# NIBB 2010-INTERNSHIP PROGRAM

**Prepared by**

**Walaa A. Adly**

**Under supervision of:**

**Prof. Dr. Ikuo Uchiyama**

**Genome informatics laboratory.**

# Report

I am Walaa A. Adly, I am from Egypt, a graduate student of faculty of Agriculture, Cairo University, Biotechnology group, also I have graduated from information technology institute (iti), Bioinformatics department as fellowship, and finally I was an international intern in NIBB 2010 internship. Really it would be my pleasure to have the chance to be an international intern of this respectable institute in Okazaki city in Japan, which was very different place from where I have always lived. Personally, I have enjoyed visiting and staying in it. In addition, I had the chance to know such valuable civilization, beside that the Mishima lodge, the place I stayed in it, was very comfortable and safe, which I didn't faced any problems or troubles.

My internship helped me to gain a lot of experiences and skills in a variety of biological fields especially in genome informatics field as well as being able to communicate with different culture and society. My work in the laboratory was very interesting and enthusiastic which we able to combine biological sciences with informational technology in addition to implement scientific programs such as detect genes responsible for diseases by using highly qualified software. The team was very helpful and friendly. Really I liked all people and respected them. Also I would like to thank my professor for his support to me which taught me how I think about things differently.

I am honor and proud of developing my experience and skills of the scientific research in the ideal environment for creativity. Finally I hope to study and complete my PhD. in this innovative and pioneer climate. Now I will summarize my work in genome informatics laboratory.

# Introduction

Comparative genome analysis has played an important role in understanding microbial life from a genomic prospective. Also in genomics context and gene functions are conserved between orthologs, whereas paralogs generally acquire different functions; therefore ortholog identification is crucial for identyfing a pair of genes that have the same function in different genomes, especially in genome annotation. Here we will make an orthologous table containing genes in a specified pathway by using RECOG system to figure out the relationship between orthologous groups involved in this pathway (1).

# Background

RECOG stands for Research Environment for Comparative Genomic. It is a software to integrate a lot of methods of comparative genomics and incorporate the knowledge of individual genomes and gene functions into one analysis. RECOG software has a client – server architecture which allow users to choose a set of organisms to compare and constructs orthologous groups among those organisms and predict the function related (1).

MBGD database stands for microbial genome database for comparative analysis. It provides orthology information for every gene stored in it, which identifies orthologous relationships among the genomes. MBGD database is the basis of the RECOG system and we can access this database through it (1).

# The first week

I downloaded the RECOG software and started to use it to make some analysis, through which I learned how to predict the new pathway function using orthologous table from the software.

Functions of RECOG system:

1- Phylogenetic profile:
Phylogenetic pattern is defined as a binary vector for each orthologous group that represents the presence or absence of genes in each genome, and in this method, we can predict the related function, if two orthologous groups have similar phylogenetic patterns.

2- Domain fusion method:
We can find two genes are fused to each other in one gene. If you find such a pattern, you can infer that these two genes have a related function.

3- Gene neighborhood method:
Genes appeared in close position in two genomes in multiple chromosomes then we can predict that they have a related function.

4- Neighboring gene cluster function in the program:
Identifies a set of genes that are located in both orthologous table and genomic sequence and assign the same color to each gene in that set.

5- Zoom: to see a detailed map of a specific orthologous group for identifying gene names.
6- Search: to search for keywords in cluster annotation and gene annotation.
7- Connect to MBGD database where we can click on compare maps and identify conserved arrangement of the same orthologous groups between different organisms. Choose gene list from database &

choose ORFID to display a comparative genome region map containing only the specified genes.

## Task 1:

### Make list of genes with species names involved in TCA cycle:

Gene context and gene fusion are important in function prediction. If there are two genes with the same color and closed to each other, then this is a case of gene order conservation and they may have the same function. And if there are two genes enclosed with frame, which indicates there is a gene fusion and thus they have a related function and work together.

