MSOAR 2.0: INCORPORATING TANDEM DUPLICATIONS INTO ORTHOLOG ASSIGNMENT BASED ON GENOME REARRANGEMENT

Guanqun Shi*
Department of Computer Science, University of California, Riverside, CA 92521, USA
**Email: gshi@cs.ucr.edu*

Liqing Zhang
Department of Computer Science, Virginia Tech, Blacksburg, VA 24060, USA
Email: lqzhang@vt.edu

Tao Jiang
Department of Computer Science, University of California, Riverside, CA 92521, USA
Email: jiang@cs.ucr.edu

Ortholog assignment is a critical and fundamental problem in comparative genomics, since orthologs are considered to be functional counterparts in different species and can be used to infer molecular functions of one species from those of other species. MSOAR is a recently developed high-throughput system for assigning orthologs between closely related species on a genome scale. It attempts to reconstruct the evolutionary history of input genomes in terms of genome rearrangement and gene duplication events. It assumes that a gene duplication event inserts a duplicated gene into the genome of interest at a random location (*i.e.*, the random duplication model). However, in practice, biologists believe that genes are often duplicated by tandem duplications, where a duplicated gene is located next to the original copy (*i.e.*, the tandem duplication model). In this paper, we develop MSOAR 2.0, an improved system for ortholog assignment. For a pair of input genomes, the system first focuses on the tandemly duplicated genes of each genome and tries to identify among them those that were duplicated after the speciation (*i.e.*, the so-called inparalogs), using a simple phylogenetic tree reconciliation method. For each such set of tandemly duplicated inparalogs, all but one gene will be deleted from the concerned genome (because they cannot possibly appear in any ortholog pairs), and MSOAR is invoked. Using both simulated and real data experiments, we show that MSOAR 2.0 is able to achieve a better sensitivity and specificity than MSOAR. In comparison with two well-known genome-scale ortholog assignment tools, the InParanoid program and the Ensembl ortholog database, MSOAR 2.0 shows the highest sensitivity. Although the specificity of MSOAR 2.0 is slightly worse than that of InParanoid in the real data experiments, it is actually better than that of InParanoid in the simulation tests. These experimental results demonstrate that MSOAR 2.0 is a highly accurate tool for ortholog assignment.

1. Introduction

Orthologs and *paralogs* are two different types of homologous genes that differ in the way that they evolved. Orthologs are genes in different species that evolved from a common ancestral gene due to speciation events while paralogs are duplicated genes in the same genome[1] . To better understand the evolutionary process, paralogs are further divided into two subtypes: *outparalogs* and *inparalogs*[2]. With respect to a given speciation event, outparalogs are genes duplicated before the speciation while inparalogs are genes duplicated after the speciation. Note that, the orthology between two species is in general a many-to-many relationship. In other words, for a pair of genomes, an ortholog group consists of a pair of sets of inparalogs, one from each genome. The inparalogs in one set are co-orthologous to all the inparalogs in the other. However, one may select a representative for each set of inparalogs (*e.g.*, the *exemplar gene*[3]) and define an ortholog pair for each ortholog group consisting of the two representatives. Such an ortholog pair may contain the two genes, one from each set, that correspond the best in terms of their

*Corresponding author.

positions on the genomes[4] or sequence similarity[2]. This allows us to think of orthology as a one-to-one relationship, which could help simplify the discussion in many cases and makes it possible to evaluate an ortholog assignment result against the orthology benchmark defined by gene symbols (which is a one-to-one relationship). Note that, once an ortholog pair is defined for an ortholog group, all other pairs of genes from the group will be regarded as false positives.

Clearly, it is easy to identify ortholog pairs between two species if the duplication history of the genes on the two genomes is given (relative to their speciation event). Unfortunately, this evolutionary process is unknown. What we know is all the genes in the contemporary genomes. In order to find the most probable ortholog assignment between two genomes, we need to reconstruct the true evolutionary history.

1.1. Existing Work on Ortholog Assignment

There exist many algorithms and tools for ortholog assignment, including the well-known COG system[5], InParanoid[2, 6], OrthoMCL[7], HomoloGene[8], TreeFam[9], PhyOP[10], and Ensembl Compara[11], just to name a few. A recent comprehensive review on ortholog assignment tools in the public domain can be found in Ref. 12. The first four tools, *i.e.*, COG, InParanoid, OrthoMCL and HomoloGene, are basically sequence similarity based methods that calculate pairwise similarity scores and employ some simple clustering algorithms to identify ortholog groups. For example, InParanoid assigns *main ortholog pairs* as the pairs of protein sequences with the highest bidirectional BLASTp scores (*i.e.*, *bidirectional best hits*, or *BBHs*), and uses them as "seeds" to identify inparalogs from both species by applying a heuristic clustering algorithm[2]. TreeFam, PhyOP and Ensembl Compara, on the other hand, explicitly reconstruct phylogenetic trees to infer the orthology relationship. Ensembl Compara, in particular, is a computational pipeline that combines some clustering method with phylogenetic tree reconciliation. It provides one-to-one, one-to-many, and many-to-many orthology relationships for more than 30 eukaryotic species[11]. However, none of these methods take gene order and genome rearrangement into account when they assign orthologs. It has been shown that genome rearrangement is very common between two closely related genomes[13–16], and thus the gene order information may help improve the accuracy of ortholog assignment.

By combining both sequence similarity and gene order information, a high-throughput ortholog assignment system called MSOAR[4, 17] has recently been developed. The system attempts to reconstruct the evolutionary history of the genes in the input genomes in terms of genome rearrangement and gene duplication events, and tries to minimize the *RD* (rearrangement and duplication) distance under the parsimony principle. MSOAR considers four genome rearrangement events including reversal (*i.e.*, inversion), translocation, fusion, and fission, and assumes that a gene duplication event inserts a duplicated gene into the concerned genome at a random location (*i.e.*, the random duplication model).

