ALIGNSCOPE : A VISUAL MINING TOOL FOR GENE TEAM FINDING WITH WHOLE GENOME ALIGNMENT

HEE-JEONG JIN,[1] HYE-JUNG KIM,[1] JEONG-HYEON CHOI[2] AND HWAN-GUE CHO[1]

[1]*Dept.of Computer Science and Engineering, Pusan National University,South Korea*
E-mail : {hjjin,hjkim,hgcho}@pusan.ac.kr
[2]*School of Informatics, Indiana University, Bloomington, IN 47404, USA*
E-mail : jeochoi@indiana.edu

One of the main issues in comparative genomics is the study of chromosomal gene order in one or more related species. Recently identifying sets of *orthologous* genes in several genomes has become getting important, since a cluster of similar genes helps us to predict the function of unknown genes. For this purpose, the whole genome alignment is usually used to determine horizontal gene transfer, gene duplication, and gene loss between two related genomes. Also it is well known that a novel visualization tool of the whole genome alignment would be very useful for understanding genome organization and the evolutionary process. In this paper, we propose a method for identifying and visualizing the alignment of the whole genome alignment, especially for detecting *gene clusters* between two aligned genomes. Since the current rigorous algorithm for finding gene clusters has strong and artificial constraints, they are not useful for coping with "noisy" alignments. We developed the system *AlignScope* to provide a simplified structure for genome alignment at any level, and also to help us to find gene clusters. In this experiment, we have tested AlignScope on several microbial genomes.

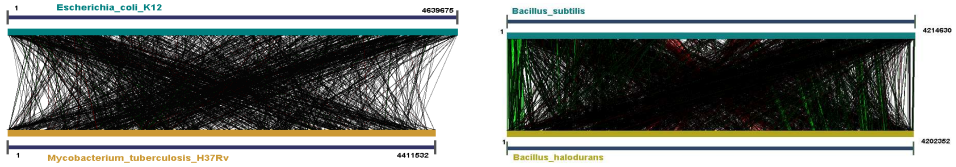
1. Introduction

Alignment is a procedure that compares two or more sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences. This procedure assists in designating the functions of unknown proteins, determining the relatedness of organisms, and identifying structurally and functionally important elements and other useful functions.[9,12] Many widely divergent organisms are descended from a common ancestor through a process called evolution. The inheritance patterns and diversities of these organisms have significant information regarding the nature of small and large-scale evolutionary events.

The complexity and the size of the genome make it difficult to analyze. Because the large amount of biological noises is present when visualizing genomes, it is not enough to simply draw the aligned pairs of various genomes. Therefore an alignment visualization tool needs to provide a method for viewing the global structure of whole genome alignment in a simplified form at any level of detail. Figure-1 clearly illustrates this problem. In Figure-1, the resolution of the snapshot is 800 by 600 pixels, so one pixel corresponds about 6000 bases of a given genome sequence.

Currently there are several systems for visualizing the alignment of genomes. The NCBI Map Viewer[14] provides graphical displays of biological features on NCBI's as-

sembly of human genomic sequence data. GeneViTo[5] is a JAVA-based computer application that serves as a workbench for genome-wide analysis through visual interaction. GenomePixelizer[1] generates custom images of the physical or genetic positions of specified sets of genes in whole genomes or parts of genomes. This method assists in comparing of specified genes between genomes. However, the available genome viewers do not conveniently display the global structure between two genomes since these viewers only draw sets of alignment or gene pairs (Figure-1) and genome alignment has a lot of aligned pairs. So, we propose a novel visualization system to establish meaningful features, especially *gene clusters*, in whole genome alignment using "zoned hierarchical clustering" method. Zoned hierarchical clustering method clusters all alignment pairs to illustrate a simplified view.



