be most sensitive to such perturbations, although SYNERGY's performance was surprisingly robust for even those orthogroups. Lack of such duplication and loss events significantly correlated with higher confidence values ($p < 10^{-4}$ in all four measures). Overall, SYNERGY was remarkably robust to perturbations in the species phylogeny or noisy gene predictions.

4.3 Comparison to curated resource

High quality resolution of orthology and paralogy is essential for tracking genomic events and understanding their evolutionary impact. A notable example is the WGD that occurred within the lineage we study here (Dietrich *et al.*, 2004; Kellis *et al.*, 2004); see Fig. 5c, star). Recently Byrne and Wolfe, (2005) reported orthologies for six of the species we investigate based on sequence similarity, chromosomal alignment and intensive manual curation. This study is limited by its assumption that the WGD is the only duplication event among this lineage, and relies predominantly on synteny to assign orthology relations. Nonetheless, this curated resource, compiled in the Yeast Gene Order Browser (YGOB, Byrne and Wolfe, 2005), provides a 'gold standard' of orthology and paralogy relations which we use to evaluate our automated methods.

When we compared SYNERGY's orthology and paralogy assignments between those species included the YGOB resource, we find that our automatic predictions conform very well to those of this 'gold standard' (see Fig. 6). For example, SYNERGY was able to automatically identify over 80% of the orthology assignments between all pairs of species. More significantly, SYNERGY was able to resolve at a similarly high level of accuracy the precise paralogy relations within orthogroups where both paralogous copies were maintained in at least one species following the WGD. This task is challenging since determining which pairs of genes that were retained in duplicate are orthologous requires disambiguating between genes sharing high degrees of sequence similarity. We also compared the quality of SYNERGY's paralogy assignments to that of INPARANOID (Remm *et al.*, 2001), a hit-clustering method designed to identify paralogous relations between genes within genomes. SYNERGY identified more known paralogs dating to the WGD than INPARANOID did (Fig. 6). Unlike INPARANOID, SYNERGY also resolved orthologies and gene trees for *multiple* species simultaneously.



YGOB comparison

	Scer	Cgla	Scas	Klac	Agos	Kwal
S. cerevisiae	**90.6%** *79.3%*	85.1% ***88.0%***	**82.2%** *78.9%*	85.6% ***95.3%***	87.1% ***96.5%***	86.1% ***94.2%***
C. glabrata	**88.2%** *74.6%*	**91.2%** *79.6%*	81.2% ***83.0%***	87.5% ***96.9%***	88.1% ***97.0%***	87.8% ***95.6%***
S. castellii	**88.4%** *74.5%*	**83.6%** *68.7%*	**88.2%** *69.5%*	84.3% ***95.5%***	84.8% ***96.1%***	84.5% ***94.2%***
K. lactis	**93.5%** *74.3%*	**94.7%** *77.6%*	**92.7%** *74.3%*		96.1% ***97.2%***	88.0% ***96.6%***
A gossypii	**94.3%** *72.6%*	**95.3%** *75.6%*	**93.5%** *72.6%*	**96.7%** *87.9%*		89.9% ***97.4%***
K. waltii	**93.5%** *71.6%*	**94.6%** *75.2%*	**90.8%** *71.6%*	**96.0%** *79.5%*	**96.8%** *84.3%*	

Fig. 6. Comparison of SYNERGY and INPARANOID (Remm *et al.*, 2001) predictions to the gold standard of YGOB (Byrne and Wolfe, 2005). The matrix displays the sensitivity (orange cells) and specificity (green cells) of orthology assignments in YGOB that were automatically identified by SYNERGY (top number) and INPARANOID (bottom, italicized) for each pair of species. Because YGOB was designed specifically to study the WGD in these yeast species using syntenic relations, SYNERGY may include many orthology assignments that were not detected by YGOB due to lack of chromosomal evidence. The diagonal shows the percent of paralogues reported by YGOB that were detected by SYNERGY and INPARANOID (blue cells).

SYNERGY also showed greater sensitivity than INPARANOID when identifying orthology relations, albeit potentially at the cost of lower specificity. Some of the reduced specificity may be the result of a limitation of our gold standard. While YGOB's annotations are highly accurate, their methodology is limited by two assumptions: (1) all duplication events originated in the WGD and thus orthology is at most a two-to-one relationship, and (2) gene order is nearly always conserved and thus can be used as the primary source of evidence for shared ancestry. These assumptions relegate a greater portion of genes as singletons without orthologs, and a far fewer proportion of their orthologous loci are ancestral to all of their species than those that SYNERGY identified (62% versus 82%). We therefore believe that many of the orthology assignments reported by SYNERGY but not by YGOB (or INPARANOID) (Fig 6, green cells) are likely to be correct assignments.