1- I prepared a list of cog id of TCA cycle from reference (2).Then I downloaded a file containing all COG entries from FTP site. I prepared a file containing the correspondence between species name of COG and that of MBGD database.
2- I wrote a Perl script which executed these files and finally produced an output file that contains a list of genes with species names involved in TCA cycle and we will used it in another Perl script to be used in RECOG system.
3- This output file can be incorporated into the RECOG system by importing the gene set file to define a gene set, which is used to filter the orthologous table by creating new gene set combination.

## Task 2:

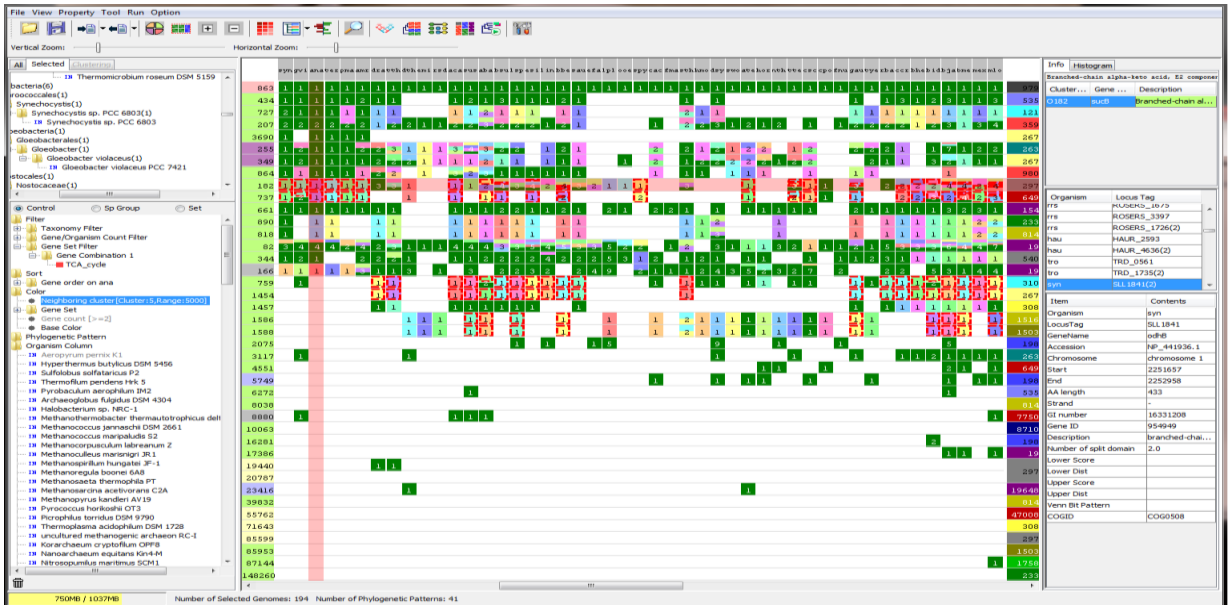
### Make gene property file:

I wrote Perl script which used the downloaded COG file, the list of COG_ID of TCA cycle. After that I can incorporate the gene property file into RECOG system by import it. The imported properties are displayed in the gene information window.

## Task 3:

After filtering out using the TCA cycle gene set generated in task 1 on the RECOG system, I examined the adjacency of genes on the chromosome using the neighboring gene cluster function to see to

what extent gene neighborhood relationships can be observed in this pathway and summarized the result.



A snapshot of the RECOG window that showing the adjacency of genes on the chromosome using the neighboring gene cluster function.