Figure 1 sketches an outline of the major algorithmic steps in MSOAR. In particular, it attempts to remove false ortholog pairs that involve genes randomly duplicated after the speciation in the "noise" gene pair detection step. Such a (false) ortholog pair usually incurs a great cost in the rearrangement distance between the genomes, and thus we would be able to reduce the RD distance by "uncoupling" (*i.e.*, removing) the pair. However, in reality, randomly duplicated genes only account for a part of all duplicated genes. Recent studies have shown that at least 30% of duplicated genes are found next to their original copies (*i.e.*, in tandem positions)[18, 19].

1.2. Gene Duplication Models

The importance of gene duplication in molecular evolution is well established[20, 21]. However, the biological mechanism behind gene duplication has been unknown for quite many years. Recently, biologists proposed three different mechanisms for gene duplication based on the size of the duplication and whether they involve an RNA intermediate[22, 23]: retrotransposition, tandem duplication, and genome duplication.

Retrotransposition describes the integration of a reverse transcribed mRNA into the genome in a random manner (see Figure 2), and is the cause of random duplications. Tandem duplication is one of

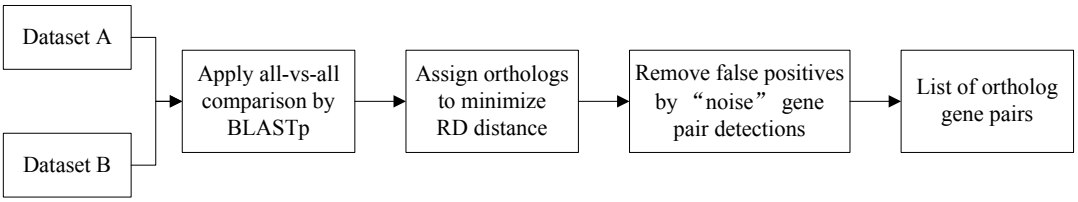

Fig. 1. An outline of MSOAR.

the possible outcomes of "unequal crossover", which results from the homologous recombination between paralogous sequences (see Figure 3). As a result, genes are duplicated next to their original copies in tandem arrays on the genome, which are known as *TAGs* (*i.e.*, *tandemly arrayed genes*)[19]. Genome duplication is probably due to the lack of disjunction between daughter chromosomes after DNA replication, and occurs more in plants than in animals. Recent studies show that there is another type of large-scale duplications, segmental duplication, which involves 1kb∼400kb nucleotides, though the molecular mechanism of segmental duplication is still unclear[23].

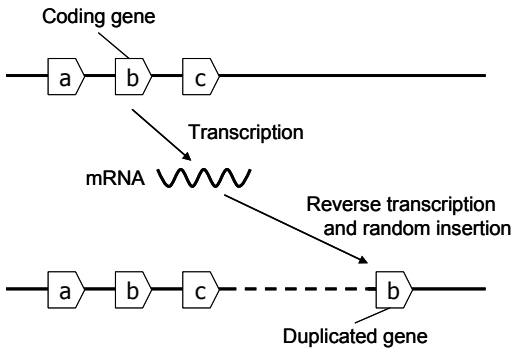

Fig. 2. Retrotransposition.

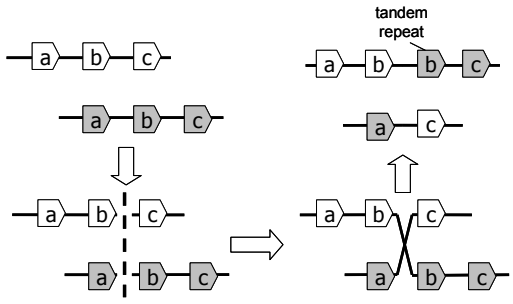

Fig. 3. Unequal crossover.

1.3. An Improved Ortholog Assignment System

Although MSOAR is able to identify most randomly duplicated inparalogs in the "noise" gene pair detection step, it is incapable of catching inparalogs that are produced by tandem duplications, which prevents MSOAR from identifying false ortholog pairs that involve two duplicated inparalogs in TAGs from both genomes. This is further illustrated in the following figures. Figure 4 shows how MSOAR identifies randomly duplicated inparalogs and Figure 5 depicts an example showing MSOAR's inability to treat the inparalogs in a TAG correctly.

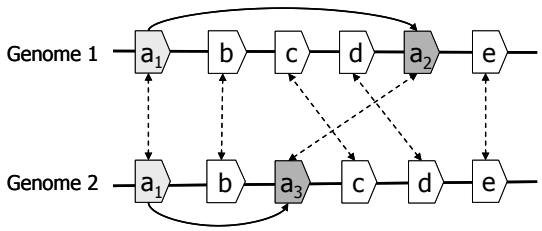

Fig. 4. Genes a_2 and a_3 are randomly duplicated from gene a_1.

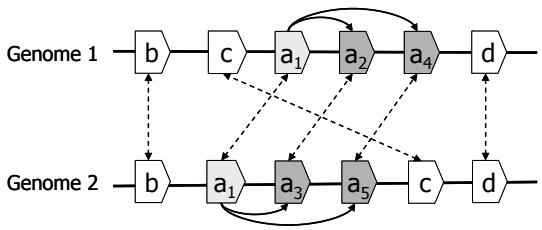

Fig. 5. Genes a_2, a_3, a_4, and a_5 are tandemly duplicated from gene a_1.

