CLUSTERING OF MAIN ORTHOLOGS FOR MULTIPLE GENOMES

ZHENG FU

Department of Computer Science and Engineering, University of California Riverside Riverside, CA 92521, USA

zfu@cs.ucr.edu

TAO JIANG

Department of Computer Science and Engineering, University of California Riverside Riverside, CA 92521, USA

jiang@cs.ucr.edu

The identification of orthologous genes shared by multiple genomes is critical for both functional and evolutionary studies in comparative genomics. While it is usually done by sequence similarity search and reconciled tree construction in practice, recently a new combinatorial approach and a high-throughput system *MSOAR* for ortholog identification between closely related genomes based on genome rearrangement and gene duplication have been proposed in [11]. MSOAR assumes that orthologous genes correspond to each other in the most parsimonious evolutionary scenario minimizing the number of genome rearrangement and (post-speciation) gene duplication events. However, the parsimony approach used by MSOAR limits it to pairwsie genome comparisons. In this paper, we extend MSOAR to multiple (closely related) genomes and propose an ortholog clustering method, called MultiMSOAR, to infer main orthologs in multiple genomes. As a preliminary experiment, we apply MultiMSOAR to rat, mouse and human genomes, and validate our results using gene annotations and gene function classifications in the public databases. We further compare our results to the ortholog clusters predicted by MultiParanoid, which is an extension of the well-known program Inparanoid for pairwise genome comparisons. The comparison reveals that MultiMSOAR gives more detailed and accurate orthology information since it can effectively distinguish main orthologs from inparalogs.

1. Introduction

According to the definition of Fitch [10], *orthologs* are genes that evolved by speciation, while *paralogs* are genes that evolved by duplication. Orthologs typically occupy the same functional niche in different species, whereas paralogs tend to evolve toward functional diversification. Hence, the identification of orthologous genes shared by multiple genomes is critical for both the functional and the evolutionary aspects of comparative genomics.

The traditional ortholog identification methods could be categorized into two types. The first is sequence similarity-based methods, such as COG/KOG [23,22,24], EGO [15], Inparanoid/MultiParanoid [19,1], OrthoMCL [17], just to name a few. The

other is tree-based methods, including RAP [6], TreeFam [16], PhyOP [12], Orthostrapper [21], RIO [26], OrthologID [4], etc. The main assumption behind sequence similarity-based methods is that the evolutionary rates of all genes in a homologous family are equal and thus the divergence time could be estimated by comparing the DNA or protein sequences of genes. However, incorrect ortholog assignments might be obtained if the real rates of evolution vary significantly between homologs, and methods that rely on sequence similarity alone are highly subject to artificial association of slowly evolving paralogs and to erroneous exclusion of the more rapidly evolving genes [5]. Tree-based analysis is very intuitive and informative for ortholog identification, since it visually presents the history of a gene family[7]. Usually, orthologs and paralogs are identified by a *reconciled* tree, which is constructed to reconcile the incongruent gene and species trees by taking into consideration gene duplication events. However, tree-based approaches critically rely on the correctness of reconstructed gene and species trees. Moreover, reconstructing accurate gene trees for genome-wide scale analysis is very computation-intensive.

Recently, a new combinatorial approach and a high-throughput system MSOAR for genome-wide ortholog identification for closely related genomes based on genome rearrangement and gene duplication were proposed in [11]. MSOAR focuses on the assignment of a subtype of orthologs, called *main orthologs* which are formed by the *true exemplars* [20] from each pair of corresponding sets of inparalogous genes, [a] by computing the rearrangement/duplication distance between two genomes. The assumption is that main orthologs correspond to each other in the most parsimonious evolutionary scenario involving genome rearrangement and (post-speciation) gene duplication events. Since the true exemplar gene of an inparalogous set is the direct descendant of the ancestral gene of the set, it best reflects the original position and function of the ancestral gene in the ancestral genome. Hence, a reliable assignment of main orthologs is an important step toward the general identification of orthologs. The extensive tests on simulated data and real human and mouse genomes in [11] demonstrate that MSOAR has a comparable performance as Inaparanoid [19] and is able to find ortholog pairs that would be missed by Inparanoid (or any sequence similarity based methods). Moreover, its assignment result on human and mouse gonomes is well supported by the six methods listed on the HGNC Comparison of Orthology Predictions (HCOP) website (http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/hcop.pl)[9], Jackson Lab's human-mouse ortholog database [8], and the protein functions defined in Protein Analysis Through Evolutionary Relationships (PANTHER) classification system (http://www.pantherdb.org/) [25]. However, MSOAR requires the computation of the so called *RD* distance (*i.e.* genome rearrangement/duplication distance) between two given genomes [11], and is thus limited to pairwise comparisons.