## Task 4:

## Make Cluster property file:

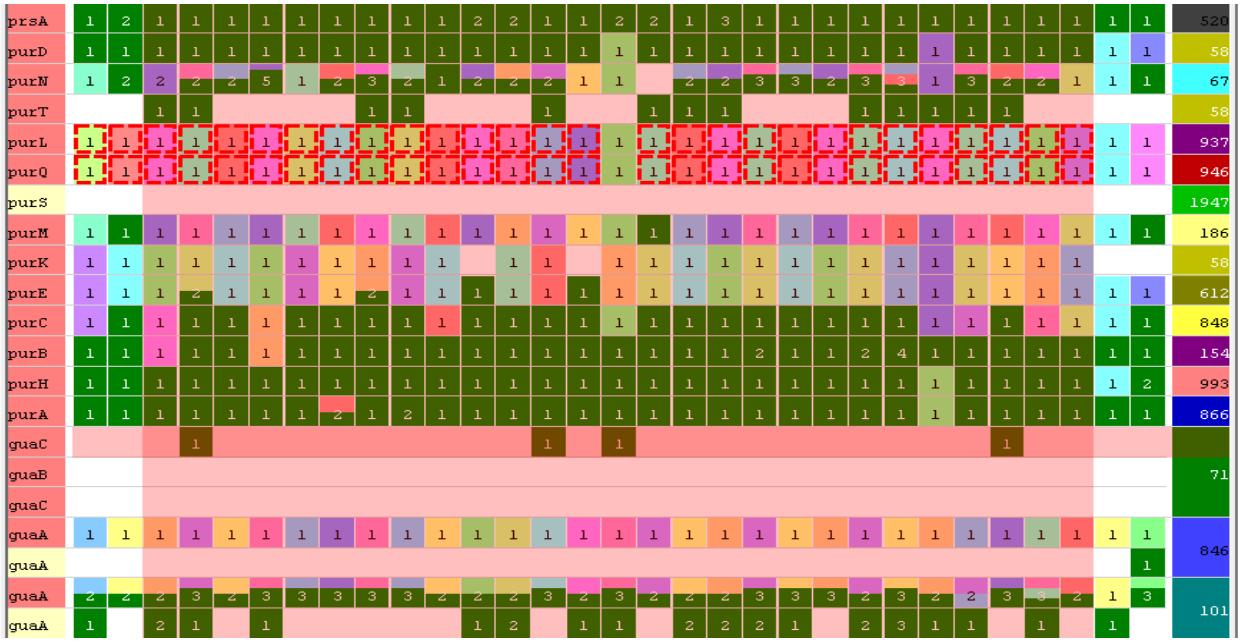
Aim: to make correspondence between RECOG ortholog id and COG id automatically. For this purpose, maximally assigned COGID is identified for each orthologous group.

## Task 5:

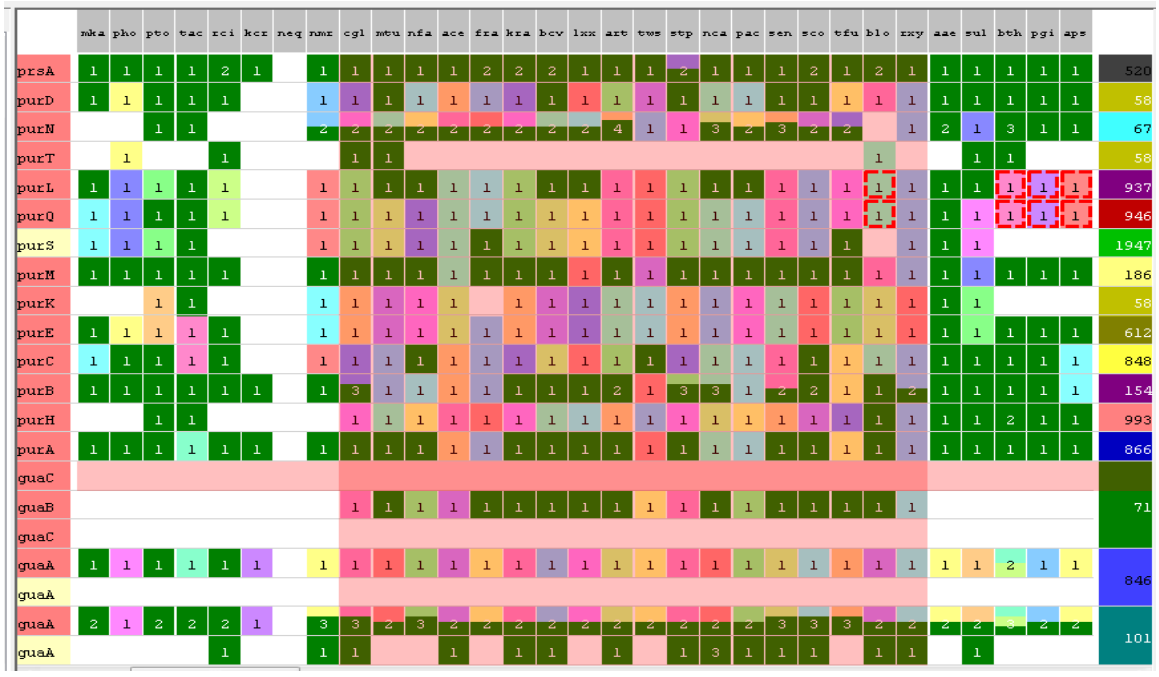
Make analysis to Amino acid biosynthesis in Purine which explain the relationship between orthologous groups

