English
Research Express@NCKU Volume 28 Issue 2 - December 19, 2014
[http://research.ncku.edu.tw/re/articles/e/20141219/4.html]

Combining Phylogenetic Profiling-Based and Machine Learning-Based Techniques to Predict Functional Related Proteins

Tzu-Wen Lin, Jian-Wei Wu and Darby Tien-Hao Chang*
Department of Electrical Engineering, National Cheng Kung University
darby@mail.ncku.edu.tw

Optics Express, Vol. 18, No. 1, 165-172 (2010)

Annotating protein functions and linking proteins with similar functions are important in systems biology. The rapid growth rate of newly sequenced genomes calls for the development of computational methods to help experimental techniques. Phylogenetic profiling (PP) is a method that exploits the evolutionary co-occurrence pattern to identify functional related proteins. However, PP-based methods delivered satisfactory performance only on prokaryotes but not on eukaryotes. This study proposed a two-stage framework to predict protein functional linkages, which successfully enhances a PP-based method with machine learning (ML).

Figure 1 shows the workflow of the proposed method. A PP-based approach is employed, where only protein pairs with high phylogenetic similarity are submitted to the second stage, to reduce the data at the first-stage. A unique feature of the first stage of this study to other PP-based approaches is a non-zero filter (marked by an asterisk in Figure 1), which verifies if the phylogenetic similarity is reliable. Next, a ML-based approach is applied on the reduced data for the final prediction.

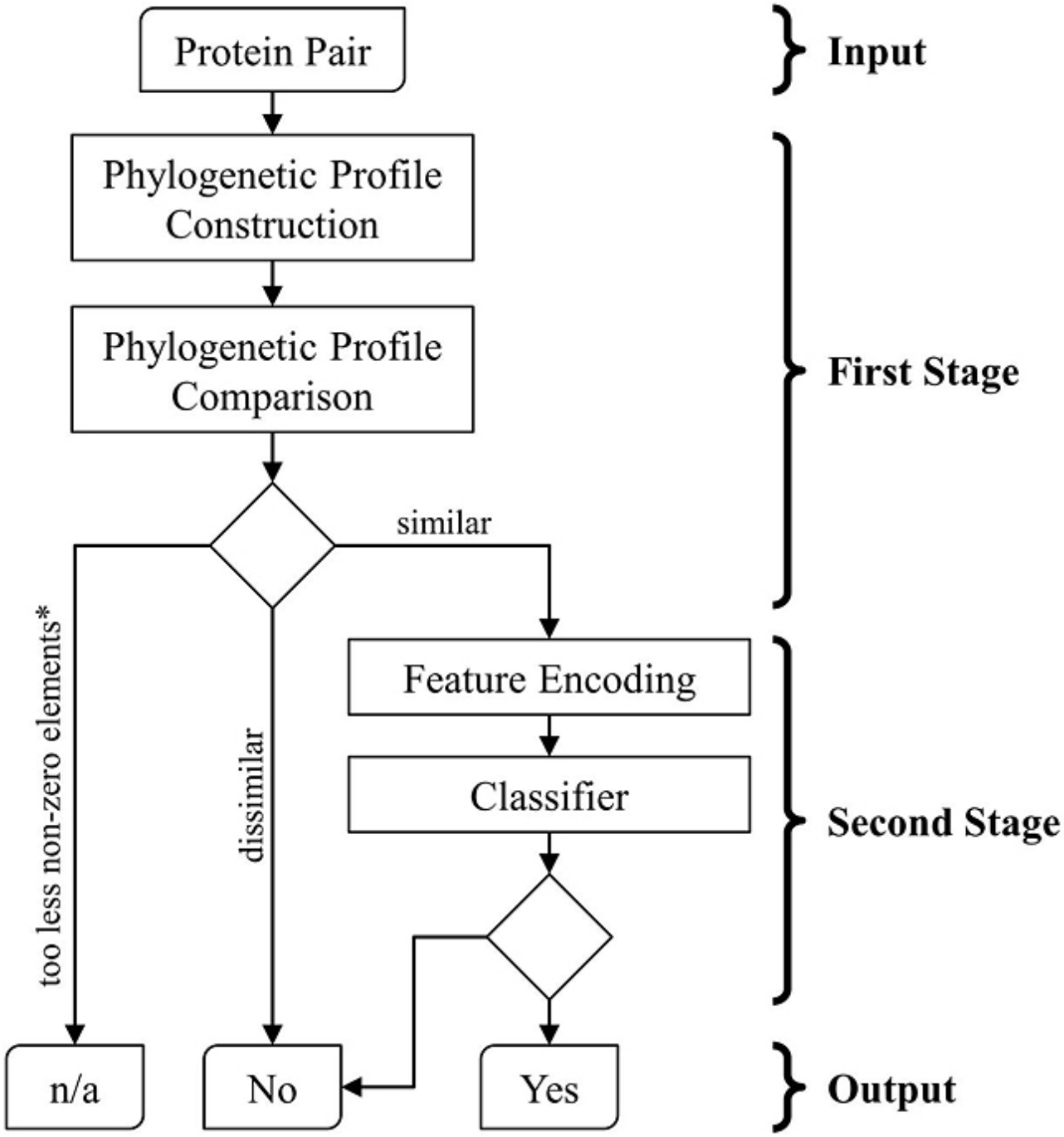

Figure 1 Workflow of the proposed two-stage framework of functional linkage prediction.