In Figures 4 and 5, we assume that the genes with the same letter from the two genomes represent true orthologs, and all duplications happened after the speciation in both genomes. For example, in Fig-

ure 5, (a_1, a_1) is a true ortholog pair while (a_2, a_3) and (a_4, a_5) are not. The genes a_2 and a_3 in Figure 4 and genes a_2, a_3, a_4 and a_5 in Figure 5 are all duplicated from gene a_1 after the speciation, and thus are inparalogs of a_1. In both cases, MSOAR first tries to assign orthology between all pairs of genes and calculates the RD distance between the two genomes. However, in the "noise" gene pair detection step, MSOAR is able to identify the false ortholog pair (a_2, a_3) in Figure 4 since the RD distance between the two genomes will decrease by 1 (*i.e.*, 3 fewer reversals and 2 more duplications) if this pair is removed. However, if the duplicated genes are in TAGs, as shown in Figure 5, removing any of the pairs (a_2, a_3) and (a_4, a_5) will not affect the number of reversals but will increase the number of duplications by 2, thus increasing the RD distance between the two genomes. Since MSOAR tries to find an assignment to minimize the RD distance between the two genomes, it will correctly identify the false ortholog pair (a_2, a_3) in Figure 4 while incorrectly keep both false ortholog pairs (a_2, a_3) and (a_4, a_5) in Figure 5 in the assignment.

In this paper, we incorporate the tandem duplication model into MSOAR, and develop an improved system to assign orthologs, simply called MSOAR 2.0. The idea is to consider tandemly duplicated genes first and try to identify the inparalogy relationship among them using a simple phylogenetic tree reconciliation method. For each set of inparalogs (on the same genome), all but one gene will be deleted from the concerned genome before MSOAR is invoked. Our experimental results demonstrate that this pre-processing step could indeed remove many false positives correctly and thus greatly improve the specificity of MSOAR.

The rest of the paper is organized as follows. The next section describes the details of MSOAR 2.0. The experimental results are presented in Section 3. Some concluding remarks are given in Section 4.

2. Methods

The pipeline of MSOAR 2.0 is outlined in Figure 6. The main steps in the pipeline are explained in detail in the following subsections.

2.1. Gene Family Definition and Construction

A gene family is defined to be the set of genes that are all descended from a common ancestral gene[4, 9]. Given two input genomes, our improved system starts by constructing gene families for all the genes on both genomes. We mix all protein sequences on both genomes and calculate the pairwise similarity scores by applying an all-versus-all BLASTp comparison[24]. By analyzing the results of BLASTp, we obtain a square similarity matrix, whose elements contain sequence similarity measurements for each pair of proteins in the dataset. Gene families can be calculated using the MCL (Markov clustering) algorithm[25] with default parameters.

Based on probability and graph flow theory, MCL simulates random walks on a graph using Markov matrices to determine the transition probabilities among the vertices of the graph. Unlike many other protein sequence clustering algorithms, MCL is able to deal with the presence of multi-domain proteins, promiscuous domains and fragmented proteins, making it one of the most widely used clustering algorithms in bioinformatics[25, 26]. Some papers use MCL directly to identify ortholog groups such as OrthoMCL[7], while some others use TribeMCL (an extension of MCL) as a tool to find paralogs within a genome[19]. In our system, we apply MCL to cluster all homologous genes on both genomes (including all possible orthologs and paralogs) into gene families.

2.2. dS-based Distance Matrix Generation

In order to construct a gene tree, we need to measure the pairwise distances within a gene family. This is done by performing a multiple alignment using ClustalW[27] first and then calculating a distance matrix for each gene family. Unlike most of the other phylogenetic approaches that measure pairwise gene distances based on amino acid substitutions, such as TreeFam[9], we choose to use the synonymous substitution rate (*i.e.*, the dS value) between two genes as the distance proxy following Ref. 10. Since silent mutations in coding DNA sequences do not lead to changes in their protein products, synonymous substitutions are under less selective constraint than

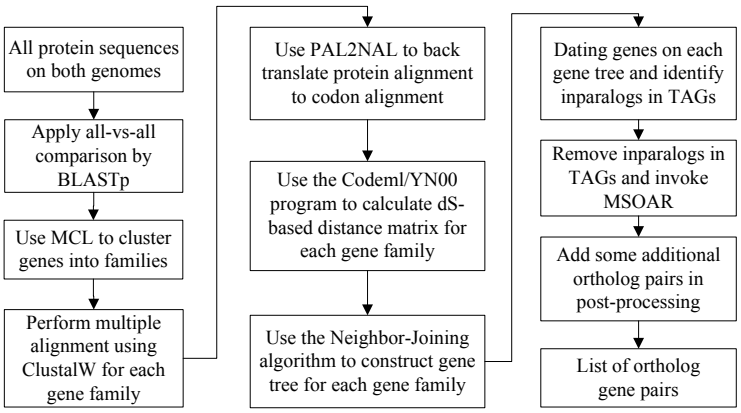

Fig. 6. An outline of MSOAR 2.0.

other coding sites. Therefore, they could more accurately reflect the neutral mutation rate between two genes.

In order to calculate the pairwise dS values, we reverse translate each multiple protein sequence alignment into a multiple codon alignment using the program PAL2NAL[28], which is a pattern matching algorithm that maps each amino acid to its corresponding codon sequence. Finally, the distance matrix based on the synonymous substitution rate is calculated by applying the YN00 program in the PAML package[29].

2.3. Gene Tree Reconstruction and Duplication Dating

The gene tree for each gene family is reconstructed by running the Neighbor-Joining algorithm[30], which is a pretty fast algorithm even for large gene families. In the process of gene tree reconstruction, we manually introduce a gene that is equally distant from all the other genes in a family as the outgroup in order to root the gene tree for each family. Once a gene tree is reconstructed, we need to label each of its internal nodes as either a duplication event or a speciation event. This process is a special case of the *gene duplication dating* problem, or the problem of reconciling a gene tree with a species tree. The phylogenetic tree reconciliation problem has been studied extensively in the literature, and many exact and heuristic algorithms have been proposed (see, *e.g.*, Ref. 31). In our case, since only two species are involved, we propose a straightforward algorithm to date the duplication events in linear time.