Example: Gamaproteobacteria, Actinobacteria and Bacilli.

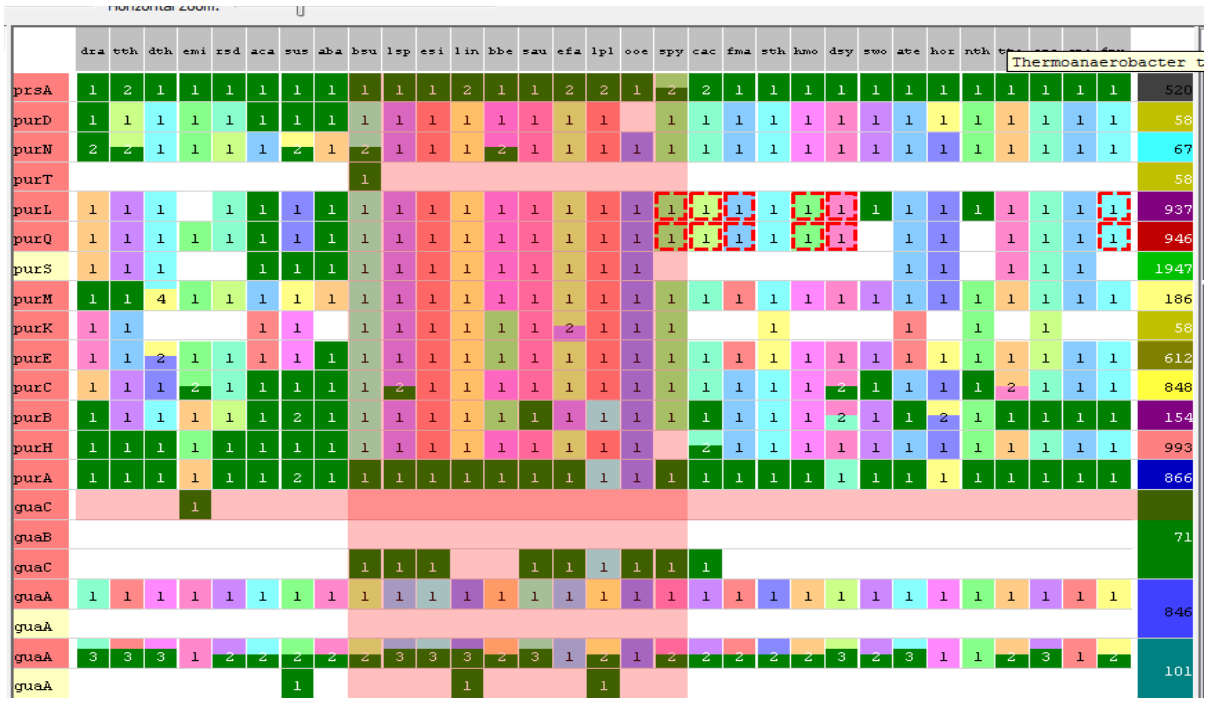
## *Gammaproteobacteria*




## *Actinobacteria*



***Bacilli***




Observation:

*Gammaproteobacteria:*

I observed that limited number of genes is clustered

*Actinobacteria:*

I observed that more genes are clustered than Gammaproteobacteria, but they are separated into multiple small clusters.

*Bacilli:*

I observed that larger clusters of genes are existed.

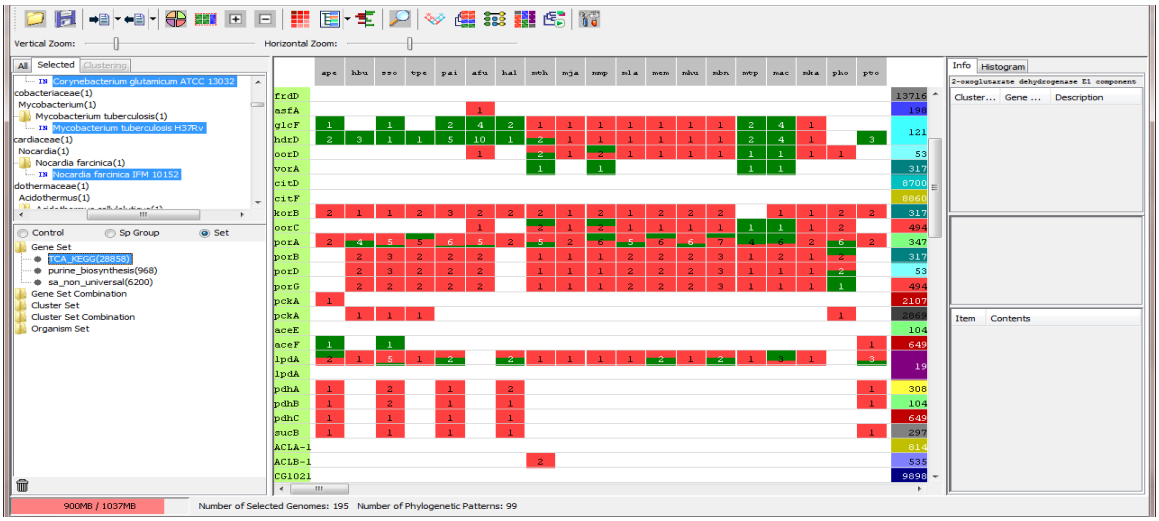
# The second week

## Background:

KEGG represents molecular functions using pathway map diagram which provides a hierarchal gene function classification scheme that is much simpler than GO which is more useful for microbial genome annotation and comparison. Also COG database is clusters of orthologous group in microbial comparative genomics. (1)

## The task:

Use KEGG database and use gene property file and cluster property file and list of KEGG database which obtained from KEGG website (FTP) and introduce them to RECOG system. Rewrite Perl scripts for using KEGG database instead of the COG database.



A snapshot of RECOG system showing the orthologous table of genes in TCA cycle according to KEGG database.

## The result:

Related genes have the same orthologous pattern, located in the chromosome across to each other. Also phylogenetic pattern clustering based on the pattern occurrence in each orthologous group within a genome.

# The third week

In this week I prepared & presented two presentations. The first was titled "Bioinformatics Algorithms" from book "An introduction to Bioinformatics Algorithms"(3)and I presented the second chapter "Algorithms and complexity" and defined the meaning of Algorithms and its types, the differences between biological algorithms & computer algorithms, the difference between iterative and recursive algorithms as Recursive algorithms can often be rewritten to use iterative loops instead, and vice versa; it is a matter of clarity that dictates which technique is easier to use. Also I presented different algorithm design techniques like Exhaustive search or Brute force Algorithm, Branch and Bound Algorithms, Greedy Algorithm, Dynamic programming Algorithm, Divide and conquer Algorithm, Machine learning Algorithm and Randomized Algorithm.