This study also elaborates the effects of the non-zero filer and the reference collections used to construct PPs. The issue of the reference collection of a PP-based could be further divided as (i) the size of the reference collection and (ii) the evolutionary distance of the reference collection to the query species. To elucidate these issues, the prediction performances of using different non-zero filters were also analyzed on two other reference collections. Figure 2a used a prokaryotic reference collection of 829 prokaryotes while Figure 2b used a eukaryotic reference collection of 132 eukaryotes. The best non-zero threshold in Figure 2a in the high-precision (namely, few-prediction) region was relatively large. This echoes that the size of the reference collection does affect the suitable non-zero filter. On the other hand, the best non-zero threshold in Figure 2b is consistent without depending on the number of predictions. With the eukaryotic reference collection, the proposed method achieved >90% and >80% precision in the top 50 and 100 prediction, respectively. All these results outperformed all the compared methods. In addition, since different applications might require different number of predictions, Figure 2 provides a useful

clues for future studies to identify the best non-zero threshold.

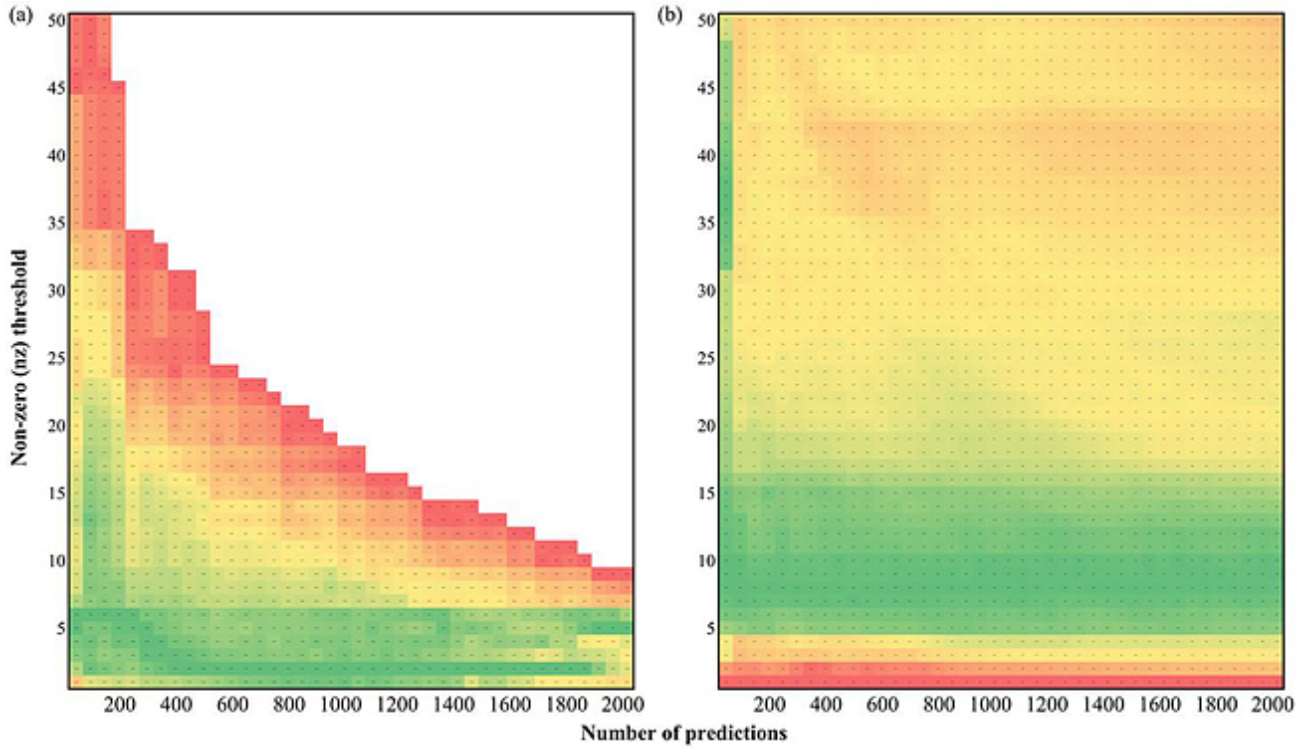

Figure 2 The relationships among the number of predictions (x-axis), the non-zero thresholds (y-axis) and the prediction performance (color) with different reference collections of (a) 829 prokaryotes and (b) 132 eukaryotes.

The proposed two-stage framework achieved good performance and preserved the advantages of both categories of techniques: (i) high performance in the top predictions of phylogenetic profiling and (ii) stable performance of machine learning. This is critical in practical applications. In addition, the proposed non-zero filter has been shown that phylogenetic profiling-based methods are promising for eukaryotes based on currently available eukaryotic genomes. The discovery of this study helps analyzing protein functional linkages and encourages developing hybrid framework in the future.