Which pathways cannot be reconstructed using protein phylogenetic profiles?

Yohan Kim[1], Shankar Subramaniam[1,2]

Keywords: genomic context, protein networks, computational proteomics

1 Introduction

Phylogenetic profile methods have shown that those proteins that share similar profiles are more likely to have same functions than those that do not [1]. After fine-tuning of these methods, their applications on a number of complete genomes have uncovered novel cellular systems [2,3]. However, what has received relatively little attention in these studies is providing explanations for why phylogenetic profile based methods cannot confidently assign functional relationships to a significant number of those proteins that are annotated in the KEGG database [4]. There are three scenarios that fit this observation. One scenario is that the number of complete genomes from which profiles were derived was not sufficient. Consequently, even though two proteins were functionally related, their profiles did not have enough number of co-varying profile elements for them to be considered similar. In the second scenario, two proteins considered were universally shared across genomes and thus the method could not pick up strong signals from comparisons of their profiles to assign functional relationships. Finally in the third scenario, a profile of one protein did show enough variations in its elements but there were no proteins with similar profiles even with 'sufficient' number of sequenced genomes being used. In order to better assess the performance of protein phylogenetic profiles based methods, pathways for *E. coli* K12 in the KEGG database and those reconstructed using the methods are compared. It is our hope that by identifying which types or classes of proteins are less amenable to phylogenetic profile based methods, we are more likely to come up with a new generation of algorithms that can assign functions to them with greater confidence.

2 Figures and Tables

total # of proteins	4311
# of protein with at least one KEGG pathway entry	1153
total # of unique protein pairs that share at least one KEGG pathway entry	38751

Table 1: *E. coli* K12 statistics.

[1] Dept. of Chemistry and Biochemistry, UCSD, 9500 Gilman Dr., San Diego, California, USA. E-mail: ykim@ucsd.edu

[2] Dept. of Bioengineering, UCSD, 9500 Gilman Dr., San Diego, California, USA. E-mail: shankar@sdsc.edu

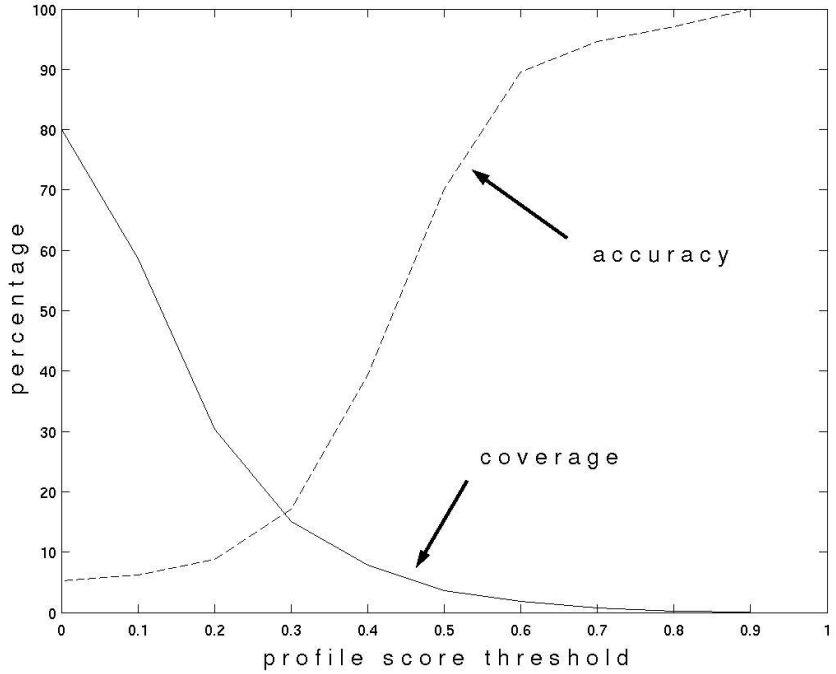

Figure 1: Accuracy and coverage of pathway predictions for *E. coli* K12 using protein phylogenetic profiles.

References

[2] Date, S.V. and Marcotte, E.M. 2003. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nature Biotechnology* 21:1055-1062.

[4] Kanehisa, M. and Goto, S. 2000. KEGG: Kyoto encyclopedia of genes and genomes.*Nucleic Acids Research* 28:27-30.

[1] Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. 1999. Assigning protein functions by comparative analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences USA* 96:4285-4288.

[3] von Mering, C., Zdobnov, E.M., Tsoka, S., Ciccarelli, F.D., Pereira-Leal, J.B., Ouzounis, C.A., Bork, P. 2003. Genome evolution reveals biochemical networks and functional modules. *Proceedings of the National Academy of Sciences USA* 26:15428-15433.