(a) (b)

Figure 1. A snapshot of whole genome alignment using AlignScope. The snapshot (a) shows the alignment between *E. coli K12 (4.6Mbp)* and *M. tuberculosis H37Rv (4.4Mbp)*. The snapshot (b) shows the alignment between *Bacillus subtilis (4.2Mbp)* and *Bacillus halodurans (4.2Mbp)*.

A fundamental question in genomics is the relationship between chromosomal gene order and function. Current evidence suggests that genes are not randomly distributed on the chromosomes and that genes which are physically close to each other tend to represent groups of genes with a functional relationship even if they are not contiguous. These groups of genes are called "gene clusters" or "gene teams" in two or more genomes.[3,6,7] Identifying conserved gene clusters is important for many biological problems, such as genome comparative mapping, studying transcriptional neighborhoods, predicting gene functions and other problems. For this purpose, Steffen[7] and Takeaki[13] proposed that a common interval in a sequence be used, but in practice, they also considered assumptions. Corteel[3] introduced the concept of a "gene team", which is a set of orthologous genes that appear in two or more species, possibly in a different order yet with the distance between the genes on the team for each chromosome is always within a certain threshold. For this model to function properly each gene can have at most one orthologous partner in the other chromosome. Xin[6] removes this constraint in the original model, and thus allows the analysis of complex prokaryotic or eukaryotic genomes with extensive paralogs. Kim[8] searches for gene clusters with and/or without physical proximity constraint. Other work identifies functional modules from genomic association of genes using protein interaction networks.[11]

Currently the rigorous algorithms for finding gene clusters require strong and artificial constraints. For example, distances between adjacent genes within a cluster are smaller

than given threshold. Since they allow a few "noise" genes which don't included in a cluster, they are not useful for coping with the "noisy" genome that occurs in practice. Our method enables not only to provide a simplified structure of genome alignment but also to assist in detecting of gene clusters.

2. Clustering for Gene Pairs

Since the previous visualization tools for genome alignment mainly consider genetic information such as ORF and gene prediction and annotation, they are not useful in the study of the relationships.Our AlignScope visualizes relationships at any simplified level and also predicts gene clusters using a zoned hierarchical clustering algorithm.

2.1. *Preliminary*

In this paper, we only consider a pairwise whole genome alignment. Many local alignment tools such as BLAST, FASTA, MUMer and GAME can be used to obtain the alignment. Without loss of generality, we denote an upper genome by U and a lower genome by L.

- $u_i = [v_i,\ w_i]$ ($l_j = [p_j, q_j]$) is a subsequence of U (L) where v_i and w_i are the start and end positions of u_i at U and p_j and q_j are those of l_j at L.
- A geometry center $M(u_i) = \frac{v_i+w_i}{2}$ and $M(l_j) = \frac{p_j+q_j}{2}$.
- An alignment pair $a_i = (u_i, l_j)$ where u_i is the opponent of l_j in terms of an aligned pair.
- The geometry center $M(a_i) = (M(u_i), M(l_j))$ where $a_i = (u_i, l_j)$.
- The distance $\Delta(a_i, a_j)$ between two alignment pairs, $a_i = (u_{i_1}, l_{i_2})$ and $a_j = (u_{j_1}, l_{j_2})$ is defined by

$$\begin{aligned}\Delta(a_i, a_j) &= \Delta(M(a_i), M(a_j)) = |M(a_i) - M(a_j)| \\ &= |M(u_{i_1}) - M(u_{j_1})| + |M(l_{i_2}) - M(l_{j_2})| = \frac{|v_{i_1} - v_{j_1} + w_{i_1} - w_{j_1}| + |p_{i_2} - p_{j_2} + q_{i_2} - q_{j_2}|}{2}\end{aligned}$$