In this paper, we present a new method to cluster main orthologs shared by

[a]With respect to a certain speciation event, the inparalogous genes are those that were generated by post-sepciation duplications [19].

multiple genomes, by extending MSOAR to more than two genomes. Given a set of genomes, the new method, called MultiMSOAR, first applies MSOAR to each pair of input genomes, and then it combines the pairwise ortholog assignment results from MSOAR consistently, taking into account the species phylogeny, to build main orthologs clusters for the whole set of input genomes. We validate the performance of MultiMSOAR by testing the method on the genomes of rat, mouse and human and comparing its predicted main ortholog clusters using gene annotations and functional classification in public databases. We also compare our result to that of MultiParanoid's [1], which is a single-linkage based ortholog clustering approach utilizing the pairwise ortholog clusters obtained by Inparanoid [19].

2. Method

Consider k closely related genomes $G_1, G_2, \ldots, G_k$, where $k \geq 3$. Suppose that these k genomes are ordered according to their given (rooted) species tree in a post-order traversal fashion. For example, genome G_1 and G_2 share a common ancestor denoted as A_{12}, A_{12} and genome G_3 share a common ancestor denoted as A_{123}, so on and so forth, and finally all the genomes share a common ancestor denoted as $A_{12\cdots k}$. That is, the genomes are *phylogenetically ordered*. MultiMSOAR first applies MSOAR on each pair G_i, G_j of the input genomes to obtain a set of putative main ortholog pairs for G_i, G_j. Then it constructs clusters of main orthologs for all the input genomes by combining the pairwise ortholog prediction results by resolving inconsistency and taking into account the species tree and possibilitiy of gene loss.

2.1. *Main ortholog clusters for three genomes*

We first explain the idea of this method for the case of three genomes. Given three phylogenetically ordered genomes G_1, G_2 and G_3, and the sets (or tables) of putative main ortholog pairs $T(G_1, G_2)$, $T(G_1, G_3)$, and $T(G_2, G_3)$ obtained by applying MSOAR to genome pairs G_1 and G_2, G_1 and G_3, and G_2 and G_3, MultiMSOAR starts the construction of ortholog clusters by making every main ortholog pair in these three tables its own cluster. MultiMSOAR next merges clusters using the single linkage technique, *i.e.* two clusters are merged if and only if they share a common (main) orthologous gene. This procedure is repeated until no mergeable clusters exist. This first step is called *cluster initiation*, and the main ortholog clusters generated in this step are called the *initial clusters*. In the following, we will deal with each initial cluster separately.

We can use an undirected connected graph $\mathcal{G}(X, Y, Z)$ to describe the structure of an initial cluster, where X, Y, and Z are three disjoint vertex sets that contain the vertices representing genes from the three genomes involved in the initial cluster. In graph $\mathcal{G}(X, Y, Z)$ (or simply $\mathcal{G}$ for simplicity), the vertices are $X \cup Y \cup Z$ and each edge connects two vertices if they are assigned as a main ortholog pair by MSOAR in the pairwise comparisons, *i.e.* they form an entry in one of the main ortholog pair tables. Since the main orthology is an inter-genome and one-to-one relationship, $\mathcal{G}$ is

