Verifying Orthologous Paralogenes using Whole Genome Alignment

P.Y. Chan T.W. Lam S.M. Yiu

Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong

Email: {pychan, twlam, smyiu}@cs.hku.hk

ABSTRACT: Identifying orthologous paralogenes is a fundamental problem in comparative genomics and can facilitate the study of evolutionary history of the species. Existing approaches for locating paralogs make use of local alignment based algorithms such as BLAST. However, there are cases that genes with high alignment scores are not paralogenes. On the other hand, whole genome alignment tools are designed to locate orthologs. Most of these tools are based on some unique substrings (called *anchors*) in the corresponding orthologous pair to identify them. Intuitively, these tools may not be useful in identifying orthologous paralogenes as paralogenes are very similar and there may not be enough unique anchors.

However, our study shows that this is not true. Paralogenes although are similar, they have undergone different mutations. So, there are enough unique anchors for identifying them. Our contributions include the followings.

- Based on this counter-intuitive finding, we propose to employ the whole genome alignment tools to help verifying paralogenes. Our experiments on five pairs of human-mouse chromosomes show that our approach is effective and can identify most of the mis-classified paralog groups (more than 80%).
- We verify our finding that whole genome alignment tools are able to locate orthologous paralogenes through a simulation study. The result from the study confirms our finding.

1 INTRODUCTION

Paralogs [6] are genes that were duplicated from a single gene on the same genome. Orthologs are genes in different species that evolved from the same gene through speciation. Locating paralogs and orthologs is a fundamental problem in comparative genomes. This information can be used to study the evolutionary history of the species.

For finding individual orthologous genes (i.e., orthologous genes without duplication in either genome), there are quite a number of different approaches [3, 7, 11, 12]. On the other hand, for identifying paralogs, we still rely on the local alignment based methods such as BLAST [1, 8]. Although the BLAST score (local alignment score) is a good indicator whether the genes are paralogs, there are cases that genes having high BLAST alignment scores turn out to be not paralogs. For example, in Human Chromosome X, the genes NUDT10 and NUDT11 were believed to be paralogs[1] and they have a high BLAST alignment score, however, these two genes are found to be not paralogs[2].

In fact, we noticed that the groupings in NCBI are not yet stable. For example, from the two versions (Jan and May) of the NCBI database, there are 10 groups out of 16 groups that involve paralogs in either human chromosome X or mouse chromosome X are found to be misclassified. So, it is desirable to have a better method to verify whether the genes in the same paralog groups are indeed paralogs.

Whole genome alignment tools [2, 4, 5, 13] are also designed to locate individual orthologous genes. Most of these tools are based on some *unique* and highly similar short substrings (called *anchors*) between the orthologous pairs to identify them. Intuitively, it is difficult for these tools to locate the orthologs that involve some paralogs since paralogs are very similar and there may not be enough anchors (due to the uniqueness requirement). However, we observed that this is not true. Although paralogenes are very similar, they have gone through different mutations separately. As a result, enough anchors still exist for the whole genome alignment tools to identify the orthologs involving paralogs. In other words, if there are two paralogenes that are conserved to the same gene in another species, existing whole genome alignment tools should be able to locate both pairs of orthologs. Our contributions include the followings.

- Based on this counter-intuitive finding, we propose a verification process that makes use of whole genome alignment tools to verify orthologous paralogs. We have tested our approach on five human-mouse chromosome pairs. The experimental results show that our method is effective and most of the mis-classified paralog groups can be identified. In particular, our approach is able to identify more than 80% (sometimes more than 90%) of these misclassified groups. On the other hand, the paralog groups that can go through our verification process are likely to be real paralogs. We check the verification results based two versions of the NCBI databases.
- To confirm that whole genome alignment tools are able

[1] In the January version of the NCBI database.

[2] In the updated version (May) of the NCBI database, http://www.ncbi.nlm.nih.gov/HomoloGene.

to locate orthologous paralogs, we have conducted a simulation study. The result from the study is consistent with our finding and more than 80% of the simulated orthologous paralogs are successfully located by the whole genome alignment tools.

The organization of the rest of the paper is as follows. In Section 2, we describe the verification process and the experimental results of our verification process based on five pairs of human-mouse chromosomes. The simulation study to confirm that whole genome alignment tools can be used to locate orthologous paralogs is presented in Section 3. Section 4 concludes the paper.

2 THE VERIFICATION PROCESS and EXPERIMENTAL RESULTS

2.1 The Verification Procedure

Before we describe our verification process, we first assume that whole genome alignment tools are able to locate orthogous paralogs. More precisely, if genes A_1 and A_2 is a paralogous gene pair in Genome G_1 and both are orthologs to the same conserved gene B in Genome G_2, then there is a high chance that both orthologous pairs (A_1,B) and (A_2,B) are reported by the whole genome alignment tools. This assumption will be verified by a simulation study in Section 3.

In our verification process, we only focus on the orthologous paralog groups that contain paralogenes which are highly similar based on the score from BLASTn. More precisely, if the local alignment scores among the genes are lower than a threshold (in our experiment, we set the threshold to be 200), it is already very likely that these genes are not paralogs, so we will not further process these genes using our method.