- A cluster $c_x = \{a_i \mid \forall a_p, a_q \in c_x$ and $\forall a_k \notin c_x, \Delta(a_p, a_q) < \Delta(a_p,\ a_k)$ and $\Delta(a_p,\ a_q) < \Delta(a_q,\ a_k)\}$.
- $|c_x|$ denotes the cardinality of a cluster c_x.
- An interval $I(c_x) = ([v_x,\ w_x],\ [p_x,\ q_x])$ where v_x and w_x are the minimum and maximum positions of c_x at U, and p_x and q_x are those of c_x at L.
- A distance between two clusters, c_x and c_y,

$$\Delta(c_x,\ c_y) = \frac{\sum_{i,j} \Delta(\ a_i,\ a_j)}{(|c_x| \ +\ |c_y|)},\ \text{ for }\ a_i \in c_x, a_j \in c_y.$$

Figure-2 shows an example of our notations.

- In genome U, $u_1 = [1,8]$, $u_2 = [11,15]$, $u_3 = [21,26]$.
- In genome L, $l_1 = [3,7]$, $l_2 = [14,19]$, $l_3 = [23,30]$.
- $a_1 = (u_1, l_3)$, $a_2 = (u_2, l_1)$, $a_3 = (u_3, l_2)$.
- $I(a_1) = ([1,8],[23,30])$, $I(a_2) = ([11,15],[3,7])$, $I(a_3) = ([21,26],[14,19])$

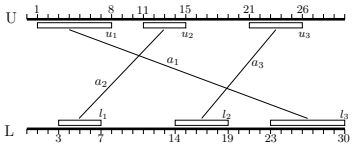


Figure 2. . The plot shows an example of genome alignment between two genomes U and L. In genome U, $u_1 = [1,8]$, $u_2 = [11,15]$, $u_3 = [21,26]$. In genome L, $l_1 = [3,7]$, $l_2 = [14,19]$, $l_3 = [23,30]$.

2.2. *Zoned Hierarchical Clustering*

Hierarchical clustering is a statistical method for finding relatively homogeneous clusters and determines clusters of similar data points in multi-dimensional spaces based on empirical data and a certain kind of distance measure.[10] It starts with each case in a separate cluster and then combines the clusters sequentially, reducing the number of clusters at each step until only one cluster is left. When there are N input data, this involves $N-1$ clustering steps, or fusions.

Now we will cluster all alignment pairs to illustrate a simplified view from the whole genome alignment in order to construct a simplified structure of alignment pairs. We use the concept of hierarchical clustering modified, "*zoned hierarchical clustering*", for clustering of alignment pairs. There are 6 stages for zoned hierarchical clustering:

(1) Sort all alignment pairs by their geometric center on the basis of upper genome.
(2) Assign each alignment pair to a base cluster, i.e., $c_x = \{a_x\}$. If we have n alignment pairs initially, we will have n clusters at the start.
(3) For each cluster, we must compute local "effecting zone" which is dependent on the structure of the on-going cluster. (This procedure will be explained later.)
(4) Find the nearest pair of clusters using $\Delta(c_x,\ c_y)$. In this procedure, we consider only "effecting zone" of each cluster for search area. This will prevent searching all candidate clusters.
(5) Merge the pair of the nearest clusters. This is the basic procedure of hierarchical clustering and procedure produced the clustering tree.
(6) Repeat steps 3 and 5 until all clusters are merged.

Effecting Zone of each cluster

Typical hierarchical clustering techniques find the nearest pair of each pair in the alignment pairs. In the case of single-link clustering, this process takes very long time, $O(N^2)$. For fast clustering, we need to consider a small local zone to find the nearest clusters. So our method takes $O(k \cdot N)$, where the number of data in a local zone is k.