a tripartite graph and have four possible topologies, called *triangle*, *2-path*, *3-path*, and *n-path* respectively (see Figure 1). We will process these topologies differently. In the case of a triangle, the corresponding cluster has three orthologous genes, one from each genome, forming exactly three pairs of main orthologs. Such a cluster will be reported as a final main ortholog cluster because of the strong support from the pairwise comparisons. Each 2-path topology describes the scenario that a main ortholog pair was found in two of the genomes, but neither of these two genes have an main ortholog counterpart found in the third genome. This main ortholog pair will also be reported as a final main ortholog cluster. Moreover, if the main ortholog pair was found between G_1 and G_3 (or G_2 and G_3), a gene loss will also be reported in G_2 (or G_1, respectively), since G_1 and G_2 are assumed to have diverged from a more recent speciation. Note that, if the main ortholog pair was found between G_1 and G_2, we will not need report a gene loss event in G_3. A 3-path topology is an acyclic path with three vertices, describing the scenario that two main ortholog pairs were found that involve one gene from each genome and share a common gene. However, none of the remaining two (unshared) genes were found to form main ortholog pairs with any other genes. This 3-path topology indicates a possible main ortholog pair (missing edge) that has been missed by MSOAR due to complications caused by multi-domain proteins or alternative splicing. Therefore, the three genes in this 3-path initial cluster will be reported as a final main ortholog cluster. Some real examples of gene losses and missing main ortholog pairs found by MultiMSOAR will be given in section 3. All other initial clusters have the n-path topology. An n-path could be a path or a cycle, as long as it involves more than three vertices. Such an initial cluster contains more than one gene from some genome, and the handling of such an initial cluster is nontrivial.

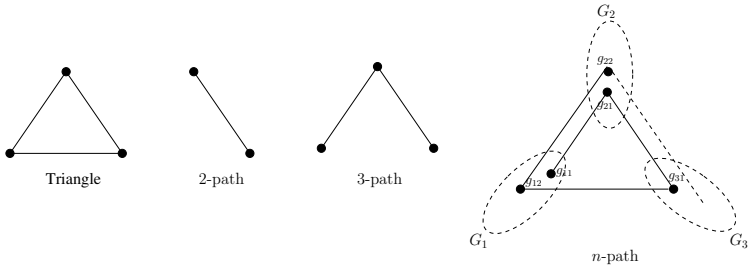

Fig. 1. Four possible topologies of the initial main ortholog clusters.

In practice, the number of initial clusters with the n-path topology should be usually very small if the pairwise comparison results are reliable. For example, the number of such initial clusters is 390 (or 2.64%)involving a total of 2688 (or 5.79%) genes from all three genomes in the rat, mouse and human comparison to be discussed in the next section. For each initial cluster $\mathcal{G}(X, Y, Z)$ with the n-

path topology, MultiMSOAR uses a heuristic algorithm, called NPATHRESOLVER, to divide the initial cluster into final main ortholog clusters, each with three ortholog genes, using a combinatorial optimization approach. This heuristic algorithm transforms $\mathcal{G}(X, Y, Z)$ into a complete weighted tripartite graph $\bar{\mathcal{G}}(\bar{X}, \bar{Y}, \bar{Z}, W)$ by adding dummy vertices and dummy edges (so that a perfect matching always exists), and then tries to find a perfect tripartite matching with the maximum weight. This tripartite matching problem is also called the *maximum three-index assignment* problem, which is known NP-hard [13]. We employ the *single-pass recursive* heuristic proposed by Bandelt *et al.* [3], which could also be applied to the maximum multi-index assignment problem. The heuristic works as follows: (i) Find the maximum weight bipartite matching $M_{\bar{X}\bar{Y}}$ between the vertex sets $\bar{X}$ and $\bar{Y}$. (ii) Let $N = \{n_{xy} | x \in \bar{X}, y \in \bar{Y}, (x, y) \in M_{\bar{X}\bar{Y}}\}$ be a new vertex set, and define the weight between vertices $n_{xy} \in N$ and $z \in Z$ as $W(n_{xy}, z) = W(x, z) + W(y, z)$. (iii) Find a maximum weight bipartite matching between the vertex sets N and Z. Note that, a maximum weight bipartite matching can be computed by the classical Hungarian method [18] in cubic time.