To study the contribution of including synteny in our approach, we re-ran SYNERGY on these data while ignoring the genes' locations. We found that synteny plays a relatively minor role in predicting a genes' correct orthologs, but contributed substantially to reconstructing the correct gene trees. For example, over 200 duplication events were detected at the root of the *sensu stricto* species when ignoring synteny, many of which should have been traced to the WGD. We believe that the contribution of synteny information to orthology prediction may be most noticeable in cases where genes are undergoing exceptionally slow or fast rates of evolution, as is often the case between paralogs undergoing gene conversion or neofunctionalization (Kellis *et al.*, 2004, data not shown). It is here that synteny can help the most when deciding how to root the gene tree in Stage 2 of the algorithm.

4.4 Simulated orthogroups

To obtain an objective measure of SYNERGY's accuracy, we simulated orthogroup evolution across multiple rounds of speciation events along a tree designed to emulate the six species we used in our YGOB comparison above. We seeded the simulation with 1000 randomly drawn genes from *S.cerevisiae* and used the SEQ-GEN (Rambaut and Grassly, 1997) program to simulate the evolution of protein sequences using the JTT model of amino acid substitutions (Jones *et al.*, 1992). In this experiment, we ignored chromosomal ordering of the simulated sequences. Our simulation included duplicating or losing a gene along its evolution with rates proportional to those observed among the fungal data. The rates of amino acid substitution between lineages were specified by the branch lengths in the simulated gene trees. These lengths were exponentially distributed according to the observed lengths in the fungal dataset.

SYNERGY identified 87.1% of the pairwise orthologous relations between these simulated orthogroups at a specificity rate of 98.6%. We compared these results to predictions made by INPARANOID on the same data and found SYNERGY to be substantially more sensitive than INPARANOID at a comparable rate of specificity: INPARANOID was only able to identify 69% of orthologous relations with a specificity of 99.2%. We did not consider SYNERGY's accuracy in reconstructing the correct phylogenetic trees for this comparison, as INPARANOID does not perform this task.

5 DISCUSSION AND FUTUREWORK

Here we present a framework for identifying groups of orthologous genes in a step-wise manner, while simultaneously reconstructing a corresponding phylogenetic gene tree for each group. We describe a novel algorithm—SYNERGY—that uses this framework to reconstruct a genome-wide catalog of gene trees across multiple species by incorporating multiple sources of information, including sequence similarity and conserved gene order (if relevant). SYNERGY's gene trees reflect the evolutionary history of each group of genes, allowing us to accurately identify orthology and paralogy relations between genes, and the duplication, loss and divergence events that underly these relations.

Our approach has several important benefits. First, its accurate automatic genome-wide resolution is unprecedented. It is typically absent from automatic 'hit clustering' methods applied on a genomic scale, which either ignore paralogs altogether (RBH), or do not make detailed distinctions between orthologs and paralogs (Tatusov *et al.*, 1997). While complementary phylogenetic methods aim to reconstruct such relations, they typically do so for a single (predefined) group of orthologs, and do not scale well to whole genomes.

More recent genomics approaches still require extensive manual curation to provide reasonable results (Byrne and Wolfe, 2005; Li *et al.*, 2006), and are hence of limited utility, especially given the rapid increase in fully sequenced genomes. Even the most recent advances combining hit-clustering and phylogenetic approaches are greatly limited by the phylogenetic reconstruction step, resulting in either low-quality trees or restrictions on the considered sequence similarities or gene family size (Blomme *et al.*, 2006; Jothi *et al.*, 2006). In contrast, since SYNERGY's gene tree reconstruction is constrained a priori by the topology of the species tree, we do not have to apply extra reconciliation steps (e.g. Durand *et al.*, 2006). For example, in orthogroups that have no duplication and loss events, our algorithm is guaranteed to yield the correct gene tree.