2.3. *Two Density Measures for Level-of-Detail and Detection of Gene Clusters*

The zoned hierarchical clustering process generates a tree or a dendrogram, where each step in the clustering process is illustrated by merging two clusters. Each internal node in

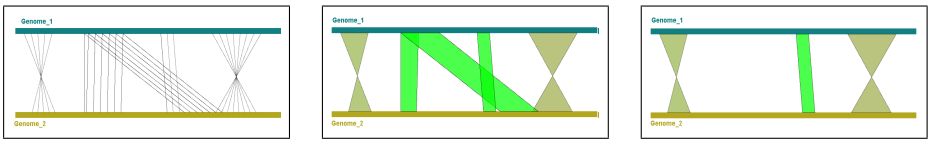
clustering tree represents a cluster. We define two *density* measures for each cluster to view the simplified structure of whole genome alignment as level of detail. One is *"area density (σ_a)"* and the other is *"line density (σ_l)"*. Area density is the number of alignment pairs divided by the sum of the two interval distances at U and L of the geometric polygon (in fact a trapezoid) that covers alignment pairs in c_x (Figure-4). Line density is the number of *properly* contained alignments divided by all alignments on either side of c_x in U and L. The formal definition of σ_a and σ_l is given in the following:

- Given $I(c_x)$ = $([a,b],[c,d])$, $\sigma_a(c_x)$ = $|c_x|/((b-a)+(d-c)+2)$.
- Let the number of alignment pairs contained in $I(c_x)$ be t, then $\sigma_l(c_x) = 2*|c_x|/t$.

In Figure-4, $\sigma_a(c_x)$ is $4/((U_e-U_s)+(L_e-L_s)+2)$ and $\sigma_l(c_x)$ is 8/12. AlignScope can visualize at any simplification level to select clusters with a certain σ_a or σ_l in a tree. And since σ_l means a noise rate of a cluster, we can detect gene clusters according to the noise rate.

2.4. *Spliced Gene Clusters*

In addition to providing a simplified structure for detecting gene clusters, AlignScope can find gene clusters. Figure-3 shows the clusters detected by our method and common interval method. Note that several alignment pairs have been mixed at *Genome_1* while most of them are pure at *Genome_2*. The clusters consist of these alignment pairs, they shall be called "*spliced clusters*". The spliced clusters are determined by zoned hierarchical clustering. We speculate that the spliced cluster between the genomes is related to evolution.



(a) (b) (c)

Figure 3. Comparison of zoned hierarchical clustering and common interval method. The plot (a) shows test data for alignment pairs. The plot (b) shows clusters obtained by zoned hierarchical clustering. The plot (c) shows clusters found by common interval method.

3. Visualization of Gene Pairs

Data Filtering

AlignScope is able to display the features of the whole genome in real-time. Since it is sometimes hard to understand genes in specific region or with certain alignment score, AlignScope supports various filtering options to analyze these data. It supports the following filtering functions:

- Filtering by alignment score: AlignScope uses a large number of alignment pairs from alignment programs such as BLAST[2] and GAME[4] with specific alignment score.
- Filtering by physical position: Since images drawn by AlignScope fit typical computer monitors without scrolling due to the large size of genome, it is difficult to view information on individual genes or alignment pairs. So, AlignScope extracts data by selected regions of interest.
- Filtering by $|c_x|$s within a cluster: Zoned hierarchical clustering generates clusters with a various $|c_i|$. AlignScope can extract clusters by a $|c_i|$.

Shape and Color of Alignment pair

There are two kinds of gene clusters, *parallel* clusters, and *reverse* clusters which mean that the order of the aligned genes is reversed. The colors and shapes of alignment pairs represent the degree that they are parallel or reverse. Let $Color(c_x)$ be a color assigned to cluster c_x, N be the number of crossing alignment pairs in c_x, and $|c_x| = M$. In Figure-4, the total number of crossing alignment pairs is 8. And two shapes, which are sand clock and rectangle (Figure-3), are used to represent clusters according to the number of crossing alignment pairs in c_x. If the number of crossing alignment pairs in c_x is larger than $\frac{M(M-1)}{4}$, the shape of the cluster is a sand clock, and in other case, it is a rectangle.