The weights W in $\bar{\mathcal{G}}(\bar{X}, \bar{Y}, \bar{Z}, W)$ are defined taking into account both sequence similarity and the main ortholog pair information from the pairwise comparisons found by MSOAR which are mostly based on gene location information.

$$W(i,j) = \begin{cases} MAXWEIGHT & Evalue(i,j) = 0 \ or (i,j) \in E(\mathcal{G}) \\ -\log(Evalue(i,j)) & 0 < Evalue(i,j) \leq 1e-20 \\ MINWEIGHT & Otherwise \end{cases} \quad (1)$$

Here, $Evalue(i, j)$ is obtained by an all-*versus*-all BLASTp comparison between each pair of genomes. $(i, j) \in E(\mathcal{G})$ indicates that i and j was assigned as a main ortholog pair by the pairwise comparisons using MSOAR. MAXWEIGHT and MINWEIGHT are two constant values, where MAXWEIGHT must be bigger than the biggest value of $-\log(Evalue(i, j))$ and MINWEIGHT must be smaller than the smallest value of $-\log(Evalue(i, j))$.

The algorithm obtains a set of triplets based on the final maximum weight matching. A triplet will be reported as a main ortholog cluster if and only if its three vertices represent real genes. In other words, as long as a triplet contain at least one dummy vertex, all the genes in this triplet will be regarded as inparalogs. The outline of algorithm NPATHRESOLVER is illustrated in Figure 2.

2.2. *Extension to the comparison of more than three genomes*

Now consider the case of $k > 3$ genomes $G_1, G_2, \ldots, G_k$. The initial clusters can be constructed in the same way as in the case of three genomes using the single linkage clustering technique. Here, the graph $\mathcal{G}(V_1, V_2, \ldots, V_k)$ has k disjoint vertex sets, which correspond to the k genomes. Similar to the above, the initial clusters are classified into three possible topologies: the k-clique, a pseudo-clique, and a nontrivial case. A k-clique consists of k genes, one from each genome, that form

Algorithm npathResolver($\mathcal{G}(X, Y, Z)$)

1. Add dummy vertices and edges to obtain a complete weighted tripartite graph $\bar{\mathcal{G}}(\bar{X}, \bar{Y}, \bar{Z})$
2. Define edge weight function W for $\bar{\mathcal{G}}$ according to equation (1)
3. Compute a tripartite matching $M(\bar{X}, \bar{Y}, \bar{Z})$ using the single-pass recursive heuristic
4. **for each** $m \in M(\bar{X}, \bar{Y}, \bar{Z})$
5. **if** m contains no dummy vertices
6. **then** output m as a final main ortholog cluster

Fig. 2. The heuristic algorithm to resolve initial clusters with the n-path topology.

exactly $k(k-1)/2$ main ortholog pairs as found by the pairwise comparisons. This cluster will be reported as a final main ortholog cluster. A pseudo-clique is a graph with $m \leq k$ vertices, with each vertex from a different genome. If the pseudo-clique contains e edges, we use a parameter $q = 2e/m(m-1)$ to measure its *cliqueness* (or edge density). When m and q are greater than some user-defined thresholds, the corresponding initial clusters will be reported as a final main ortholog clusters, and some gene loss events will be reported according to the species phylogeny. In a nontrivial case, the initial cluster contains multiple genes from the same genome. A maximum weight k-partite matching will be used on $\mathcal{G}(V_1, V_2, \ldots, V_k)$ to distinguish main orthologs from inparalogs, similar to the above algorithm npathResolve for three genomes. Note that the single-pass recursive heuristic for finding a maximum weight matching can be extended to $k > 3$ genomes in a straightforward way [3]. Again, this approach will be quite effective since the number of nontrivial cases are expected to be very small.