The second presentation was titled "Scalar Data of Perl programing language" from book "Learning Perl Book" (4). I presented the definition of Scalar and the difference between Scalar numbers & Scalar strings, Single –Quoted String Literals & Double –Quoted String Literals. Also the definition of Scalar Variables with different operators, and finally the if Control structure & the while control structure with scalars.

# The fourth week

## Background:

Some species have the same content of pathogenic element; we compared different strains but closely related species. We compared such strains of some species & identified the differences among strains (we compared species included in the same genus).

Core genome analysis is a useful approach for comparison of closely related genomes. Core genome is the conserved region among the same species but each strain has the additional sequences in each variable region so the difference between strains can be found in variable region. Genes related to pathogenicity is often found in variable region.

Variable region often has different characteristics from core region because prokaryotes often acquire new sequences from horizontal gene transfer. For example, the following characters may be different between variable and core regions.

Example: *Staphylococcal* food poisoning, methicillin resistant *staphylococcus aureus.*

## Method:

Identify variable region and fix region among *Staphylococcus aureus* genomes. For this purpose, first we identified core region based on the order of orthologous genes among strains using RECOG system with domclust and core genome alignment programs. Then, we defined variable region as genes that are not included in the core region and are not conserved universally and listed what kind of genes are included in the variable region.

We characterized functions of genes in the variable region by using gene ontology database that defines various gene functions in a hierarchal manner. We used BinGO, a plugin of the CytoScape software, for knowing a set of what kind of functions (in terms of GeneOntologyID) are enriched in the variable region of *S. aureous.*

# The fifth week

## Introduction:

Bacterial genome has a horizontal gene transfers, and some genes have different codon usage so we can examine the codon usage to identify horizontally transferred genes. There are many factors for determining codon usage such as GC content. Codon usage is generally constant within species and often correlated with the number of tRNAs in the cell for efficient translation. If codon preference is similar throughout one genome so this is efficient way but if codon preference is different between genes in one genome this is not efficient way.

## Aim:

analyzing flexible region using gene ontology and observing different aspects of flexible region by using CGAT software which stands for Comparative Genome Analysis Tool to compare nucleotide sequences instead of protein sequences. Also CGAT has features to calculate some properties such as codon usage and GC content and it is client server architecture.