For gene pairs with high BLASTn alignment scores that are in the same paralog groups, we conduct a whole genome alignment between the chromosomes of the two corresponding species. For any gene in the paralog group, if there is no corresponding orthologous pair reported by the whole genome alignment tool or there is exactly one orthologous pair is reported, then we predict that this gene should not belong to this paralog group and thus is misclassified. And if there are more than half of the genes in the group that are misclassified, we assume that this group is misclassified. In fact, based on our experiments, if a group is misclassified, almost all the genes are misclassified. Otherwise, we assume that the group is a correct paralog groups.

2.2 Experimental Results

To evaluate our verification approach, we have conducted some experiments on five human-mouse chromosome pairs. These human-mouse Chromosme pairs are Human Chromosome X vs mouse chromosome X, Human Chromosome 1 vs Mouse Chromosome 4, Human Chromosome 1 vs Mouse Chromosome 3, Human Chromosome 17 vs Mouse Chromosome 11 and Human Chromosome 11 vs Mouse Chromosome 7. These five pairs are chosen because they contain the highest number of paralog groups among all the possible chromosome pairs.

We use two whole genome alignment software tools, MSS [2, 9] and MUMmer [10] in our experiments. In fact, both of them show a very similar performance in the experiments. After obtaining the output from the whole genome alignment tools, we focus on the homology groups of the chromosome pairs that involve paralogs with high BLASTn alignment scores. We compare two versions of the NCBI databases (one obtained in January and the other one obtained in May). We count the number of paralog groups that are misclassified (that is, the groups, which were classified as paralogs in January version, but turn out to be not paralogs in the May version), and the number of groups that exist in both versions (we assume that these groups are likely to be correct). Then, see how many of them are correctly predicted by our verification process.

Table 1 shows the experimental results. From the results, both MSS and MUMmer successfully identify most of the cases. In particular, in all five cases, all the correct paralog groups can pass our verification process. On the other hand, except one case (using MUMmer for human chromosome 4 and mouse chromosome 1), our verification procedure can detect more than 80% of those misclassified groups. From the experiments, it seems that both software tools have a similar performance. To conclude, the verification process is shown to be effective.

3 THE SIMULATION STUDY

Recall that most of the whole genome alignment tools are based on some *unique* and highly similar anchors between the corresponding orthologous pair to locate them. Intuitively, if one of the genes in the ortholgous pair has a corresponding paralogene in the genome, then it may not be possible to have enough anchors since paralogenes are usually very similar so the uniqueness requirement for anchors may easily be violated. In this section, we will show that this is not true by a simulation study. The reason for this counter-intuitive finding is that although paralogenes are similar, they have gone through a lot of mutations separately. As a result, we still can find enough anchors for the tools to identify their locations. In fact, we have illustrated how to make use of this counter-intuitive finding to verify orthologous paralogenes in Section 2.

The simulation study is conducted as follows. Based on a set of real paralog groups, we have obtained distributions for the following parameters: (a) length of a gene; (b) the similarity of a gene and its paralogene; (c) the similarity of a gene and its ortholog. For (b), given a gene A, when we generate its paralog, we select regions from A. The number of regions

(a) Human Chromosome X vs Mouse Chromosome X

	NCBI (Jan vs May)	Correctly Predicted by MSS	Correctly Predicted by MUMmer
No. of *correct* paralog groups	4	4	4
No. of *misclassified* paralog groups	8	7	7

(b) Human Chromosome 1 vs Mouse Chromosome 4

	NCBI (Jan vs May)	Correctly Predicted by MSS	Correctly Predicted by MUMmer
No. of *correct* paralog groups	5	5	5
No. of *misclassified* paralog groups	7	6	5

(c) Human Chromosome 1 vs Mouse Chromosome 3

	NCBI (Jan vs May)	Correctly Predicted by MSS	Correctly Predicted by MUMmer
No. of *correct* paralog groups	3	3	3
No. of *misclassified* paralog groups	10	8	9

(d) Human Chromosome 17 vs Mouse Chromosome 11

	NCBI (Jan vs May)	Correctly Predicted by MSS	Correctly Predicted by MUMmer
No. of *correct* paralog groups	4	4	4
No. of *misclassified* paralog groups	15	14	13

(e) Human Chromosome 11 vs Mouse Chromosome 7

	NCBI (Jan vs May)	Correctly Predicted by MSS	Correctly Predicted by MUMmer
No. of *correct* paralog groups	3	3	3
No. of *misclassified* paralog groups	16	14	13

Table 1: Evaluation of the Verification Process

and the length of each region are selected according to a distribution. These regions will be duplicated with mutations in its paralog, then gaps between these regions are generated also according to a distribution to form the paralogene. We allow two types of mutations: The single nucleotide polymorphism (SNP), that is, only one nucleotide is modified (insertion, deletion, or substition), and the block mutation. The number of SNP, the number of block mutations together with the lengths of the blocks are generated according to a distribution. A similar generation procedure is applied to (c) with a different set of distributions. The following shows the steps for generating a pair of paralogenes and the corresponding ortholog in a pair of genomes.

