Searching biologically plausible synteny blocks among multiple genomes

Tsuyoshi Hachiya[1] **Yasubumi Sakakibara**[1]
[1] *Department of Biosciences and Informatics, Keio University, Japan*
Email: hacchy@dna.bio.keio.ac.jp, yasu@bio.keio.ac.jp

ABSTRACT: In comparative genome analyses, *synteny blocks* play important roles for finding ortholog genes, reconstructing phylogenetic tree and predicting genome rearrangement events. In this paper, we propose a novel method to search biologically plausible synteny blocks not only from the viewpoint of finding highly preserved regions but also from the viewpoint of analyzing genome rearrangements. We have applied the method to our experiments on four fungal organisms, and succeeded to obtain some biologically interesting results.

1 INTRODUCTION

Whole genomes for more than two hundred species have been sequenced, and more sequencing data are rapidly generated. Comparative genomics provides effective tools for genome analyses, and an important feature of comparative genome analyses is that it allows not only *micro-scopic* analyses such as comparison of gene sequences, but also *macro-scopic* analyses such as prediction of genome rearrangement events. *Synteny block* is usually defined as a highly preserved region between two or multiple genomes. Further, synteny block is considered as a unit of genome rearrangement and it may play a role of a bridge between microscopic analyses and macroscopic analyses.

Mauve [3] and GRIMM-Synteny algorithm [8] have been developed to estimate synteny blocks mostly based on sequence similarity analysis. Although the GRIMM-Synteny algorithm is a typical method to calculate synteny blocks, one crucial problem to use GRIMM-Synteny algorithm is that it requires determining the values for two parameters, gap size G and cluster size C, because those parameters significantly affect the results of synteny blocks. In this work, we propose a novel method to determine biologically plausible values for these two parameters and to search synteny blocks which are plausible for both micro-scopic view and macro-scopic view by making use of GRIMM-Synteny algorithm which is based on sequence similarity analysis and MGR algorithm [2] which is based on genome rearrangement analysis. Our fundamental strategy to search biologically plausible values for G and C is based on two ideas: the first idea is that plausible "genome" phylogenetic tree which is based on genome rearrangement analysis must be *consistent with* "molecular" phylogenetic tree which is based on sequence similarity analysis, and the second idea is that plausible "genome" phylogenetic tree should not be largely affected by change of the two parameters if the change is within a fixed range.

We have applied our algorithm to four fungal organisms: *S. cerevisiae*, *A. gossypii*, *S. pombe* and *A. oryzae*. Our algorithm has found that plausible gap size $\hat{G}$ is equal to 200,000 bp and plausible cluster size $\hat{C}$ is equal to 5,210 bp, and 33 plausible synteny blocks have been obtained. 32 synteny blocks among the 33 synteny blocks include only one gene or only one functional sequence, and the remaining one synteny block includes two genes, *tub1* and *tub3*. This result indicates that it is essential for fungal organisms to maintain functional cluster of *tub1* and *tub3* while other genes are fully shuffled by numerous rearrangements.

2 METHODS

2.1 GRIMM-Synteny and MGR algorithm

Before going into the details of our method, we briefly summarize the GRIMM-Synteny algorithm and Multiple Genome Rearrangement (MGR) algorithm.

The input of GRIMM-Synteny algorithm is a set of bidirectional local alignments (also called anchors) among multiple genomes. First, the algorithm finds close anchors whose distance is smaller than gap size G and joins those anchors to be a cluster. Second, it removes small clusters whose size is smaller than cluster size C. Finally, it outputs remaining clusters which have not been removed as synteny blocks.

Locus information of synteny blocks can be converted into an order of synteny blocks on multiple genomes, and the order information is the input of MGR algorithm. The algorithm is extension of duality theorem for genomic distance problem [7] which calculates a distance between two genomes based on an order of synteny blocks, and predicts genome rearrangement events among multiple genomes. The algorithm simultaneously reconstructs a phylogenetic tree based on genome rearrangement analysis.

2.2 Definition of *reliable*, *robust* and *plausible*

Although, the GRIMM-Synteny algorithm is a typical method to calculate synteny blocks, one crucial problem to use GRIMM-Synteny algorithm is to determine the values for two parameters, gap size G and cluster size C. Now, we propose a novel method which determines the two parameters which are plausible for both sequence similarity analysis (micro-scopic view) and genome rearrangement analysis (macro-scopic view). Our fundamental strategy to search biologically plausible values for the two parameters is based on two ideas: the first idea is that plausible "genome" phylogenetic tree must be *consistent with* "molecular" phylogenetic tree, and the second idea is that plausible "genome" phylogenetic tree should not be largely affected by change of the two parameters if the change is within a fixed range.