SYNERGY strikes an important balance between orthogroup completeness (sensitivity) and soundness (specificity). We ensure completeness by allowing many edges (candidate homology relations) into the input gene similarity graph and by applying a lenient criterion to derive candidate orthogroups. Then, we achieve soundness by refining these coarse relations as we progress through the species tree, breaking orthogroups using phylogenetic principles at each Stage. The bottom-up design of our algorithm also renders it scalable, allowing us to handle a large number of species and genes.

To evaluate the quality of our results, we scored the robustness of orthogroup soundness and completeness to perturbations in either species or gene content. In addition to informing us on orthogroup quality, this framework can pinpoint 'sensitive' places in the phylogeny where additional species should be sequenced or where open reading frame predictions should be improved.

We tested SYNERGY on a set of nine *Ascomycota* fungal genomes, obtaining mostly high quality (robustly sound and complete) orthogroups, that covered over 90% of the genes in these genomes and were of comparable quality to a manually curated gold standard of orthology and paralogy assignments in this lineage. The orthogroups identified by SYNERGY provide better coverage than standard hit-clustering approaches, and their detailed gene trees provide invaluable information that allows us to trace individual gene families, as well as identify the histories of sets of orthogroups.

While our bottom-up approach provides high-quality results, it is nevertheless a greedy algorithm and can occasionally misassign genes. This greediness could be relaxed by adding top-down reassignments after the bottom-down reconstruction is completed. Formulating the orthology resolution problem within the framework of bottom-up orthogroup identification should provide an important paradigm for additional algorithmic solutions.

SYNERGY opens the way to a host of comparative genomics studies. Many groups of species are now being extensively sequenced (e.g. vertebrates) and can be tackled with our scalable algorithm. As shared history is one of the best indicators for shared function, using orthology and paralogy relations one can map individual genes, entire molecular systems, and whole chromosomal regions between species, and discover how their corresponding sequences, topologies and function change over evolutionary time.

ACKNOWLEDGEMENTS

We thank D Pe'er, I Nachman and R Nelken for critical reading of this manuscript, and M Ethier, B Matenuto and A Karger for excellent technical support. A.R. holds a Career Award at the Scientific Interface from the the Burroughs Wellcome Fund. N.F. was supported in part by an Israel Science Foundation grant.

Conflict of Interest: none declared.

REFERENCES

Arvestad,L. *et al.* (2003) Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics*, **19** (Suppl. 1), 7–15.

Blomme,T. *et al.* (2006) The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.*, **7**, R43.

Byrne,K.P. and Wolfe,K.H. (2005) The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.*, **15**, 1456–1461.

Cliften,P. *et al.* (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science*, **301**, 71–76.

Dietrich,F.S. *et al.* (2004) The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. *Science*, **304**, 304–307.

Dufayard,J.-F. *et al.* (2005) Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, **21**, 2596–2603.

Dujon,B. *et al.* (2004) Genome evolution in yeasts. *Nature*, **430**, 35–44.

Durand,D. *et al.* (2006) A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J. Comput. Biol.*, **13**, 320–335.

Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.

Goodman,M. *et al.* (1979) Fitting the gene lineage into its species lineage, a parsimony strategy Illustrated by cladorams constructed from globin sequences. *Syst. Zool.*, **28**, 132–163.

Goodstadt,L. and Ponting,C. (2006) Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput. Biol.*, **2**, e133.

Jones,D.T. *et al.* (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, **8**, 275–282.

Jothi,R. *et al.* (2006) COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics*, **22**, 779–788.

Kellis,M. *et al.* (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.

Kellis,M. *et al.* (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature*, **428**, 617–624.

Koski,L.B. and Golding,G.B. (2001) The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.*, **52**, 540–542.

Li,H. *et al.* (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.*, **34**, 572–580.

Li,L. *et al.* (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.

Lynch,M. and Katju,V. (2004) The altered evolutionary trajectories of gene duplicates. *Trends Genet.*, **20**, 544–549.

Ohno,S. (1970) *Evolution by Gene Duplication*. George Allen and Unwin, London.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Rambaut,A. and Grassly,N.C. (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, **13**, 235–238.

Remm,M. *et al*. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol*., **314**, 1041–1052.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*., **4**, 406–425.

Scannell,D. *et al*. (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, **440** (7082), 341–345.

Tatusov,R.L. *et al*. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Thompson,J.D. *et al*. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.

Wall,D.P. *et al*. (2003) Detecting putative orthologs. *Bioinformatics*, **19**, 1710–1711.

Zmasek,C.M. and Eddy,S.R. (2001) A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics*, **17**, 821–828.