ORF Clustering Analysis of 17 Complete Genomes

Kenji Suzuki [12]
suzuken@kuicr.kyoto-u.ac.jp
Takeshi Ara [1]
takeshi@bs.aist-nara.ac.jp
Hideo Matsuda [3]
matsuda@ics.es.osaka-u.ac.jp
Hirotada Mori [1]
hmori@gtc.aist-nara.ac.jp

[1] Research and Education Center for Genetic Information, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0101, Japan
[2] Institute for Chemical Research, Kyoto University, Gokasho Uji, Kyoto 611-0011, Japan
[3] Department of Informatics and Mathematical Science, Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan

1 Introduction

With thanks to the rapid improvement of sequencing technique, many micro organisms have been revealed the entire structure of their genomes. At present, the complete genome sequences of 17 bacteria and 1 eukaryote (Baker's Yeast) have been determined. The new research area, comparative genomics of complete genome sequences is certainly being established. We just rowed out into the genome biology.

In the beginning of 1997, we showed the complete genome structure of *Escherichia coli* W3110 strain. At the same time, US group also showed that of MG1655 strain. About 4,300 orfs were predicted on its chromosome. Although *E. coli* is undoubtedly one of the most intensively studied organisms, less than 2,000 orfs had been analyzed in the past. As the consequence, more than 50% of *E. coli* orfs were function unknown. To elucidate the function of these orfs is the next target of our project. As one of the approaches to predict their function or structure, we have tried to identify the functional, evolutional or structural unit of orfs, named orf component [1]. Genes from the same ancestral origin could be shown similarities. On the basis of this hypothesis, we performed a robust similarity analysis by all v.s. all pairwise alignment. We clustered all of the orfs predicted using single linkage clustering algorithm and picked out the region which showed local similarity as component. From this analysis, 25% of *E. coli* orfs were clustered. This approach, however, has a strict limitation when the analysis is restricted within *E. coli* orfs. This depends on the numbers of paralogous orfs. To overcome this limitation, we expand our analysis to the 17 complete genome sequences including eukaryotic organism, *Saccharomyces cerevisiae.*

2 Method

The method adopted was essentially the same as we reported previously [1] except the final step to extract components.

The steps of our approach are as follows:

1. All of 38,301 orfs were submitted to pairwise similarity analysis using BLASTP.
2. All orfs were clustered by single linkage clustering algorithm by 1e-05 of P value as a threshold.
3. To extract components using Maximum density subgraph method (MDS) algorithm.

2.1 MDS Algorithm

The outline of our algorithm for finding local homologous sequence is as follows.

Step 1. Generating the blocks.

Given n protein sequences P_i $(1 \leq i \leq n)$ and a block length l, all possible blocks $P_i(1), P_i(2), \cdots, P_i(|P_i| - l + 1)$ $(1 \leq i \leq n)$ are extracted from the sequences, where $|P_i|$ is the length of P_i.

Step 2. Constructing block graph.

The block similarity between every pair of blocks generated in Step 1 is computed. Then, given a cutoff score, a weighted edge is drawn between any two blocks whose block similarity is higher than or equal to the score (where the weight of the edge is equal to block similarity), otherwise no edge is drawn.

Step 3. Finding a maximum-density subgraph.

Given the block graph constructed in Step 2, search for the maximum-density subgraph in the graph. Note that the block graph may be partitioned into several connected components due to the cutoff. In this case, the maximum-density subgraph is determined for each connected component.

Step 4. Combining overlapping blocks.

Some blocks in the subgraphs obtained in Step 3 may overlap each other in some sequences. These overlapping blocks are combined into larger segments.

3 Result and Discussion

28,952 orfs out of 38,301 were classified to 2,684 clusters by single linkage clustering using BLAST program. The biggest cluster consists of 14,289 orfs and 12,935 component were extracted by MDS methods from our analysis. However, the extraction of component candidate in some clusters failed because of computational problem. From the clustering analysis of *E. coli* orfs, only 25% orfs of *E. coli* could be clustered. As described above, this absolutely depends on the number of paralogous orfs. So we expand our analysis to the orfs predicted from 17 complete genome sequences. As the result, 75% of *E. coli* orfs were observed to show orthologous relation to the orfs predicted in other organisms. 269 *E. coli* orfs were observed to be *E. coli* specific because no orfs of other species did not consist in these groups. On the other hand, 79 clusters were observed in all of the organisms analyzed in common and 49 orfs of *E. coli* were found in these clusters. 25 out of 49 orfs function as translation and 5 orfs were function unknown. Further analysis to identify components is now under way.

Acknowledgments

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, 'Genome Science', from The Ministry of Education, Science, Sports and Culture of Japan

References

[1] Suharnan, S., Itoh, T., Matsuda, H., Mori, H., Clustering Molecular Sequences with Their Components, *Genome Informatics 1997*, Universal Academy Press, 125–131, 1997.