Poster A-23

Phylogenetic profiling approach to generate RNA-regulated protein networks in E. coli

Authors:
Nozomu Yachie (*Institute for Advanced Biosciences, Keio University*)
Koji Numata (*Institute for Advanced Biosciences, Keio University*)
Yoshiteru Negishi (*Institute for Advanced Biosciences, Keio University*)
Hiroyuki Nakamura (*Institute for Advanced Biosciences, Keio University*)
Junichi Sugahara (*Institute for Advanced Biosciences, Keio University*)
Rintaro Saito (*Institute for Advanced Biosciences, Keio University*)
Masaru Tomita (*Institute for Advanced Biosciences, Keio University*)

Short Abstract: Although investigation of whole-cell network of RNAs and proteins are all-important in order to understand cellular roles of RNAs, no significant methodologies have been established. We here present a novel approach on the basis of phylogenetic profiling to generate RNA-protein functional linkages using numerous complete genome sequences.

Long Abstract:
Recent genome and transcriptome projects have identified many examples of non-protein-coding RNA (ncRNA) genes in various species. NcRNA do not encode protein products but rather encode structural, regulatory or catalytic RNA molecules. However, cellular functions of these molecules are still largely unclear except well-known RNA families such as tRNA, rRNA and micro RNA. Although cell-wide examination of functional linkages between RNAs and proteins are required in order to understand cellular roles of RNAs, no bioinformatics or experiments based methodologies to discover them have been hitherto established.
Here, we developed a computational system to generate whole-cell network of RNA-protein functional linkages in Escherichia coli K12 strain MG1655 mainly based on phylogenetic profiling. In the system, RNA- and protein-encoded regions within E. coli genomic DNA were initially compared with genome sequences of the other 307 species (264 prokaryotes, 15 eukaryotes and 28 archaeas were retrieved via NCBI ftp server ftp://ftp.ncbi.nlm.nih.gov/) using NCBI BLAST program (comparative genomics), and homology profiles which consists of 307 homology scores (E-value) for each RNA and protein was created (homology profile). Then every pairs of two molecules were evaluated on the basis of mutual information calculated using corresponding pair of E-values. The mutual information for each pair was statistically assessed according to those calculated from shuffled homology profile and statistically significant pairs were assumed to have a functional linkage. The system runs on a CPU cluster through Sun Grid Engine, grid-computing environment. The programs are written in PERL programming language version 5 with the topical combination of G-language Genome Analysis Environment (http://www.g-language.org) package.
The functional linkage network (RNA-protein and protein-protein) of E. coli was generated full-automatically by using ~100 CPUs. The map contained 295 RNA-protein and 2,615 protein-protein linkages among 92 RNA and 748 protein nodes. Previously reported protein complexes such as those related to transport systems, transcriptional regulators and enzyme complexes were clustered in the map. In addition, some previously reported RNA-protein

linkages such those of ribosomal components (composed of large subunit and small subunit), tRNA and their corresponding aminoacyl-tRNA synthetase, and RNaseP complex were detected.

The generated network map was further validated by comparing it with the documented protein-protein interactions. With 1,127 interactions in EcoCyc database and 161 in DIP database, the coverage and accuracy of our generated network were significantly different from random network and also higher than those generated by previous works using phylogenetic profiling. We also mapped these links into 111 unique metabolic pathways in KEGG database and found that the linked proteins and RNAs are likely to participate in the same pathway. These data suggested that the generated map indeed reflects cellular network of E. coli. Based on our results and the other genome-scale data, e.g. microarray expression profiles and results of comprehensive gene knockout experiments, possible function and metabolisms of ncRNAs are discussed.