Then we define two values for G and C are *reliable* if a "genome" phylogenetic tree which is reconstructed by MGR algorithm based on the two values is similar to "molecular" phylogenetic tree. We also define two values for G and C are *robust* if the "genome" phylogenetic tree is not largely affected by change of the two parameters. Further, we intuitively define two values for G and C are *plausible* if the two values are both reliable and robust.

2.3 Distance between two phylogenetic trees

In order to effectively evaluate whether two parameters, gap size G and cluster size C, are plausible or not, we introduce a formal measure to calculate a distance between two phylogenetic trees. In general, it is a hard task to define and calculate a distance between two phylogenetic trees of different topologies and even of a same topology.

Then we compare *"genome" distance matrix* with *"molecular" distance matrix* in spite of comparing "genome" phylogenetic tree with "molecular" phylogenetic tree directly. *Distance matrix* consists of pairwise distances among multiple genomes, and "genome" distance matrix and "molecular" distance matrix are distance matrices which are calculated in the process of reconstructing "genome" phylogenetic tree and "molecular" phylogenetic tree, respectively. An example of distance matrix is shown in Table 1.

	genome1	genome2	genome3
genome1	0	d_{12}	d_{13}
genome2	$d_{21}(=d_{12})$	0	d_{23}
genome3	$d_{31}(=d_{13})$	$d_{32}(=d_{23})$	0

Table 1: Example of distance matrix.

Diagonal elements of distance matrix are always zero and distance between genome i and genome j (d_{ij}) is equal to distance between genome j and genome i (d_{ji}) in both "genome" distance matrix and "molecular" distance matrix. Therefore we converted information of distance matrix into a *distance vector* $\mathbf{d}$ defined as $\mathbf{d} = (d_{12}, d_{13}, \cdots, d_{ij}, \cdots, d_{(N-1)N})$, where $i < j$ and N is number of genomes.

One of standard measures to evaluate a similarity (correlation coefficient) is Pearson correlation coefficient. Correlation coefficient $r_{\mathbf{d}^1\mathbf{d}^2}$ between two feature vectors, $\mathbf{d}^1$ and $\mathbf{d}^2$, is defined as follows:

$$r_{\mathbf{d}^1\mathbf{d}^2} = \frac{(\mathbf{d}^1 - \overline{\mathbf{d}^1}) \cdot (\mathbf{d}^2 - \overline{\mathbf{d}^2})}{\|\mathbf{d}^1 - \overline{\mathbf{d}^1}\| \times \|\mathbf{d}^2 - \overline{\mathbf{d}^2}\|} \qquad (1)$$

Then we define distance between two phylogenetic trees as Pearson correlation coefficient r. The correlation coefficient is always between -1 and $+1$ and correlation coefficients which are close to $+1$ indicate a strong positive correlation and which are close to -1 indicates a strong negative correlation. Values close to 0 indicate a weak correlation, with 0 itself indicating no correlation at all.

2.4 Search plausible parameters

We have intuitively defined two values for two parameters, gap size G and cluster size C, are plausible if the two values are both reliable and robust. Now, we propose a novel method to determine plausible values for the two parameters.

First, our method calculates "genome" distance matrix for any two parameters, G_i and C_i, by solving GRIMM-Synteny algorithm and MGR problem. We define a vector of this "genome" distance matrix which is based on genome rearrangement analysis as $\mathbf{d}^g(G_i, C_i)$. Second, CLASTAL W which is based on progressive alignment methods calculates "molecular" distance matrix by comparing orthologous genes, such as small ribosomal RNA sequences. We define a vector of this "molecular" distance matrix as $\mathbf{d}^m$. Third, our method calculates correlation coefficient $r(\mathbf{d}^g(G_i, C_i), \mathbf{d}^m)$ between two vectors, $\mathbf{d}^g(G_i, C_i)$ and $\mathbf{d}^m$. We define $r(\mathbf{d}^g(G_i, C_i), \mathbf{d}^m)$ as an indicator of reliability and call it $reliability(G_i, C_i)$. The larger $reliability(G_i, C_i)$ becomes, the more reliable two parameters are. Fourth, our method calculates correlation coefficients $r(\mathbf{d}^g(G_i, C_i), \mathbf{d}^g(G_i + \Delta G, C_i))$, $r(\mathbf{d}^g(G_i, C_i), \mathbf{d}^g(G_i - \Delta G, C_i))$, $r(\mathbf{d}^g(G_i, C_i), \mathbf{d}^g(G_i, C_i + \Delta C))$ and $r(\mathbf{d}^g(G_i, C_i), \mathbf{d}^g(G_i, C_i - \Delta C))$ which are correlation coefficients between original genome phylogenetic tree and phylogenetic trees when G or C is changed within a fixed range. We define an average of those correlation coefficients as an indicator of robustness and call it $robustness(G_i, C_i)$. The larger $robustness(G_i, C_i)$ becomes, the more robust two parameters are. Fifth, we define plausibility $plu(G_i, C_i)$ for the two parameters as weighted sum of $reliability(G_i, C_i)$ and $robustness(G_i, C_i)$ such as follows:

$$plu(G_i, C_i) = w_{re} \cdot reliability(G_i, C_i) + w_{ro} \cdot robustness(G_i, Ci) \qquad (2)$$

, where w_{re} and w_{ro} are weight of reliability and robustness, respectively. Finally, we search plausible parameters, $\hat{G}$ and $\hat{C}$, which maximize $plu(\hat{G}, \hat{C})$ among all region of gap size G and cluster size C.

3 RESULTS

3.1 Organisms

All organisms are divided into two groups: the first is prokaryote and the second is eukaryote. Further, eukaryote includes four kingdoms: animal kingdom, plant kingdom, fungi kingdom and protist kingdom.

Fungi kingdom is an interesting kingdom because organisms in fungi kingdom have both aspects of higher eukaryote and prokaryote. For example, some organisms in fungi kingdom are unicellular like prokaryote and some organisms in fungi kingdom are multicellular like higher eukaryote. Further, the ratio of genes which have introns in some organisms in fungi kingdom is exceedingly low (less than 10 %) like prokaryote and the ratio of genes which have introns in some organisms in fungi kingdom is extremely high (higher than 80 %) like higher eukaryote.

We applied our method to the analyses of four fungal organisms: *S. cerevisiae*, *A. gossypii*, *S. pombe* and *A. oryzae*. Features of these four organisms are shown in Table 2.

organisms	unicellular or multicellular	introns
S. cerevisiae	unicelluar	few (5.3 %)
A. gossypii	multicelluar	few (4.6 %)
S. pombe	multicellular	many (45.9 %)
A. oryzae	multicellular	many (80.8 %)

Table 2: Feature of four fungal organisms.

S. cerevisiae is a well-known organism whose popular name is budding yeast. *S. cerevisiae* has been studied well as a simple model of eukaryote and whole genome sequence of *S. cerevisiae* was sequenced in 1997 for the first time as a eukaryote organisms. *A. gossypii* has been studied as a homologous organism of *S. cerevisiae* and whole genome sequence of *A. gossypii* was sequenced in 2004. Less than 10 % of genes have introns in these two organisms and therefore DNA processing mechanism of these two organisms is closer to prokaryote than higher eukaryote.

On the other hand, *S. pombe* whose genome was sequenced in 2002 and *A. oryzae* whose genome was sequenced in 2004 both are closer to higher eukaryote than prokaryote in the sense that about 50 % of *S. pombe* genes and more that 80 % of *A. oryzae* genes have introns.

3.2 Ribosomal RNA distance matrix

Our method requires "molecular" distance matrix to evaluate plausibility for two parameters, gap size G and cluster size C. In this work, we construct "molecular" distance matrix by aligning four 25S ribosomal rRNA sequences which are orthologous among *S. cerevisiae*, *A. gossypii*, *S. pombe* and *A. oryzae*. CLUSTAL W can calculate similarity scores between any pair of four rRNA sequences and therefore we define distances between two rRNA sequences as the reciprocal number of similarity score between two sequences. According to this definition, a "molecular" distance matrix which is based on rRNA sequence analysis is obtained and is shown is Table 3.

	S. cerevisiae	*A. gossypii*	*S. pombe*	*A. oryzae*
Sce	0	0.0484	0.0774	0.2354
Ago	0.0484	0	0.0806	0.2643
Spo	0.0774	0.0806	0	0.2940
Aor	0.2354	0.2643	0.2940	0

Table 3: "Molecular" distance matrix based on rRNA sequence analysis.

CLUSTAL W also reconstructs unrooted phylogenetic tree by solving progressive alignment methods. The phylogenetic tree reconstructed by CLUSTAL W is shown in Figure 1.

3.3 Plausible parameters

We so far have constructed "molecular" distance matrix which is based on rRNA sequence analysis. Now we evaluate plausibility $Plu(G,C)$ for two parameters, gap size G and cluster size C. Our method calculates plausibility for all regions of two parameters. The result of calculating plausibility is shown in Figure 2.