3. Experimental results

In order to test the performance of MultiMSOAR as a tool of clustering main orthologs shared by multiple genomes, we have applied it to three model genomes: Rat (*Rattus norvegicus*), mouse (*Mus musculus*) and human (*Homo sapiens*). Gene positions, transcripts and translations were downloaded from the UCSC Genome Browser [14] website (http://genome.ucsc.edu). We use the canonical splice variants from the November 2004 update of the rat genome (UCSC rn4, Nov. 2004, version 3.4), the build 36 "essentially finished" assembly of the mouse genome (UCSC mm8, February 2006) and the build 36.1 finished human genome assembly (UCSC hg18, March 2006). There are 7066 protein sequences in the rat genome assembly rn4, 19199 sequences in mouse genome assembly mm8 and 20161 sequences in human genome assembly hg18. The pairwise main ortholog information is obtained by running MSOAR on each pair of the genomes. Specifically, there are 14306 main

ortholog pairs reported between mouse and human, 6539 main ortholog pairs between mouse and rat, and 6347 main ortholog pairs between rat and human. The distributions of the different topologies of initial main ortholog clusters are showed in Figure 3. MultiMSOAR identified 14790 main ortholog clusters in total. We validate the predicted main ortholog clusters using the gene annotation information and function classification in public databases below. We will also compare the result of MultiMSOAR with that of MultiParanoid [1] which is an ortholog clustering method solely based on sequence similarity. The comparative study shows that the prediction result of MultiMSOAR largely agrees with that of MultiParanoid, but about 7.17% of MultiMSOAR's predicted main ortholog clusters properly refine their corresponding MultiParanoid clusters.

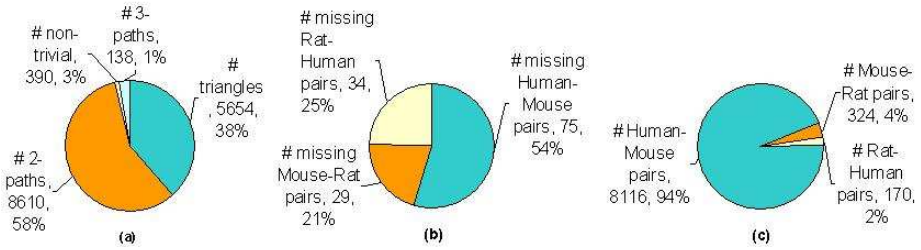

Fig. 3. Some statistics in the comparison of human, mouse,and rat genomes. (a) The distribution of the four topologies of the initial main ortholog clusters. (b) The distribution of the three types of 3-path topologies. (c) The distribution of the three types of 2-path topologies.

3.1. *Validation using gene annotation*

First, we use gene annotation information (in particular, gene symbols or names) to validate the main ortholog clusters found by MultiMSOAR. The hypothesis is that genes with identical symbols are most likely to be main orthologs, since a gene symbol usually conveys the character or function of the gene. We extracted the gene annotation information from UniProtKB/Swiss-Prot [2] Release 52.1. Recall that MultiMSOAR output 14790 main ortholog clusters for rat, mouse and human, among which only 12598 clusters have complete annotations. Out of the 12598 main ortholog clusters, 10605 (84.18%) clusters are true positives (*i.e.* all the genes in the cluster have completely identical gene symbols). Among the 10605 true positives, 6176 clusters have size two and 4429 clusters have size three. Since there are 12455 assignable main ortholog clusters (*i.e.* the total number of clusters of genes with identical symbols), MultiMSOAR achieved a sensitivity of 85.15% for the rat, mouse and human comparison. The detailed results are also summarized in Table 1.