To avoid postulating unnecessary gene losses, every internal node with descendant genes from the same species is labeled as a duplication event. Then, the lowest internal nodes with descendant genes from both species are labeled as speciation events. All ancestral nodes of the speciation nodes must be labeled as duplication events since there are only two species. An example of such a gene duplication dating algorithm is shown in Figure 7.

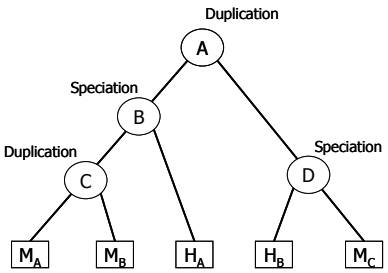

Fig. 7. An example of the gene duplication dating algorithm. Node C is a duplication event since M_A and M_B are both from the same species. Node B and D correspond to speciation events since they have descendant genes from two species. Node A is a duplication event since it is the ancestral node of speciation nodes B and D.

2.4. Identification of Inparalogs in TAGs

After dating duplications in a gene tree, we may deem each set of genes duplicated after the speciation event as a potential set of inparalogs (*e.g.*, M_A and M_B in Figure 7). In order to confirm a potential set of inparalogs, we need to consider the positions of the genes on the concerned genome. If the potential inparalogs are adjacent to each other on the genome, *i.e.*, they appear in the same TAG, then we define

them as inparalogs. For each such set of inparalogs, at most one gene can be included in an ortholog pair. Since these genes appear in tandem, it would make no difference to the RD distance (which is the objective function of MSOAR) which of them is chosen to represent the set in the ortholog pair. Thus, we will keep the gene that has the highest similarity score against any gene in the other genome and remove the other inparalogs in the same set so they will not be considered by MSOAR later on. If some potential inparalogs are separated by other genes on the genome, they will all be kept at this step and dealt with by MSOAR later on.

2.5. Invocation of MSOAR and Post-Processing

After removing duplicated inparalogs in TAGs on each genome, MSOAR is now invoked on the remaining genes. To further improve the performance of MSOAR, we use a post-processing step. If we consider the positions of the orthologs assigned by MSOAR on each genome, we find that in many cases a large consecutive block of assigned genes on one genome are orthologous to a consecutive block of assigned genes on the other genome with the same or reverse orientation. However, in some cases, there is a single unassigned gene (called a "gap") in each of the blocks forming an orthologous pair, and the gap appears at the same relative location in both blocks (see Figure 8). If the sequences of the two genes in the corresponding gaps are sufficiently similar (*e.g.*, at least one of the genes is the best hit of the other), then we deem that two genes as an ortholog pair and add it to the output list.

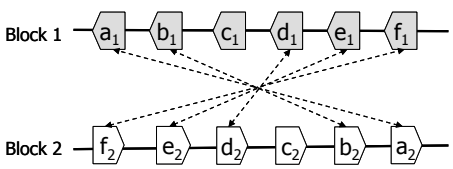

Fig. 8. An example of the post-processing. Block 1 and block 2 are orthologous blocks between two genomes, where (a_1, a_2), (b_1, b_2), (d_1, d_2), (e_1, e_2), (f_1, f_2) are ortholog pairs assigned by MSOAR. c_1 and c_2 are the corresponding gaps that have not be assigned orthology. If one of them is the best hit of the other, then we deem (c_1, c_2) as an additional ortholog pair and add it to the output.

3. Experiments and Results

In order to test the performance of MSOAR 2.0, we apply it to both simulated and real data, and compare our results with MSOAR[4], the popular ortholog assignment tool InParanoid[6], and the Ensembl ortholog database[11].

3.1. Simulation Results

To assess the accuracy of ortholog assignment, we simulate two input (single-chromosomal) genomes by using random duplications, reversals, and point mutations. The simulation is controlled by a set of 4 parameters (k, p, α, β), where k denotes the number of duplications in the ancestral genome before the speciation, p is the total number of genome-level evolutionary events (*i.e.*, duplications and reversals) on each genome after the speciation, α is the percentage of duplications among the p events, and β is the percentage of tandem duplications among all duplications.

The simulation is performed as follows. We first generate an ancestral genome G with 100 genes, each of which is a random sequence of 1,000 amino acids. We randomly perform k duplications in G to obtain another genome H. Then, a speciation happens and the genome H evolves into two contemporary genomes H_1 and H_2. The evolution from genome H to each of the contemporary genomes involves p evolutionary events, including $p \cdot \alpha$ duplications and $p \cdot (1 - \alpha)$ reversals. Among all duplications, β of them are tandem (*i.e.*, we randomly choose a gene and insert its copy next to it) while the others are random (*i.e.*, we randomly choose a gene and insert its copy randomly into the genome). In order to simulate the sequence change of each gene along the evolutionary process, we set a constant mutation rate $\mu = 1\%$ to allow each gene on the genome to have up to μ mutations of its amino acids between every two evolutionary events.