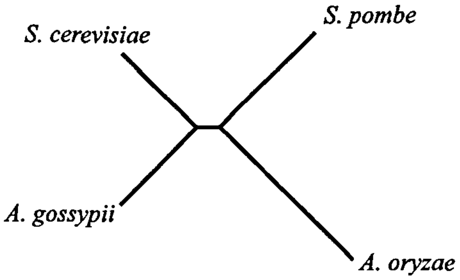

Figure 1: "Molecular" phylogenetic tree based on rRNA sequence analysis.

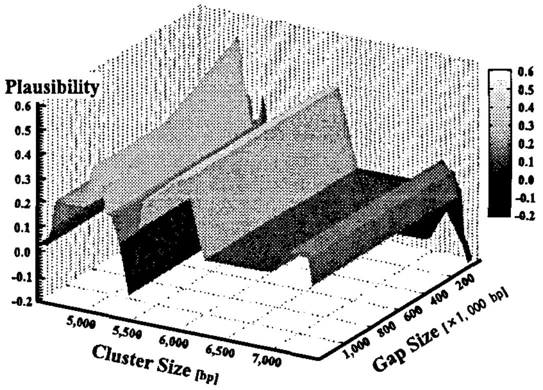

Figure 2: Change of plausibility for two parameters.

As a result of searching all regions of two parameters, our algorithm has found that plausible gap size $\hat{G}$ is equal to $200,000$ bp and plausible cluster size $\hat{C}$ is equal to $5,210$ bp. These two plausible values for two parameters are applied to GRIMM-Synteny algorithm and 33 synteny blocks are predicted. Then MGR algorithm calculates plausible "genome" distance matrix which is shown in Table 4 and reconstructs "genome" phylogenetic tree which is shown in Figure 3.

	S. cerevisiae	*A. gossypii*	*S. pombe*	*A. oryzae*
Sce	0	9	14	13
Ago	9	0	14	14
Spo	14	14	0	15
Aor	13	14	15	0

Table 4: "Genome" distance matrix based on rearrangement analysis.

The length of each edge in Figure 3 represents the number of genome rearrangement events (the rearrangement events are inversion, fission, fusion and translocation) between two genomes connected by the edge. According to Figure 3, a total of 54 genome rearrangements have occurred since the divergence of four fungal organisms. We have also counted

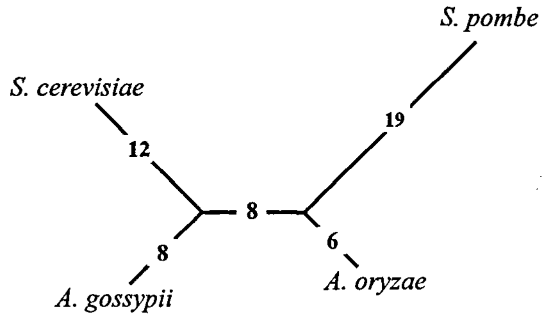

Figure 3: "Genome" phylogenetic tree based on rearrangement analysis.

the number of each event: inversion, fission, fusion or translocation, which are common events of genome rearrangement and numbers and ratio of these events are shown in Table 5. Table 5 indicates that translocation is most common genome rearrangement event and fusion hardly occurs among fungi organisms.

rearrangement type	# of events	ratio of events
inversion	13	24.1 %
fusion	8	14.8 %
fission	2	3.7 %
translocation	31	57.4 %

Table 5: Number and ratio of each event used in genome rearrangements.

3.4 Gene functions in plausible synteny blocks

So far, we have obtained 33 plausible synteny blocks. Detailed annotations of CDS regions and functional sequence regions on *S. cerevisiae* genome make it possible to know what genes or what functional sequences are included in synteny blocks. Then we propose a hypothesis that genes which have some specific functions are more likely to be included in synteny blocks than other genes which don't have specific functions and we test whether our hypothesis is true or not by χ^2 test.

To test our hypothesis, we count number of:

- Genes which are included in synteny blocks and have a specific function (group A)
- Genes which are not included in synteny blocks and have a specific function (group B)
- Genes which are included in synteny blocks and don't have a specific function (group C)
- Genes which are not included in synteny blocks and don't have a specific function (group D)

for all functions which are registered in MIPS functional categories. χ^2 test finds significance when the ratio of group A to group B is significantly higher than the ratio of group C to group D.

We define significance level p as $p = 0.01$ and test our hypothesis by χ^2 test. χ^2 test reveals that function which is likely to be included in synteny blocks is only CELL CYCLE AND DNA PROCESSING and that other functions are not significantly likely to be included in synteny blocks. Important results of χ^2 text are shown in Table 6.