Table 1. Validation of the main ortholog clusters found by MultiMSOAR using gene annotation

	assignable	assigned	unknown	true positive
Main ortholog clusters of size two	7700	8610	1392	6176
Main ortholog clusters of size three	4755	6180	719	4429

3.2. *Validation using gene functions*

Besides gene annotation, we also use gene functional classification to validate our clustering result. PANTHER (Protein Analysis Through Evolutionary Relationships) classification system [25] is an online resource that classifies genes by their functions. It is based on a method that uses published scientific experimental evidence or evolutionary relationship to predict functions in the absence of direct experimental evidence. Proteins that belong to the same functional family and subfamily are assigned the same PANTHER ID. We examine the consistency between the main ortholog clusters output by MultiMSOAR and the PANTHER IDs of the involved genes. Out of the 14297 main ortholog clusters of rat, mouse and human found by MultiMSOAR with valid Entrez gene IDs, 11667 (or 81.6%) clusters consist of genes with the same PANTHER IDs, including 6703 clusters of size two and 4964 clusters of size three. This result demonstrates that the main ortholog clusters obtained by MultiMSOAR are very much in agreement with the gene functional classification provided by PANTHER.

3.3. *Comparison with MultiParanoid*

MultiParanoid is a genome-scale analysis program that clusters orthologs and inparalogs shared by multiple genomes [1]. It is a straightforward extension of the well-known Inparanoid program [19], which identifies orthologs and inparalogs between a pair of genomes solely based on sequence similarity. To ensure a direct comparison between MultiMSOAR and MultiParanoid, we run MultiParanoid on the same dataset (*i.e.* UCSC hg18, UCSC mm8, and UCSC rn4). Since MultiParanoid only reports clusters of co-orthologous genes and it dose not distinguish main orthologs from their inparalogs, the size of a MultiParanoid cluster might exceed three. After comparing with the MultiParanoid clusters, the main ortholog clusters identified by MultiMSOAR are divided into four categories: match, subset, absence, and mismatch. Among the 14790 main ortholog clusters generated by MultiMSOAR for rat, mouse and human, 13109 (or 89.12%) clusters found identical matches in MultiParanoid's output, 1054 (or 7.17%) clusters are contained in the corresponding MultiParanoid clusters as proper subsets, 297 (or 2.02%) clusters are absent in MultiParanoid's output (including those clusters that are proper supersets of some MultiParanoid clusters), and 330 (or 2.59%) clusters are mismatched, *i.e.* each of them partially overlaps with some MultiParanoid cluster. Note that, when a MultiMSOAR cluster C_1 is properly contained in some MultiParanoid cluster C_2, the additional elements in C_2 are likely inparalogs (as identified by MultiM-

SOAR) rather than main orthologs, and thus C_1 could represent a more accurate main ortholog cluster than C_2. In other words, C_1 could be viewed as a refinement of C_2. The distribution of these four types of main ortholog clusters is illustrated in Figure 4. This comparison shows that the main ortholog clusters identified by MultiMSOAR are very consistent with the ortholog clusters generated by MultiParanoid. Furthermore, MultiMSOAR gives more detailed and accurate orthology information since it distinguishes main orthologs from inparalogs.

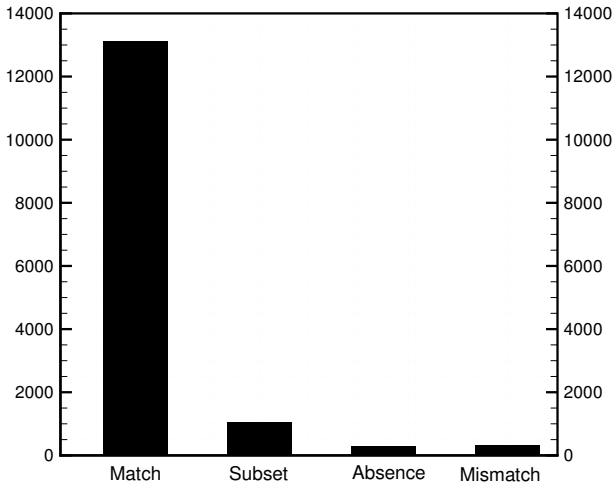

Fig. 4. Comparing the prediction results of MultiMSOAR and MultiParanoid.