Using H_1 and H_2 as input genomes, we run MSOAR 2.0, MSOAR, and InParanoid separately. From the outputs of the three programs, we can easily compare their prediction accuracies in terms of sensitivity and specificity. Note that, InParanoid actually outputs ortholog groups. For each ortholog group, we take the first pair of genes in the group as

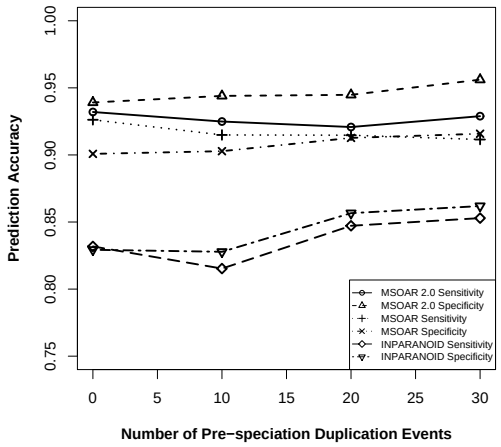

Fig. 9. Simulation results on the parameter set $(*, 50, 75\%, 50\%)$ where the parameter k is varied.

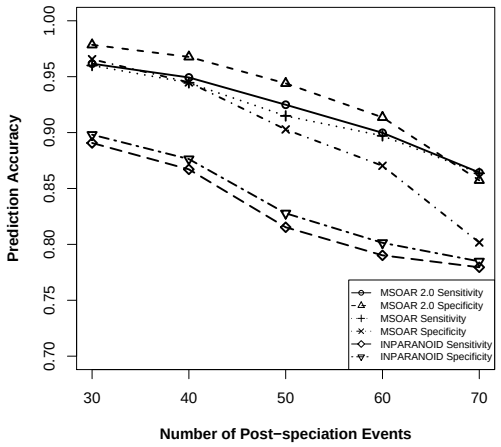

Fig. 10. Simulation results on the parameter set $(10, *, 75\%, 50\%)$ where the parameter p is varied.

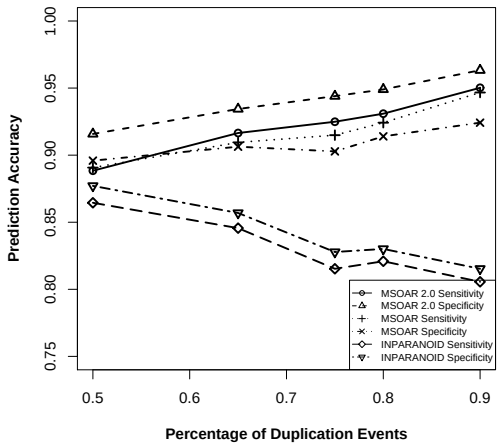

Fig. 11. Simulation results on the parameter set $(10, 50, *, 50\%)$ where the parameter α is varied.

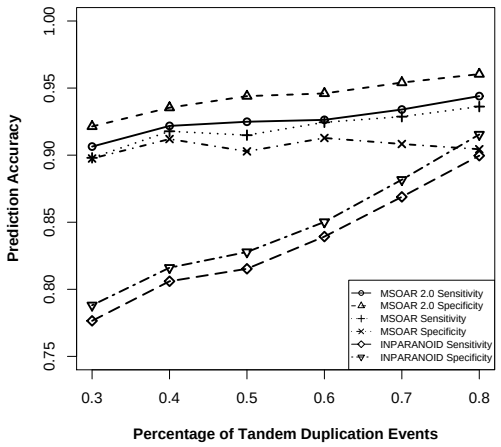

Fig. 12. Simulation results on the parameter set $(10, 50, 75\%, *)$ where the parameter β is varied.

the ortholog pair (which is referred to as the *main ortholog pair* in Ref. 2).

Since different parameters produce different input genomes, which may affect the prediction accuracies of the three programs, the parameters are varied as follows. We use a default parameter set and change the value of one parameter at one time. Based on recent studies on the relative ratios of various genome-level evolutionary events[19, 32], we choose to use $(10, 50, 75\%, 50\%)$ as our default parameter set. For each parameter set, 50 random datasets are simulated and the average prediction accuracies of the three programs are calculated. The performance of the three programs on various parameter sets are shown in Figures 9-12.

From Figures 9-12, we can see that parameter k has little effect on the prediction accuracies of the three programs as it only defines the number of out-paralogs. Parameter p, on the other hand, has a great impact on the performance of all the programs. With the increase of p, the prediction accuracies of all the three programs sharply decrease. This is because

when the number of evolutionary events increases, it is more difficult for MSOAR and MSOAR 2.0 to correctly reconstruct the evolutionary history based on the parsimony principle. Also orthologous genes may become less similar to each other for InParanoid to correctly identify them based on sequence similarity. Parameter α defines the ratio between duplications and reversals. As α goes up, the number of duplications increases while the number of reversals decreases. It becomes easier for MSOAR and MSOAR 2.0 to correctly identify reversals and assign orthologs while it becomes harder for InParanoid to differentiate main orthologs from their duplicated inparalogs due to the large number of duplications. Parameter β defines the ratio between tandem duplications and random duplications. As the ratio of tandem duplications goes up, the sensitivity and specificity of MSOAR 2.0 and InParanoid increase greatly while the performance of MSOAR remains almost unchanged. Note that, when $\beta > 80\%$, the specificity of InParanoid even beats that of MSOAR. This is due to MSOAR's inability to deal with tandem duplications. While MSOAR 2.0 may correctly identify many inparalogs in TAGs based on phylogenetic analysis and remove most of them before assigning orthologs and InParanoid clusters most inparalogs into their ortholog group and outputs only the main ortholog pair for each group, MSOAR tends to assign inparalogs in TAGs as ortholog pairs, introducing many false positive pairs.

The figures show that, in general, MSOAR 2.0 and MSOAR are more accurate than InParanoid in terms of both sensitivity and specificity on randomly simulated data. The sensitivity of MSOAR 2.0 is a little bit better than that of MSOAR while its specificity is much higher than that of MSOAR.

