

Stochastic Local Genome Alignment and Comprehensive Search for Conserved Gene Clusters

Tsuyoshi Hachiya

hacchy@dna.bio.keio.ac.jp

Yasubumi Sakakibara

yasu@bio.keio.ac.jp

Dept. Biosciences and Informatics, Keio University, 3-14-1 Hiyoshi, Kohoku-ku,
Yokohama-city, Kanagawa 223-8522, Japan

Keywords: comparative genomics, conserved gene cluster, genomic alignment, stochastic model

1 Introduction

Since the *H. influenzae* genome project had completed in 1995, subjects of early-state genome project were representative organisms such as *E. coli*, *C. elegans*, *D. melanogaster* or *H. sapiens*. After almost all representative genome projects had completed, next stage of genome project which sequences neighbor species of representative organisms started. Thus comparative genomics which compares pairwise or multiple neighbor genome sequences and discusses common feature and different feature among compared genomes are increasingly attended.

One of comparative genomics contentions is the searching conserved gene clusters which are gene clusters whose gene order is conserved among neighbor organisms. So far the biological meaning of gene order has not been discussed sufficiently. Searching conserved gene clusters comprehensively and analyzing common feature among conserved gene clusters provides evidence which suggests biological meanings of gene order. Additively searching conserved gene clusters leads to find synteny blocks which are genome-wide conserved regions and to identify genome rearrangement scenario.

2 Method and Results

Several methods such as [1] or [3] are developed to search conserved gene clusters. [1] is based on graph algorithm and detect conserved gene clusters among multiple genomes. [3] incorporates the information captured by a genome phylogenetic tree and detect conserved gene clusters among multiple microbial genomes. Both methods and other representative methods require precise annotation on all comparing genomes because those methods compare gene order directly. However, as for higher eukaryotic genomes, only a few genomes precisely annotated because complicated gene structure such as alternative splicing makes it difficult to annotate ORF information precisely.

Therefore we propose a new novel method to search conserved gene clusters, which requires precise annotation on only one genome. Our method first compares genome sequences and detects large conserved regions using stochastic model. Let the large conserved region be *local genome alignment region*. Second it searches ORF annotations which are included in local genome alignment region on one representative genome. If multiple ORFs are included in the same local genome alignment region, those genes are defined as conserved gene cluster. The detail of first step are described below.

Detect Local Genome Alignment Region In this study conserved region which is separated by non-conserved regions is defined as “local genome alignment” region. Our method detects local genome alignment region according to following algorithm.

1. Compare multiple genome sequences and detect highly conserved regions whose length is about from 100 bp to 10000 bp. Let those highly conserved regions be *anchors*. Anchors can be detected by preexisting tools such as PatternHunter [2] or Murasaki.
2. Combine neighbor and same direction anchors if distance of two anchors is “close” and each anchor size is “large”. To judge “close” distance and “large” size, we propose a new stochastic model (Figure 1). The model provides probability to a region which consist of two anchor regions and a non-conserved region according to the length of anchor region and non-conserved region. Our method learns each parameter for local genome alignment region and let it “conserved region model”. Moreover our method learns each parameter for not local genome alignment region and let it “non-conserved region model”. Our method compares two likelihood which are probabilities based on conserved region model and non-conserved region model and if likelihood based on conserved region model is larger than one based on non-conserved region model, two anchors are combined.
3. Define combined anchor regions as local genome alignment region.

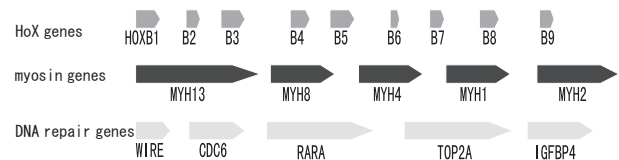
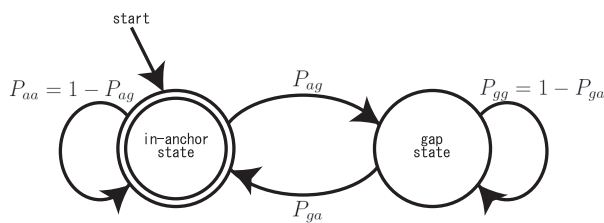


Figure 1: Stochastic model to detect local genome alignment. Figure 2: Detected conserved gene clusters whose genes have significantly similar function.

We applied our method to human 17 chromosome, mouse 11 chromosome and rat 10 chromosome comparison. As a result 37 local genome alignment regions were detected and eleven conserved gene clusters were detected according to human gene annotation.

3 Discussion

Analysis of gene functions using GO-term and χ -square test discovers that genes in the same conserved gene cluster have significantly similar function among three of eleven conserved gene clusters (Figure 2). Those conserved gene clusters are the evidence of tandem gene duplication and acquisition of new biological function.

References

- [1] Fujibuchi, W., Ogata, H., Matsuda, H., and Kanehisa, M., Automatic detection of conserved gene clusters in multiple genomes by graph comparison and P-quasi grouping, *Nucleic Acids Research*, 28:4029–4036, 2000.
- [2] Ma, B., Tromp, J., and Li, M., PatternHunter: Faster and more sensitive homology search, *Bioinformatics*, 18:440–45, 2002.
- [3] Zheng, Y., Anton, P. B., Roberts, J. R., and Kasif, S., Phylogenetic detection of conserved gene clusters in microbial genomes, *BMC Bioinformatics*, 243(6), 2005.