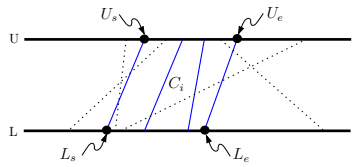


Figure 4. The plot shows an example of a cluster c_x where all solid lines are aligned pairs and contained in c_x. The line density $\sigma_l(c_x)$ = 8/12, since $|c_x|$ = 4 and the number of aligned pairs contained in $I(c_x)$ at U and L, respectively, is 6.

- $Color(c_x) = rgb(0,\ 255,\ 0)$, if $N = 0$.
- $Color(c_x) = rgb(255,\ 0,\ 0)$, if $N = M(M-1)/2$.
- $Color(c_x) = rgb(255 \cdot 2k/M(M-1),\ 255 \cdot (1-2k/M(M-1),\ 0)$, for any.

4. Experiment Result

We have tested the performance of AlignScope on two data sets. AlignScope uses a large number of alignment pairs from alignment programs such as BLAST and GAME, or gene pairs between two genomes. The first data set is a group of three prokaryote genomes, *A. fulgidus*, *P. abyssi*, and *M. thermautotrophicus*. We generated by COGs database alignment pairs from *A. fulgidus* vs. *P. abyssi* and *A. fulgidus* vs. *M. thermautotrophicus*, respectively, i.e., an alignment pair is made if a pair of genes in each genome is belonged to the same

Table 1. The longest cluster produced by AlignScope with σ_l = 0.9 for a pair of *A. fulgidus* vs. *M. thermautotrophicus*.

Index	Our	Kim[8]	COG	gene of A. fulgidus	gene of M. thermautotrophicus	description
1	O	O	COG0201	secY	secY	protein translocase, subunit SEC61 alpha (secY)
.	.	.	.	.	.	.
16	O	O	COG1588	-	-	RNAse P protein subunit P29
17	O	X	COG0092	rps3P	rps3P	SSU ribosomal protein S3P (rps3P)
18	O	X	COG0091	rpl22P	rpl22P	LSU ribosomal protein L22P (rpl22P)
19	O	X	COG0185	rps19P	rps19P	SSU ribosomal protein S19P (rps19P)
20	O	X	COG0090	rpl2P	rpl2P	LSU ribosomal protein L2P (rpl2P)
21	O	X	COG0089	rpl23P	rpl23P	LSU ribosomal protein L23P (rpl23P)
22	O	X	COG0088	rpl4P	rpl4P	LSU ribosomal protein L4P (rpl4P)
23	O	X	COG0087	rpl3P	rpl3P	LSU ribosomal protein L3P (rpl3P)

COG. Figure-5 shows an example of the steps to detect gene cluster using AlignScope. Table-1 shows the longest cluster produced by AlignScope with $\sigma_l \geq 0.9$. While Kim et al.[8] detected 16 genes, AlignScope detected 23 genes. As the result of comparing