3.4. *Examples of identified gene losses and main ortholog pairs missed in pairwise comparisons*

As described above, by taking into account the species tree of the genomes under consideration, MultiMSOAR is able to identify possible gene losses. In the case of rat, mouse and human comparison, if a main ortholog pair was found between mouse and human (or rat and human) without a corresponding orthologous gene found in rat (or mouse, respectively), a gene loss will be reported in rat (or mouse, respectively), since mouse and rat were separated by a more recent speciation. Figure 5 shows a segment of rat chromosome 5 (169,624,099 - 169,349,727), a segment of mouse chromosome 4 (151,234,544 - 150,964,681) and a segment of human chromosome 1 (6,028,567 - 6,407,434). Based on the gene location information and gene sequence similarity, MultiMSOAR successfully identified 9 main ortholog clusters within these chromosome segments and a possible gene loss in rat (*i.e.* chd5).

Besides, a main ortholog pair between human and mouse (*i.e.* ESPN) missed by MSOAR in the pairwise comparisons was identified by MultiMSOAR. This pair of main orthologs was missed by MSOAR because their sequences match different segments of their orthologous gene in rat and thus have insufficient similarity between themselves.

In the rat, mouse and human comparison, a total of 8286 genes were found to have been lost by MultiMSOAR and 138 pairs of main ortholog pairs that were missed by MSOAR in the pairwise comparison were imputed.

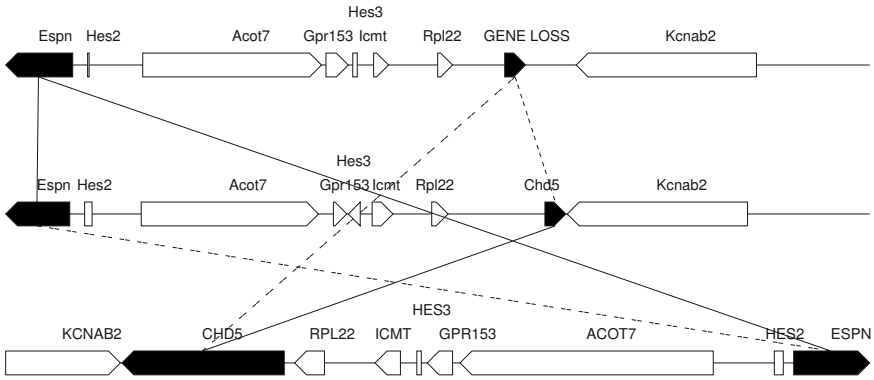

Fig. 5. An example of gene loss and missing main ortholog pairs. In the figure, the rat, mouse and human chromosomal segments are ordered top down. Solid lines indicate main ortholog pairs found by pairwise comparisons. Dashed lines indicate the missing orthology information identified by MultiMSOAR.

4. Concluding remarks

The ortholog clustering method that we presented here extends the pairwise method MSOAR [11] and enables the identification of main ortholog clusters for multiple closely related genomes. Our preliminary experiment on a three genome comparison demonstrates that our method performs consistently with the gene annotation and funcational classification information in public databases and a published program in the literature. Some interesting future work includes more extensive testing on four or more genomes and elaborate (and in-depth) handling of gene losses (*e.g.* using pseudo gene information). We plan to make this a program a public server in the near future.

5. Acknowledgment

This project is supported in part by NSF grant CCR-0309902, National Key Project for Basic Research (973) grant 2002CB512801, NSFC grant 60528001, and a Changjiang Visiting Professorship at Tsinghua University.