We used Mackintosh machine to install the CGAT server

## Results:

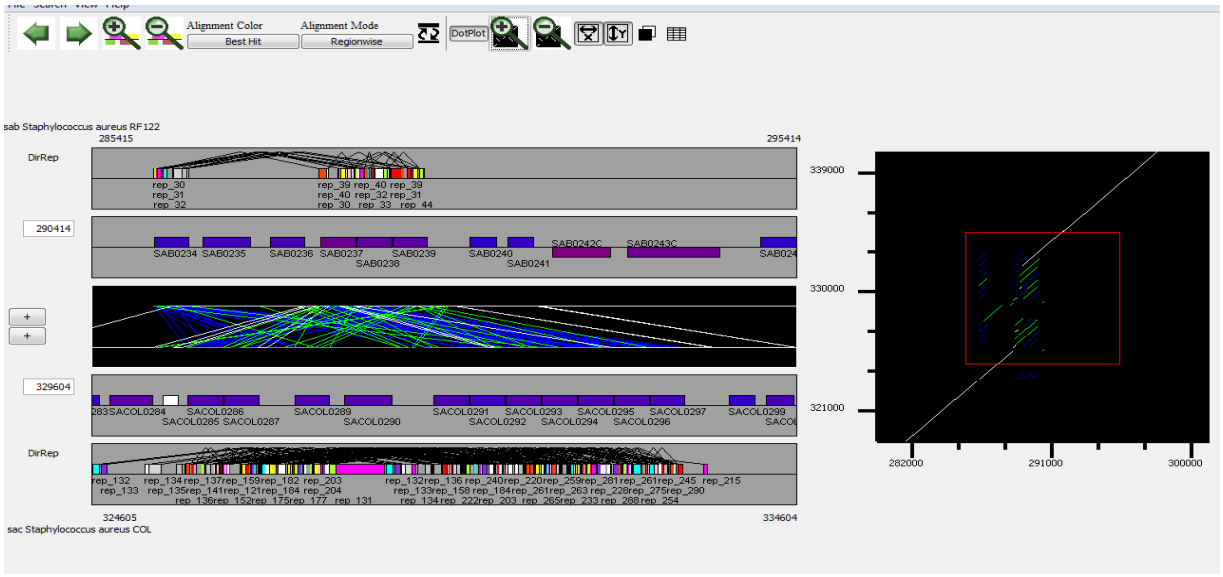
We compared *Staphylococcus aureus* COL (sac) and *Staphylococcus aureus* ED98 (sad) strains using CGAT. We found a highly repetitive region between sac (400174-581424+) and sad (338351-519601+). By examining the paralogous relationships within this region using the RECOG system I found that : In the species of *staphylococcus aureus* COL (400174-581424+), all genes from SACOL0481 to SACOL0486 have similarity to each other, which corresponds to the paralog cluster (homcluster id 4) according to the RECOG system, and also with comparison with *staphylococcus aureus* ED98 (338351-519601+), I found all genes from SAAV_0379 to SAAV_0388 have similarity to each other which

correspond to the  same paralog cluster  in the RECOG system . The number of paralogous genes are changed between strains.

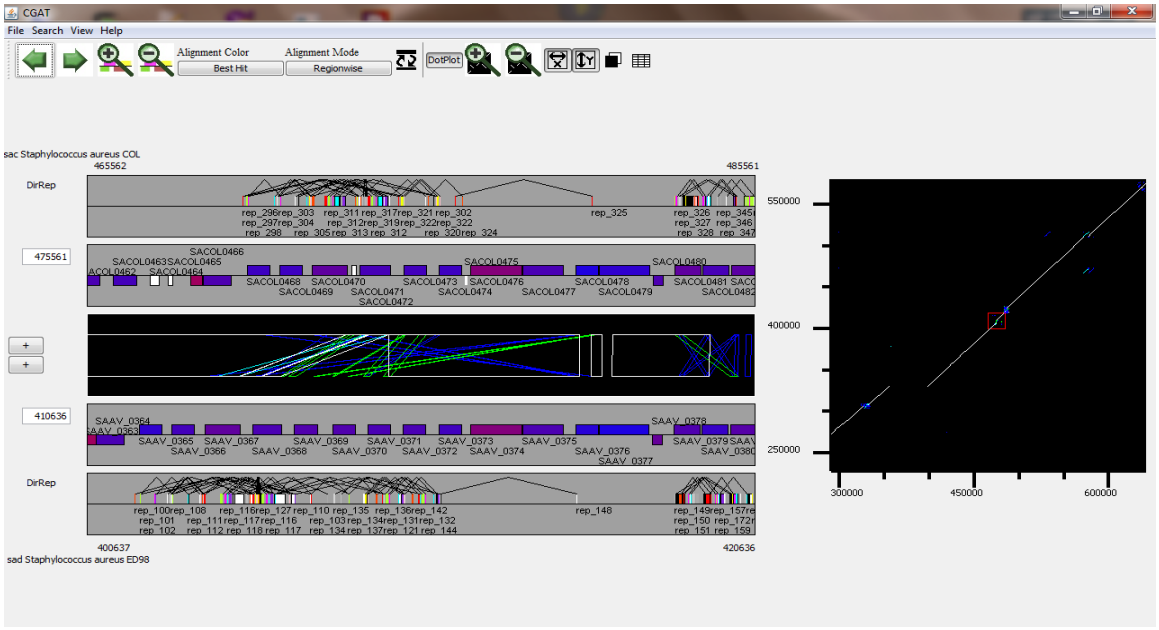
Next we examined the phylogenetic relationship for each paralogous group (homcluster id 4) in the RECOG system to check whether there is a significant difference in the topologies of phylogenetic trees among them.