3.2. Real Data Experiments

In order to evaluate the performance of MSOAR 2.0 on real data, we apply MSOAR 2.0 to several real datasets. Since the human genome is the best annotated genome and has been used as the reference genome to assign gene symbols for other species, we use it as the "center" in our pairwise comparisons and compare it with four other mammalian genomes, mouse, rat, chimpanzee, and macaque that have been completely sequenced. Protein sequences, transcripts, and gene locations for all five species, human (*Homo sapiens*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), chimpanzee (*Pan troglodytes*) and macaque (*Macaca mulatta*) (version 52, December 2008) were downloaded from Ensembl genome browser (http://www.ensembl.org/). Genes annotated as novel, supercontig, or mitochondrial are removed, and only protein-coding genes with known chromosome locations are kept. For genes with alternative splicing variants, we use their longest transcripts. Similar methods have been used in the previous studies[19, 33]. After such data pre-processing, we obtained 21,164, 23,228, 22,490, 18,572, and 21,023 genes for human, mouse, rat, chimpanzee, and macaque, respectively.

3.2.1. *Results on Human, Mouse and Rat*

For the ortholog assignments between human and mouse and between human and rat, Table 1 shows the contributions of each major step in MSOAR 2.0. The phylogenetic analysis step is able to identify more than 1,000 duplicated inparalogs in TAGs in each species (1,118/2,524 for human-mouse and 1,211/2,030 for human-rat), and remove most of them before MSOAR is invoked. Then orthology is assigned by MSOAR on the remaining genes on each genome. Finally, in the post-processing step, MSOAR 2.0 is able to catch a few hundred ortholog pairs (125 for human-mouse and 128 for human-rat) from the "gaps" between consecutive ortholog blocks on each genome.

In order to validate the prediction results of MSOAR 2.0, we choose to use gene symbols. Gene symbols are used by researchers to refer to a specific gene of interest across species. Each symbol for a species should be unique and each gene within a genome should be given only one approved gene symbol[34]. The nomenclature of a gene is done by the nomenclature committees for each species. At present, there are only three official nomenclature committees in the world, for human, mouse, and rat respectively. So only these three species have official gene symbols. To obtain the most accurate gene symbol lists, we download the most recent gene symbols for human, mouse, rat from HGNC (http://www.genenames.org/), MGI (http://www.informatics.jax.org/), and RGD

Table 1. Contributions of the major steps in MSOAR 2.0.

Pair of Species	Inparalogs in TAGs Identified by Phylogenetic Analysis	Orthologs Assigned by MSOAR	Orthologs Assigned after Post-Processing
human vs mouse	1,118 / 2,524	16,635	16,760
human vs rat	1,211 / 2,030	15,781	15,909

(http://rgd.mcw.edu/) respectively, all of which are the official nomenclature committees for the involved species.

To compare the performance of MSOAR 2.0 with MSOAR, InParanoid and the Ensembl ortholog database, we consider the gene symbols of each output ortholog pair. Some genes may not have official gene symbols. Some symbols may not be meaningful, *e.g.*, when they are composed of "LOC" and gene ID, or when the gene functions have not yet been validated. In the latter case, the genes only have transcript identifiers (*e.g.*, gene symbols with the prefix "OTTMUSG" or the suffix "RIK" in the mouse genome). For each pair of orthologs, if both genes have identical official gene symbols, we count it as a true positive pair (*i.e.*, *TP*). If the genes have different official gene symbols, we count it as a false positive pair (*i.e.*, *FP*). If only one gene in the pair has an official gene symbol and another gene on the other genome (which is not in the pair) has the same gene symbol, then this pair is also considered as a false positive pair. For all other cases, we deem the pair as an unknown pair and ignore it in the accuracy assessment. We also calculate the assignable true ortholog pairs between two species by counting the number of identical gene symbols. The performance of the four methods validated using gene symbols is shown in Table 2. The actual ortholog assignment results of MSOAR 2.0 can be downloaded from the MSOAR website (http://msoar.cs.ucr.edu/).

Table 2 suggests that MSOAR 2.0 achieves the best sensitivity among the four programs although its specificity is slightly worse than that of InParanoid. A detailed analysis on differences in the ortholog assignment results of the four programs is given in Table 3.

Since InParanoid is a sequence similarity based method, all of the orthologs assigned by InParanoid are BBHs. Although many of the true orthologs may be indeed BBHs, some of them are not. In fact, more than 80% of the true ortholog pairs assigned by MSOAR 2.0 but missed by InParanoid in the human-mouse and human-rat comparisons (412/491 for human-mouse and 401/430 for human-rat) are not BBHs as shown in Table 3 (the first two columns).

While we define orthology between two genomes as a one-to-one relationship, the Ensembl ortholog database presents orthology in general as a many-to-many relationship. Thus, for each ortholog group, it outputs all pairs of genes consisting of one gene from one genome and another from the other. As a result, the specificity of the Ensembl ortholog database is quite low because each large ortholog group may re-

Table 2. Comparison of the performance of four programs using gene symbol validation. Again, to assess the accuracy of InParanoid, we take the first pair of genes in each ortholog group (*i.e.*, the main ortholog pair of the group) as an ortholog pair. For the Ensembl ortholog database, we directly download all the ortholog pairs from Ensembl Biomart Browser, which includes one-to-one, one-to-many, and many-to-many orthology relationships.