Functional Categories	Significance
METABOLISM	n. s.
ENERGY	n. s.
CELL CYCLE AND DNA PROCESSING	$p < 0.001$
TRANSCRIPTION	n. s.
PROTEIN SYNTHESIS	n. s.
BIOGENESIS OF CELLULAR COMPONENTS	n. s.

Table 6: Whether each function is likely to be included in synteny blocks, or not.

Synteny blocks are calculated based on not only genome rearrangement analysis but also sequence similarity analysis and therefore genes or sequences which are included in synteny blocks are highly conserved. This indicates that genes which are involved in CELL CYCLE AND DNA PROCESSING are more likely to be conserved than other genes which are not involved in the mechanism. Therefore it is concluded that CELL CYCLE AND DNA PROCESSING mechanism is less easy to be changed than other mechanism in fungal evolutionary process. Further, it may be concluded that CELL CYCLE AND DNA PROCESSING mechanism is common not only among fungal organisms but also all organisms, considering that *S. cerevisiae* and *A. gossypii* are close to prokaryote and *S. pombe* and *A. oryzae* are close to higher eukaryote.

3.5 Synteny blocks and functional clusters

Synteny blocks are calculated based on not only sequence similarity analysis but also genome rearrangement analysis and therefore we can also discuss from macro-scopic view. 32 synteny blocks among 33 synteny blocks which are obtained in this work include only one gene or only one functional sequence and the remaining one synteny block includes two genes, *tub1* and *tub3*. Both *tub1* and *tub3* are alpha tubulin proteins which are involved in mitosis and mating (fertilization) and functions of these two proteins are remarkably similar to each other. It is also confirmed by genetic mutation experiment that TUB1 and TUB3 interacts with each other.

The result that almost synteny blocks include one gene or one functional sequence indicates that gene orders among four fungal organisms are fully shuffled by numerous rearrangements. Nonetheless, two genes *tub1* and *tub3* are included in the same synteny block. What biological significance does this result indicate?

It is known that some genes which have similar functions are closely located and construct a functional cluster. A typical example of the functional cluster is a set of genes, *ena1*, *ena2* and *ena5*, which are located on the long arm of chromosome V of *S. cerevisiae*. Although such functional clusters are founded among other organisms, it is not clear whether such functional clusters are just a result of tandem duplication, or such functional clusters are biologically necessary clusters in the sense that all genes in the same cluster are controlled by a similar transcriptional mechanism.

Our result that *tub1* and *tub3* are included in the same synteny blocks shed light on this problem. Our result means that the *tub1* and *tub3* are closely located on all of four fungal genomes and this indicates that to separate *tub1* from *tub3* are depressed in evolutionary process. Therefore we conclude that it is essential for fungal organisms to maintain functional clusters and those functional clusters are not just a result of tandem duplication, but are biologically necessary clusters.

REFERENCES

[1] Bourque, G., Pevzner, P. and G. Tesler. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. Genome Research, 14(4), 2004, 507–516.

[2] Bourque, G. and P. Pevzner. Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome Research, 12, 2002, 26–36.

[3] Darilng, A.E., Mau, B., Blattner, F.R., and N.T. Perna. Mauve: multiple alignment of conserved genomic sequence with rearrangement. Genome Research, 14, 2004, 1394–1403.

[4] Darilng, A.E., Mau, B., Blattner, F.R., and N.T. Perna. GRIL: Genome rearrangement and inversion locator. Bioinformatics, 20, 2004, 122–124.

[5] Dobzhansky, T., and A.H. Sturtevant. Inversions in the chromosomes of Drosophila pseudoobscura. Genetics, 23, 1938, 28–64.

[6] Hannenhalli, S. and P. Pevzner. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). Proceedings of the 27th annual ACM symposium on the Theory of Computing, pp, 1995, 178–189.

[7] Hannenhalli, S. and P. Pevzner. Transforming men into mice (polynomial algorithm for genomic distance problem). Proceedings of the 36th annual IEEE symposium on Foundations of Computer Science, pp, 1995, 581–592.

[8] Pevzner, P. and G. Tesler. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. Genome Research, 13, 2003, 37–45.

[9] Ma, B., Tromp, J. and M. Li. PatternHunter: faster and more sensitive homology search. Bioinformatics, 18(3), 2002, 440–445.

[10] Thompson, J. L., Higgins, D. G. and T. J. Gibson. CLUSTAL W: improving the sensivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Research, 22, 1994, 4673–4680.