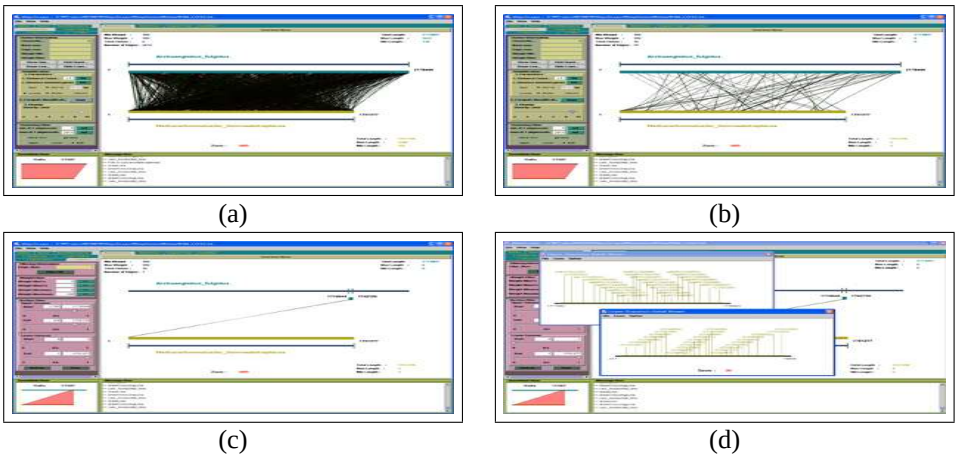


Figure 5. Example of the steps to detect a gene cluster which is shown at Table-1 using AlignScope. The plot (a) shows a snapshot of all alignment pairs of *A. fulgidus* and *M. thermautotrophicus*. The plot (b) shows a snapshot of detected gene clusters with $\sigma_l \geq 0.9$. The plot (c) shows a snapshot after filtering by physical position $[1710044, 1742720]$ at *A. fulgidus*. The plot (d) shows a snapshot of genes information in a cluster of (c).

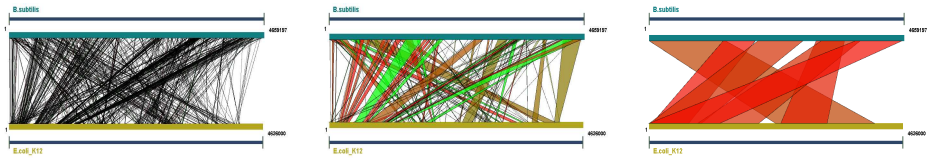
A. fulgidus with *M. thermautotrophicus*, the longest cluster found by AlignScope has the same genes except COG0255 as shown in Table-2. It is interesting that all genes except COG0255 in *A. fulgidus* are parallel with those in *M. thermautotrophicus*, but all genes in *A. fulgidus* cross completely to those in *P. abyssi*. Note that AlignScope represents the rate of crossings of genes in a cluster to its color and shape (See Section 3).

The second data-set is a pair of *E.coli K12* vs. *B.subtilis*. Alignment pairs are made in the case that a pair of genes in each genome has the same gene name. AlignScope enables to adjust the simplified level of clusters (See Section 2.4). Figure-6 shows global structures

Table 2. Example of a conserved gene cluster within three archaebacteria, *A. fulgidus*, *M. thermautotrophicus* and *P. abyssi*.

	1	2	3	4	5	6	7	8	9
A. fulgidus	COG0201	COG0200	COG1841	COG0098	COG0256	COG2147	COG1717	COG0097	COG0096
M. thermautotrophicus	COG0201	COG0200	COG1841	COG0098	COG0256	COG2147	COG1717	COG0097	COG0096
P. abyssi	COG0087	COG0088	COG0089	COG0090	COG0185	COG0091	COG0092	COG0255	COG1588
	10	11	12	13	14	15	16	17	18
A. fulgidus	COG0199	COG0094	COG1471	COG0198	COG0093	COG0186	COG1588	COG0255	COG0092
M. thermautotrophicus	COG0199	COG0094	COG1471	COG0198	COG0093	COG0186	COG1588		COG0092
P. abyssi	COG0186	COG0093	COG0198	COG1471	COG0194	COG0199	COG0096	COG0097	COG1717
	19	20	21	22	23	24			
A. fulgidus	COG0091	COG0185	COG0090	COG0089	COG0088	COG0087			
M. thermautotrophicus	COG0091	COG0185	COG0090	COG0089	COG0088	COG0087			
P. abyssi	COG2147	COG0256	COG0098	COG1841	COG0200	COG0201			

of clusters at several simplified level, $Sim(level)$; 0, 0.4, 0.8, and 1. It is worth notifying that the global structure of clusters becomes clear and vivid. Table-3 shows the result of



(a) $Sim(0)$ (b) $Sim(0.4)$ (c) $Sim(0.8)$

Figure 6. The simplified structure of gene clusters for alignment pairs between *E.coli K12* and *B.subtilis*. The plot (a) shows input alignment pairs. The plot (b) shows clusters which the smallest size is 10 and the largest size is 50. The plot (c) shows clusters which the smallest size is 45 and the largest size is 120.

