Abstract

Availability of various genome-wide datasets provides the opportunity to study the whole genome behavior of the organisms as well as predicting new functions for unknown genes. With the advent of the *omics* data, molecular biology has evolved from a rather data-poor to an extremely data-rich era. Several smaller scale studies already have shown how the integration of different omics data can result in a better mechanistic understanding of the cellular organism. In addition to omics data, co-expression cross-species comparison is also useful to expand the available information from better studied organism to other organisms or strains where the available data is limited. In the first part of the study, we described a new co-expression cross-species comparison method to analyze microarray datasets comparatively across species and to identify the co-expressed modules of genes. For this aim, we developed a method referred to as COMODO (COnserved MODules across Organisms) that uses an objective selection criterium to identify conserved expression modules between two species. The method uses as input microarray data and a gene homology map and provides as output pairs of conserved modules and searches for the pair of modules for which the number of sharing homologs is statistically most significant relative to the size of the linked modules. We demonstrated the performance of COMODO using distantly related two model bacterial systems, *Escherichia coli* and *Bacillus subtilis*. As a notable result, we identified larger size of conserved co-expressed modules than previously predicted to exist. In addition, we identified co-expressed modules of similar elementary processes with totally different regulatory mechanisms. Later, we discussed the statistics to assess co-expression conservation between two or three organisms, and we expanded COMODO to detect the coexpression conservation across three organisms. We applied COMODO to study the expressional conservation and divergence across *E. coli*, *Salmonella enterica*, and *B. subtilis*. We observed several modules just conserved in *E. coli* and *S. enterica* including many modules related to response to various stimuli and signal transductions, even though some aspects of their life style are remarkably different (pathogenicity of *S. enterica*). Moreover, based on the conserved coexpressed modules, we could predict some conservation in the regulatory interaction of *E. coli* and *S. enterica* although the regulatory network is not available for *S. enterica*. Furthermore, we also investigated the co-expression conservation of genes involved in two special functions, *quorum sensing* and pathogenicity across *E. coli* and *S. enterica*, and we could observe fair conservation for genes involved in quorum sensing, but almost no conservation for genes involved in pathogenicity. In fact, *S. enterica* contains a much larger number of genes related to pathogenicity that are considered the main causes of difference in life style of the two phylogentically close species *E. coli* and *S. enterica.* In the second part of this study, first we explored the mutual relation between the regulatory network and microarray expression compendium in *E. coli*. For this aim, we tried to detect modules in the regulatory network which may resemble combinatorial regulators by using Fisher exact test and Monte Carlo sampling. As both of the methods, Fisher exact test and Monte Carlo sampling, failed to find modules in the regulatory network, at the next attempt we tried to define a similarity measure for each pair of genes based on their common regulators; we called it coregulatory similarity. PageRank value was used as a measure to assess the importance of a regulator in the regulatory network. Based on this measure, the more important regulators were happened to be the more global regulators. This facilitated to define the co-regulatory similarity measure between each pair of genes based on the PageRank value of their common regulators. In our definition, regulators with lower PageRank values (more local regulators) contribute more in the co-regulatory similarity of their targets. We showed this co-regulatory similarity measure exhibits high correlation with the observed co-expression on the microarray expression compendium. Based on this study we could conclude that the observed co-expressed modules are the effect of the structure of the whole regulatory network rather than a set of combinatorial regulators. We also studied the mutual relation between the regulatory network as the controlling network and the other interaction networks with non-controlling roles in the cell. To process noncontrolling interaction networks, we detected biological modules. These biological modules included modules detected in protein-protein interaction network and EcoCyc cellular pathways. The average co-regulatory similarity values of all gene pairs in

each biological module were much higher than what is expected for random genes. We also performed the analysis in the other direction, we detected modules with high co-regulatory similarity values derived from the regulatory network. We found high similarity between these expected modules based on the regulatory network and actual biological modules. In addition, we also compared the hierarchy of biological modules, built by using regulatory networks, with the one, built by using functional GO terms. The regulatory similarity between each two modules could easily be calculated by averaging our defined co-regulatory similarity value between each pair of genes across two modules. For the functional similarity, we introduced new species-specific functional similarity measure for a pair of genes, and we calculated the average value of this similarity measure between each pair of genes across two modules. We could observe rather high correlation between functional similarity value and co-regulatory similarity of two modules, implying the hierarchies built by these two measures are highly related. Based on our observation, we could explain that despite the rapid evolution of the regulatory network, the rewiring in this network would be in the direction to keep the biological modules conserved and also in higher level preserve the functional hierarchy.