Detection of Co-regulated Genes by Comparative Analysis of Microbial Genomes

Hiroyuki Ogata ogata@kuicr.kyoto-u.ac.jp
Wataru Fujibuchi wataru@kuicr.kyoto-u.ac.jp
Minoru Kanehisa kanehisa@kuicr.kyoto-u.ac.jp

Institute for Chemical Research, Kyoto University
Gokasho, Uji, Kyoto 611-0011, Japan

1 Introduction

Rapid accumulation of genomic sequences is giving us a new opportunity to investigate and understand intricate biological systems, which may be represented as networks of genes and gene products. Several new experimental methods such as DNA microarray technology would be utilized to produce a huge amount of expression profiles and to determine correlated expression of genes in a wide variety of cells. The informatics based approaches should take large parts in this research area of functional genomics. Especially computational techniques to predict co-regulated genes must be developed, since they provide useful information for designing the experiments and for interpreting the correlated expression of genes. It is well known that, in some bacteria and also in archaea, several genes with functional links are often clustered on the genomes. Such a clustering of genes implies common regulation of genes, for example, by the mechanism of polycistronic transcription. In some cases, multiple transcripts carrying related functions are co-regulated by common factors. In this study, we propose a new technique to detect possible functional links between genes that are not necessarily clustered in the genome.

2 Materials and Methods

Thus far we have constructed the gene cluster database (GCDB). To construct the GCDB, we took two kinds of strategies. Firstly, conserved gene clusters were extracted by interspecies comparisons of localization of orthologous genes. Secondly, clusters of genes that are functionally related in metabolic pathways were extracted by the network comparison technique [1].

Here we introduce an index to measure the degree of cluster formation (DCF) for arbitrary pairs of genes in the genome. The DCF is based on the information of gene pairs that appear in the same gene clusters in the GCDB. Consider a pair of genes, a_0 and b_0 in a genome G_0, which is denoted by $P_0^{a,b}$, and their orthologs a_i and b_i in another genome G_i ($i = 1...n-1$; n is the number of organisms considered). If a_i and b_i are in the same cluster according to the GCDB, they are defined as a clustered pair $C_i^{a,b}$. If an appropriate measure is given to estimate the distances between $P_0^{a,b}$ and $C_i^{a,b}$, and between $C_i^{a,b}$ and $C_j^{a,b}$, DCF is defined as the following equation:

$$DCF = \sum_i Dis(P_0^{a,b}, C_i^{a,b}) + \sum_{i,j} Dis(C_i^{a,b}, C_j^{a,b}).$$

In this study, we employed the distance between small subunit rRNA sequences as the distance measure in the equation.

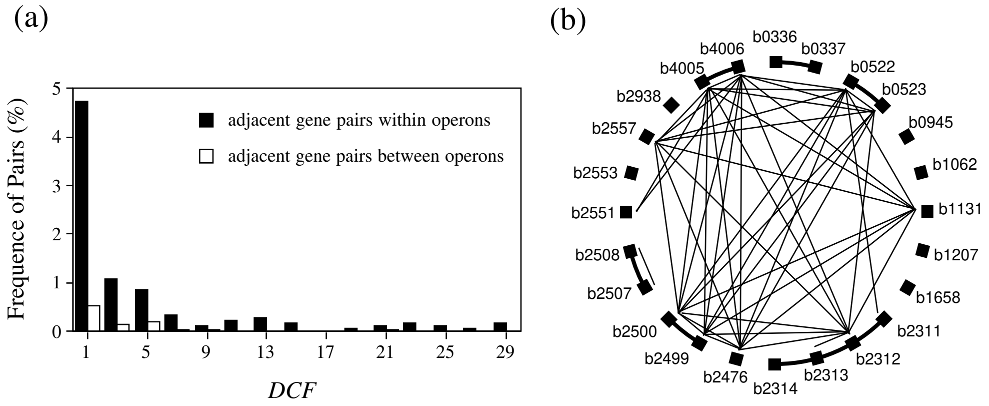

Figure 1: (a) Relative frequency of gene pairs that are adjacent along the *E. coli* genome is plotted against their DCF values. Filled bars for the adjacent gene pairs within operons, and open bars for the pairs across operons. (b) Pairs of genes with $DCF > 0$ among 24 co-regulated genes (black boxes) are connected by thin lines. Thick lines indicate operon organizations.

3 Results and Discussion

We used the GCDB constructed from the genomic data of fifteen organisms including ten bacteria and five archaea. In order to see the correlation between DCF and operon organization, we plotted the distribution of DCF values for adjacent pairs of genes both for those within operons and for those across operons by using the *E. coli* operon data (Fig. 1 (a)). It is clearly shown that the adjacent gene pairs with non-zero DCF values are more frequent within operons than across operons (Pairs with $DCF = 0$ are not shown). However, at the moment, the operon prediction utilizing DCF did not improve the prediction based only on the intergenic distances (the performance was about 74%). We expect that the performance could be improved with the increasing number of organisms used for construction of the GCDB and/or with the optimization of parameters.

The relation of DCF values with regulon organization must be another important issue. In Fig. 1 (b), relatedness by DCF values among 24 genes that are separate in the genome are represented. These genes are relevant to purine nucleotide synthesis and known to be co-regulated as a regulon. Interestingly most genes are tightly connected by lines representing DCF values larger than 0. We will further study the efficient use of DCF for detecting co-regulated genes, especially regulons, by combining with additional information of gene expression profiles.

Acknowledgements

This work was supported in part by the Grant-in-Aid for Scientific Research on the Priority Areas 'Genome Science' from the Ministry of Education, Science, Sports and Culture in Japan. The computation time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

References

[1] Ogata, H., Goto, S., Fujibuchi, W. and Kanehisa, M., Computation with the KEGG pathway database, *BioSystems*, 47:119–128, 1998.