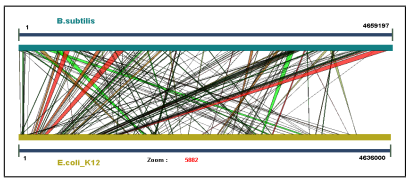
comparing the spliced clusters described Section 2.5 with Xin's.[6] We checked some clusters and compared to the result of previous work. For the selected gene clusters, AlignScope detected most of sets of gene clusters found by Xin.[6] Interestingly, there were gene clusters of candidate clusters determined by AlignScope that Xin did not detect. To the best of our knowledge, the concept of a "spliced" gene ordering or alignment has not been reported before. Now we are investigating the biological meaning of the "spliced genes" found by AlignScope. Figure-7 (b) shows a set of spliced clusters in Table-3.

5. Conclusion

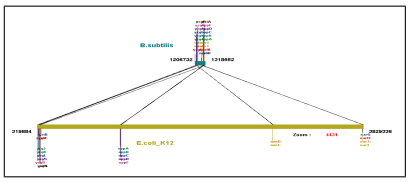
In this paper, we proposed a new method for visualization of whole genome alignment. A novel visualization tool for the whole genome alignment would be very useful for understanding genome organization and the evolutionary process. Several genome viewers already exist, each one serves a different need and research interest. AlignScope is easy to use in a computer environment and helps understanding the relationship and gene clusters between two genomes. Our system is freely available on http://jade.cs.pusan.ac.kr/∼alignscope. The main features of AlignScope are as follows:

Table 3. Example of a set of spliced clusters. Since AlignScope does not consider "noisy" alignment pairs at clustering procedure, we can obtain the fine spliced clusters.

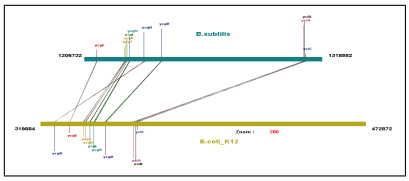
OUR	Xin [6]	COG	gene	product
O	X	COG0789	ycgE	putative transcriptional regulator
		-	ycgJ	orf, hypothetical protein
		-	ycgK	orf, hypothetical protein
		COG3100	ycgL	orf, hypothetical protein
		COG2983	ycgN	orf, hypothetical protein
		COG2719	ycgB	putative sporulation protein
		-	ycgR	orf, hypothetical protein
		COG1607	yciA	Acyl-CoA hydrolase
		COG2917	yciB	Intracellular septation protein A
		-	yciC	orf, hypothetical protein
O	O	COG2894	minD	cell division inhibitor, a membrane ATPase, activates minC
		COG0850	minC	cell division inhibitor, inhibits ftsZ ring formation
O	O	COG0243	narG	nitrate reductase 1, alpha subunit
		COG1140	narH	nitrate reductase 1, beta subunit
		COG2180	narJ	nitrate reductase 1, delta subunit, assembly function
		COG2181	narI	nitrate reductase 1, cytochrome b(NR), gamma subunit
O	O	COG0747	oppA	oligopeptide transport; periplasmic binding protein
		COG0601	oppB	oligopeptide transport permease protein
		COG1173	oppC	homolog of Salmonella oligopeptide transport permease protein
		COG0444	oppD	homolog of Salmonella ATP-binding protein of oligopeptide ABC transport system
		COG1124	oppF	homolog of Salmonella ATP-binding protein of oligopeptide ABC transport system



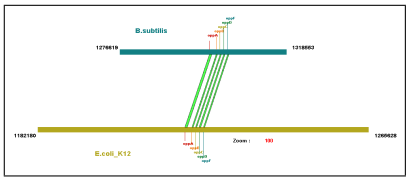
(a) all clusters between *E.coli K12* and *B.subtilis*