Pair of Species	Program	Assignable	Total Assigned	True Positives	Unknowns	Sensitivity	Specificity
human vs mouse	InParanoid	14,341	16,058	13,216	1,394	92.16%	90.13%
	Ensembl	14,341	20,670	13,619	2,850	94.97%	76.43%
	MSOAR	14,341	16,769	13,528	1,554	94.33%	88.91%
	MSOAR 2.0	14,341	16,760	13,629	1,542	95.04%	89.56%
human vs rat	InParanoid	12,688	15,197	11,750	1,529	92.61%	85.97%
	Ensembl	12,688	18,814	12,004	2,490	94.61%	73.54%
	MSOAR	12,688	15,883	11,970	1,723	94.34%	84.53%
	MSOAR 2.0	12,688	15,909	12,074	1,730	95.16%	85.15%

Table 3. Differences between the ortholog pairs assigned by MSOAR 2.0 and those by the other three programs. (a) This column lists the number of TPs found by MSOAR 2.0 but missed by InParanoid. (b) This column lists the number of TPs in the previous column that are not BBHs. (c) This column lists the number of FPs found by Ensembl but not by MSOAR 2.0. (d) This column lists the number of FPs in the previous column that are inparalogs occurring in TAGs. (e) This column lists the number of FPs found by MSOAR but not by MSOAR 2.0. (f) This column lists the number of FPs in the previous column that are inparalogs occurring in TAGs.

Pair of Species	MSOAR 2.0 vs InParanoid		MSOAR 2.0 vs Ensembl		MSOAR 2.0 vs MSOAR	
	TPs in M2-I[a]	Not BBHs[b]	FPs in E-M2[c]	Inparalogs in TAGs[d]	FPs in M-M2[e]	Inparalogs in TAGs[f]
human vs mouse	491	412	2,981	2,614	314	293
human vs rat	430	401	2,646	2,295	257	245

sult in many false positives. What is interesting is that even though it outputs a large number of ortholog pairs, its sensitivity is still a little bit worse than that of MSOAR 2.0 in both human-mouse and human-rat comparisons as shown in Table 2. It is interesting to observe that most of the false positive pairs output by Ensembl but not by MSOAR 2.0 (*i.e.*, 2,614/2,981 for the human-mouse comparison and 2,295/2,646 for the human-rat comparison) were actually found by MSOAR 2.0 to be inparalogs that appear in some TAGs, as shown in Table 3 (the two middle columns).

The last two columns of Table 3 clearly demonstrate that MSOAR 2.0 achieves a better specificity than MSOAR because of its treatment of TAGs, since most of the false positives output by MSOAR but not by MSOAR 2.0 (293/314 and 245/257 for the human-mouse and human-rat comparisons, respectively) were identified as inparalogs in TAGs by MSOAR 2.0.

3.2.2. *Results on Human, Chimpanzee and Macaque*

Since chimpanzee and macaque do not have official gene symbols, we only compare our assignment results with those of InParanoid and the Ensembl ortholog database. Figures 13 and 14 use Venn diagrams to show the commonality and difference among the ortholog pairs assigned by MSOAR 2.0, InParanoid, and the Ensembl ortholog database. We see that the three programs share more than 75% of the ortholog pairs. InParanoid outputs the least number of unique ortholog pairs while Ensembl has the most. More than 70% of the ortholog pairs unique to Ensembl are found to be inparalogs in TAGs (result not shown).

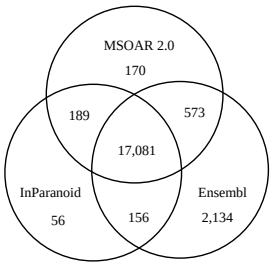

Fig. 13. Orthologs assigned between human and chimpanzee.

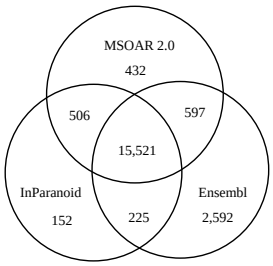

Fig. 14. Orthologs assigned between human and macaque.

Table 4 shows the number of ortholog pairs output by MSOAR 2.0 that are shared by at least one of the other two programs. We observe that the closer the compared species is to human, the more support the result of MSOAR 2.0 receives from the other programs. For a pair of very closely related species, such as human and chimpanzee, the ortholog pairs assigned by MSOAR 2.0 have more than 99% support from at least one of the other two programs, which is consistent with our expectations and confirms that MSOAR 2.0 is a highly accurate tool for ortholog assignment between closely related species.

Finally, we also observe that the number of inparalogs found in human by MSOAR 2.0 increases with the increase of evolutionary distance between human and the other species, as shown in Table 5. This is consistent with the definition of inparalogs.

Table 4. Support of the MSOAR 2.0 ortholog pairs by the other two programs.

Support	human vs chimpanzee	human vs macaque	human vs mouse	human vs rat
By both programs	94.83%	91.00%	90.08%	87.97%
By at least one program	99.06%	97.47%	97.05%	96.79%

Table 5. Inparalogs found in human and the other species by MSOAR 2.0.

Inparalogs found by MSOAR 2.0	human vs chimpanzee	human vs macaque	human vs mouse	human vs rat
Inparalogs in human	3,151	4,108	4,404	5,255
Inparalogs in the other species	559	3,967	6,468	6,581

4. Conclusion and Discussion

In this paper, we have incorporated a new gene duplication model, the tandem duplication model, into MSOAR, and developed an improved system of ortholog assignment by combining gene phylogeny and genome rearrangement. By comparison with MSOAR, InParanoid, and the Ensembl ortholog database on both simulated and real data, we showed that MSOAR 2.0 achieves the best overall prediction accuracy. Although MSOAR 2.0 has a slightly lower specificity as compared to InParanoid on real data using gene symbols as the benchmark (*e.g.*, in the human-mouse comparison, 90.13% for InParanoid vs. 89.56% for MSOAR 2.0), it nevertheless identified several hundred of true ortholog pairs that were missed by InParanoid. Because the majority of the "missed" orthologs are not BBHs, which are what the InParanoid assignment is based on, MSOAR 2.0 clearly addresses a weakness of InParanoid. Moreover, MSOAR 2.0 shows a better specificity in the simulation tests. Note that, MSOAR 2.0 also reconstructs the evolutionary history in terms of gene duplication and genome rearrangement, which could be of independent interest. Although Ensembl tends to assign a higher number of ortholog pairs than both InParanoid and MSOAR 2.0, MSOAR 2.0 outperforms it in terms of not only specificity but also sensitivity.