References

1. A Alexeyenko, I Tamas, G Liu, and E L L Sonnhammer. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, 22:9–15, 2006.
2. A Bairoch, R Apweiler, C H Wu, W C Barker, B Boeckmann, S Ferro, E Gasteiger, H Huang, R Lopez, M Magrane, and et al. The universal protein resource (uniprot). *Nucleic Acids Res.*, 33:D154–D159, 2005.
3. H Bandelt, Y Crama, and Spieksma F. Approximation algorithms for multi-dimentional assignment problems with decomposable costs. *Discrete Applied Mathematics*, 49:25–50, 1994.
4. J C Chiu, E K Lee, M G Egan, I N Sarkar, G M Coruzzi, and R DeSalle. Orthologid: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics*, 22:699–707, 2006.
5. P Dehal and J Boore. Two rounds of whole genome duplication in the ancestral vertebrate. *PLOS Biology*, 3:1700–1708, 2005.
6. J Dufayard, L Duret, S Penel, M Gouy, F Rechenmann, and G Perriere. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence detabase. *Bioinformatics*, 21:2596–2603, 2005.
7. J A Eisen. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research*, 8:163–167, 1998.
8. J T Eppig, C J Bult, J A Kadin, J E Richardson, J A Blake, A Anagnostopoulos, Baldarelli. R M, M Baya, J S Beal, S M Bello, and et al. The mouse genome database (mgd): from genes to mice: a community resource for mouse biology. *Nucleic Acids Res.*, 33:D471–D475, 2005.
9. T A Eyre, M W Wright, M J Lush, and Bruford E A. Hcop: a searchable database of human orthology predictions. *Brief Bioinform.*, 2006.
10. W M Fitch. Distinguishing homologous from analogous proteins. *Syst. Zool.*, 19:99–113, 1970.
11. Z Fu, X Chen, V Vacic, P Nan, Y Zhong, and T Jiang. A parsimony approach to genome-wide ortholog assignment. In *Proceedings of 10th Annual International Conference, RECOMB (Venice, Italy, April 2006)*, pages 578–594, 2006.
12. L Goodstadt and C Ponting. Phylogenetic reconstruction of orthology, paralogy and conserved synteny for dog and human. *PLOS Biology*, 1:e45, 2003.
13. V Kann. Maximum bounded 3-dimensional matching is max snp-complete. *Inform. Process. Lett.*, 37:27–35, 1991.
14. D Karolchik, R Baertsch, M Diekhans, T S Furey, A Hinrichs, YT Lu, K M Roskin, Schwartz M, C W Sugnet, D J Thomas, and et al. The ucsc genome browser database. *Nucleic Acids Res.*, 31:51–54, 2003.
15. Y Lee, R Sultana, G Pertea, J Cho, S Karamycheva, J Tsai, B Parvizi, F Cheung, V Antonescu, J White, and et al. Cross-referencing eukaryotic genomes: Tigr orthologous gene alignments (toga). *Genome Research*, 12:493–502, 2002.
16. H Li, A Coghlan, J Ruan, L J Coin, J Heriche, L Osmotherly, R Li, T Liu, Z Zhang, L Bolund, and et al. Treefam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acide Res.*, 34:D572–D580, 2006.

17. L Li, C Stoeckert, and D Roos. Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13:2178–2189, 2003.
18. C H Papadimitriou and K Steiglitz. *Combinatorial optimization: algorithms and complexity*. 2004.
19. M Remm, C E Storm, and E L Sonnhammer. Automatic clustering of orthologs and inparalogs from pairwise species comparisons. *Journal of Molecular Biology*, 314:1041–1052, 2001.
20. D Sankoff. Genome rearrangement with gene families. *Bioinformatics*, 15:909–917, 1999.
21. C E Storm and E L Sonnhammer. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, 18:92–99, 2002.
22. R L Tatusov, M Y Galperin, D A Natale, and E V Koonin. The cog database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, 28:33–36, 2000.
23. R L Tatusov, E Koonin, , and D J Lipman. A genomic perspective on protein families. *Science*, 278:631–637, 1997.
24. R L Tatusov, D A Natale, J D Jackson, A R Jacobs, B Kiryutin, E V Koonin, D M Krylov, R Mazumder, S L Mekhedov, A N Nikolskaya, and et al. The cog database: an update version includes eukaryotes. *BMC Bioinformatics*, 4:41–54, 2003.
25. P D Thomas, M J Campbell, A Kejariwal, H Mi, B Karlak, R Daverman, K Diemer, A Muruganujan, and A Narechania. Panther: a library of protein families and subfamilies indexed by function. *Genome Research*, 13:2129–2141, 2003.
26. C M Zmasek and S R Eddy. Rio: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, 3:14, 2002.