Step 1: Generate a random sequence as a gene with length following a distribution, we call it Gene A.

Step 2: Based on A, we generate its paralogene, call it Gene P. The similarity of A and P follows some distribution based on the information we obtain from the set of real data.

Step 3: Similarily, we generate the orthologene for A, call it Gene O.

Step 4: The pair of paralogenes is placed in one of the genomes. The locations of the genes are randomly generated. The corresponding orthologene is in the other genome with a randomly generated position. Gaps are randomly generated to fill the whole genomes. The lengths of the genomes are given by the user.

We can increase the number of orthologous paralog groups in the genomes and the number of genes in each group by repeating Steps 2 and 3 accordingly. The locations of all the genes are randomly generated. Based on the above generation process, we have generated nine pairs of genomes. Table 2 shows the details of the simulated data. Note that most of the paralog groups contain 2 paralogenes and 1 orthologene, some groups may contain more paralogenes and orthologenes. This matches the statistics we obtained from the real data. Also, the similarity of the genes follow the information we obtained from the real data set.

The generated genomes are then submitted to the whole genome alignment software tool, MSS, for verification. Note that the simulated genomes are generated in DNA level. MSS processes them in protein level. The percentage of the paralog groups identified in each case is presented in Table 3. When we say that a group is identified, it means that MSS can report all possible orthologous pairs in the whole group. From the table, we can see that that more than 80% of the paralog groups are successfully identified. For the missing groups, we have checked the genes, we found that the reason for not able to identify them is due to the low similarity of the ortholo-

Exp. No.	Approx. Len. of Seq.	No. of paralog gp.
1	1M	7
2	1M	14
3	5M	7
4	5M	14
5	1M	3
6	1M	7
7	1M	7
8	1M	7
9	1M	7

Table 2: Details of simulation data

gous pairs, but not because of the existence of paralogenes. So, from this set of preliminary experiments, we can see that whole genome alignment tools should be able to locate orthologous paralog groups.

Exp. No.	No. of Paralog gp.	No. of Paralog gp. Identified
1	7	6
2	14	12
3	7	4
4	14	11
5	3	3
6	7	5
7	7	6
8	7	6
9	7	5

Table 3: Results of Simulation

4 Conclusion

In this paper, we propose a verification process that makes use of the whole genome alignment to verify orthologous paralogs identified by local alignment based software tools such as BLAST. Based on the experimental results of five human-mouse chromosome pairs, we show that our approach is effective and can identify more than 80% of the misclassified paralog groups. The rationale behind our verification process is based on a counter-intuitive finding on whole genome alignment tools that they can be used to identify orthologous pairs in which there are paralogenes. This counter-intuitive finding is further verified by a simulation study. For future work, a more detailed simulation study and more experiments on real data will be performed to confirm our finding. On the other hand, how to locate orthologs with low similarity with computational tools is still a challenging problem.

REFERENCES

[1] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[2] HL Chan, TW Lam, WK Sung, Prudence WH Wong, SM Yiu, and X Fan. A mutation subsequence problem and locating conserved genes. *Bioinformatics*, 21(10):2271–2278, 2005. A preliminary version appears in Proceedings of the IEEE Fourth Symposium on Bioinformatics and Bioengineering (BIBE 2004), p.545-552, 2004.

[3] X. Chen, J. Zheng, Z. Fu, P. Nan, Y. Zhong, S. Lonardi, and T. Jiang. Computing the assignment of orthologous genes via genome rearrangement. In *Proceedings of the 3rd Asia-Pacific Bioinformatics Conference (APBC 2005)*, pages 363–378, 2005.

[4] A.L. Delcher, S. Kasif, R. D. Fleischmann, J. Peterson, O. White, and S.L. Salzberg. Alignment of whole genomes. *Nucleic Acids Research*, 27(11):2369–2376, 1999.

[5] A.L. Delcher, A. Phillippy, J. Carlton, and S.L. Salzberg. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research*, 30(11):2478–2483, 2002.

[6] W.M. Fich. Distinguishing homologous from analogous proteins. *Systematic Zoology*, 19:99–113, 1970.

[7] M. Remm, C. Storm, and E. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, 314:1041–1052, 2001.

[8] The BLAST Web Site. http://www.ncbi.nlm.nih.gov/blast/.

[9] The MSS Web Site. http://www.cs.hku.hk/~mss.

[10] The MUMMER Web Site. http://www.tigr.org/software/mummer/.

[11] 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 Research*, 28:33–36, 2000.

[12] R.L. Tatusov, E.V. Koonin, and D.J. Lipman. A genomic perspective on protein families. *Science*, 278:631–637, 1997.

[13] Prudence W.H. Wong, T. W. Lam, N. Lu, H. F. Ting, and S. M. Yiu. An efficient algorithm for optimizing whole genome alignment with noise. *Bioinformatics*, 20(16):2676–2684, 2004.