Poster J-32

Phylogenetic profiles for the prediction of protein-protein interactions: how to select reference organisms?

Authors:
Jingchun Sun (*Virginia Commonwealth University*)
Yixue Li (*Bioinformation Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences*)
Zhongming Zhao (*Virginia Commonwealth University*)

Short Abstract: We performed a systematic investigation of the effects of reference organism selection on the phylogenetic profiles and the prediction of protein-protein interactions (PPIs). The results provide important guidance on the construction of phylogenetic profiles for PPI prediction and functional genomics, which has become challenging due to the large data.

Long Abstract:
Protein-protein interactions (PPIs) play key roles in the cellular processes in an organism. An accurate and efficient identification of protein-protein interactions is fundamental for us to understand the physiology, cellular functions, and complexity of an organism. The phylogenetic profile method has been widely applied in the prediction of protein-protein interactions (PPIs). Studies often use all of the available complete genomes for this method. With more than two hundred genomes complete and new ones on the horizon, it remains unclear how to select reference organisms for the profile construction and then influence the in silico prediction of PPIs. In this study, we obtained and analyzed a total of 226 complete genomes covering three domains (Bacteria, Archaea, and Eukarya) and their corresponding evolutionary trees. By choosing Escherichia coli as a target organism, we performed a systematic investigation on the effects of reference organism selection on the prediction of PPIs. First, we designed three evolutionary tree based selection methods and compared their PPI prediction. We found similar prediction performance among the three methods, indicating a robust profile when the number of organisms is large. Second, we provided the first evidence that the inclusion of reference organisms from all three domains improved the PPI prediction while the inclusion of the closely related strains or species of the target organism weakened the PPI prediction. Third, the PPI prediction could reach optimal at the fifth hierarchical level in the evolutionary tree and, at this level, by using ~ 75% of the total organisms. Our results suggest that reference organisms should be selected from the moderately and highly genetically distant organisms, rather than the closely related organisms, and by the even distribution at the fifth hierarchical level in the evolutionary tree. Our study provides important guidance on the construction of phylogenetic profiles for PPI prediction and functional genomics, which has become challenging due to the large and increasing number of available candidate organisms.