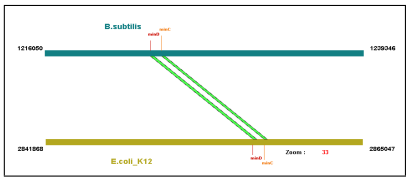
(b) a set of spliced clusters in Table-3



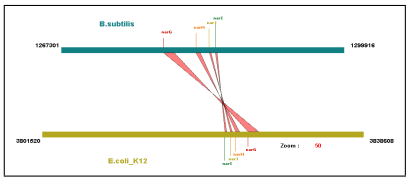
(c) information of first cluster in (b)



(d) information of second cluster in (b)



(e) information of third cluster in (b)



(f) information of fourth cluster in (b)

Figure 7. A set of "spliced clusters" found by AlignScope. The plot (a) shows detected spliced clusters between *E.coli K12* and *B.subtilis*. The plot (b) shows whole spliced clusters in Table-3 and the plot (c)∼(f) show the information of each cluster in (b).

- AlignScope provides an intuitive controls for the visualization of whole genome alignment at any simplified level.

- AlignScope is very-fast since we only consider a few "effecting zones" in each alignment and not the whole region of a genome. This improves the intractability between biologist and bioinformatics software.
- By using AlignScope, the candidate sets of gene clusters in whole genome can be easily found. In addition, AlignScope can detect the interesting "spliced" gene transfer, which was not possible in the previous approaches based on single string algorithms.

Acknowledgments

This work was supported by the Korea Research Foundation Grant(F01-2004-000-10016-0). We gratefully credit the thoughtful reviewers, who provided substantial constructive criticism on an earlier version of this note.

References

1. Kozik A, Kochetkova E, and Michelmore R. Genomepixelizer–a visualization program for comparative genomics within and between species. *Bioinformatics*, pages 335–336, 2002.
2. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, pages 403–410, 1990.
3. A. Bergeron, S. Corteel, and M. Raffinot. The algorithmic of gene teams. *In Proc. Second Annual Workshop on Algorithms in Bioinformatics*, pages 464 – 476, 2002.
4. Jeong-Hyeon Choi, Hwan-Gue Cho, and Sun Kim. Multiple genome alignment by clustering pairwise matches. *RECOMB Comparative Genomics Satellite Workshop, Lecture Notes in Bioinformatics*, pages 30–41, 2004.
5. Vernikos GS, Gkogkas CG, Promponas VJ, and Hamodrakas SJ. Genevito: visualizing gene-product functional and structural features in genomic datasets. *BMC Bioinformatics*, page 53, 2003.
6. Xin He and Michael H. Goldwasser. Identifying conserved gene clusters in the presence of orthologous groups. *In Proc. Research in Computational Molecular Biology*, pages 272–280, 2004.
7. Steffen Heber and Jens Stoye. Finding all common intervals of k permutations. *Proceedings of CPM 01,volume 2089 of Lecture Notes in Computer Science*, 2001.
8. Sun Kim, Jeong-Hyeon Choi, and Jiyoung Yang. Gene teams with relaxed proximity constraint. *IEEE Computational Systems Bioinformatics (CSB'05)*, 2005.
9. Needleman S.B. and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Moleculer Biology*, pages 443–453, 1970.
10. R. Sibson. Slink: An optimally efficient algorithm for the single-link cluster method. *Comput. J*, pages 93–95, 1973.
11. Berend Snel, Peer Bork, and Martijn A. Huynen. The identification of functional modules from the genomic association of genes. *Proceedings of the National Academy of Sciences*, pages 5890–5895, 2002.
12. Smith T.F and Waterman M.S. Identification of common molecular subsequences. *Journal of Moleculer Biology*, pages 195–197, 1981.
13. Takeaki Uno and Mutsunori Yagiura. Fast algorithms to enumerate all common intervals of two permutations. *Algorithmica*, pages 290 – 309, 2000.
14. NCBI Map Viewer. http://www.ncbi.nih.gov/mapview.