## Conclusion:

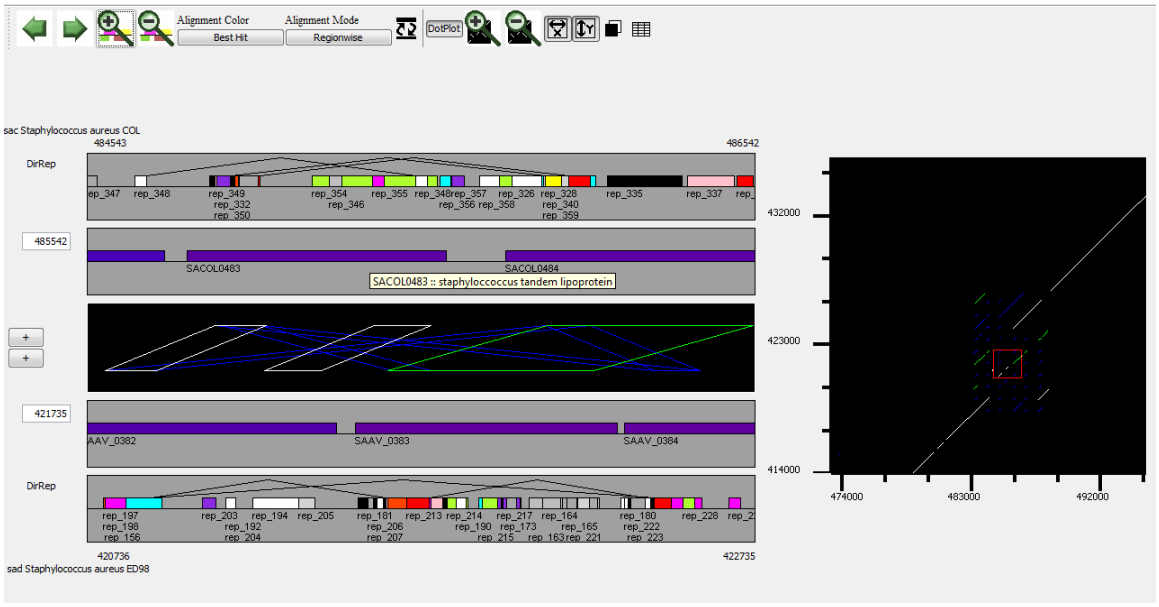
After checking the tree topologies manually, we cannot find any clear difference among tree topologies. Therefore, in this case, we cannot conclude that recombination has a large effect on the phylogenetic tree topologies of each orthologous group in this paralog cluster.




position: 290414  Sac – sab




Position: 475561 in sac and 410636 in sad




Position 485542 in sac and 421735 in sad

# References


(1) Prof. Dr. *Ikuo Uchiyama*, Book chapter, "Functional inference in microbial genomics based on large-scale comparative analysis", in "Protein function prediction for omics era", Kihara, D. ed., Springer, in press.

(2) *Eugene V.Koonin and Michael Y.Galpern*, Reference book "Sequence – Evolution – computation function approaches in comparative Genomes."

(3) *NEIL C.JONES AND PAVEL A.PEVZNER*, Reference book "An introduction to Bioinformatics Algorithms"

(4) *Randal L. Schwartz, Tom Phoenix & Briand foy*. Reference book "Learning Perl Book"