We evaluated the performance of the programs by computer simulations and gene symbols. However, simulations could be limited because the real evolutionary processes are much more complicated than what we can simulate. Furthermore, the use of gene symbols is not always feasible as many species do not have standard gene symbol assignment. We need to develop additional validation methods such as incorporating other available information, *e.g.*, gene functions. In addition, with the discovery of more mechanisms of gene evolution, new models of gene duplication (*e.g.*, segmental duplications) and genome operations (*e.g.*, *double cut and join* or DCJ), have been proposed. How to incorporate these new gene duplication models and operations into MSOAR 2.0 is our next challenge.

Acknowledgements

The research is supported in part by NSF grants IIS-0711129 and IIS-0710945, and NIH grant LM008991.

References

1. W. M. Fitch. Distinguishing homologous from analogous proteins. *Syst Zool*, 19(2):99–113, June 1970.
2. M. Remm, C. E. Storm, and E. L. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, 314(5):1041 – 1052, 2001.
3. D. Sankoff. Genome rearrangement with gene families. *Bioinformatics*, 15(11):909–917, 1999.
4. Z. Fu, X. Chen, V. Vacic, et al. MSOAR: A high-throughput ortholog assignment system based on genome rearrangement. *Journal of Computational Biology*, 14(9):1160–1175, 2007.
5. R. L. Tatusov, D. A. Natale, I. V. Garkavtsev, et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*, 29(1):22–28, 2001.
6. A. C. Berglund, E. Sjölund, G. Ostlund, et al. InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res*, 36(Database issue), January 2008.
7. L. Li, C. J. Stoeckert, and D. S. Roos. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9):2178–2189, 2003.

8. D. L. Wheeler, T. Barrett, D. A. Benson, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 34(suppl-1):D173–180, 2006.
9. H. Li, A. Coghlan, J. Ruan, et al. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res*, 34(suppl-1):D572–580, 2006.
10. L. Goodstadt and C. P. Ponting. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput Biol*, 2(9):e133, 09 2006.
11. A. J. Vilella, J. Severin, A. Ureta-Vidal, et al. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, 19(2):327–335, 2009.
12. A. Kuzniar, R. Vanham, S. Pongor, et al. The quest for orthologs: finding the corresponding gene across genomes. *Trends in Genetics*, September 2008.
13. S. Hannenhalli and P. Pevzner. Transforming men into mice (polynomial algorithm for genomic distance problem). In *FOCS '95: Proceedings of the 36th Annual Symposium on Foundations of Computer Science (FOCS'95)*, Washington, DC, USA, 1995. IEEE Computer Society.
14. W. J. Kent, R. Baertsch, A. Hinrichs, et al. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(20):11484–11489, 2003.
15. P. Pevzner and G. Tesler. Genome rearrangements in mammalian evolution: Lessons from human and mouse genomes. *Genome Research*, 13(1):37–45, 2003.
16. M. Semon and K. H. Wolfe. Rearrangement rate following the whole-genome duplication in teleosts. *Mol Biol Evol*, 24(3):860–867, 2007.
17. X. Chen, J. Zheng, Z. Fu, et al. Assignment of orthologous genes via genome rearrangement. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 2(4):302–315, 2005.
18. D. Pan and L. Zhang. Tandemly arrayed genes in vertebrate genomes. *Comparative and Functional Genomics*, 2008(545269), 2008.
19. V. Shoja and L. Zhang. A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat. *Mol Biol Evol*, 23(11):2134–2141, 2006.
20. S. Maere, S. De Bodt, J. Raes, et al. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A*, 102(15):5454–5459, April 2005.
21. S. Ohno. Evolution by gene duplication. pages 1–160, 1970.
22. M. Hurles. Gene duplication: the genomic trade in spare parts. *PLoS Biol*, 2(7), July 2004.
23. J. Zhang. Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, 18(6):292–298, June 2003.
24. S. F. Altschul, W. Gish, W. Miller, et al. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, October 1990.
25. A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucl. Acids Res.*, 30(7):1575–1584, 2002.
26. A. Alexeyenko, J. Lindberg, A. Perez-Bercoff, et al. Overview and comparison of ortholog databases. *Drug Discovery Today: Technologies*, 3(2):137–143, 2006.
27. R. Chenna, H. Sugawara, T. Koike, et al. Multiple sequence alignment with the Clustal series of programs. *Nucl. Acids Res.*, 31(13):3497–3500, 2003.
28. M. Suyama, D. Torrents, and P. Bork. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucl. Acids Res.*, 34(suppl-2):W609–612, 2006.
29. Z. Yang. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, 13(5):555–556, October 1997.
30. N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425, 1987.
31. C. Chauve, J. P. Doyon, and N. El-Mabrouk. Gene family evolution by duplication, speciation, and loss. *Journal of computational biology : a journal of computational molecular cell biology*, 15(8):1043–1062, October 2008.
32. J. M. Kidd, G. M. Cooper, W. F. Donahue, et al. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64.
33. R. Friedman and A. L. Hughes. The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Mol Biol Evol*, 20(1):154–161, 2003.
34. H. M. Wain, E. A. Bruford, R. C. Lovering, et al. Guidelines for human gene nomenclature. *Genomics*, 79(4):464–470, April 2002.