P81

Discovering novel protein-protein interactions by measuring the protein semantic similarity from biomedical literature.

Jung-Hsien Chiang and Jiun-Huang Ju.

Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan

Abstract

Protein-protein interactions (PPIs) are involved in the majority of biological processes. Identification of PPIs is therefore one of the key aims of biological research. Although there are many databases of PPIs, many other unidentified PPIs could be buried in the biomedical literature. Therefore, automated identification of PPIs from biomedical literature repositories could be used to discover otherwise hidden interactions. Search engines, such as Google, have been successfully applied to measure the relatedness among words. Inspired by such approaches, we propose a novel method to identify PPIs through semantic similarity measures among protein mentions. We define six semantic similarity measures as features based on the page counts retrieved from the MEDLINE database. A machine learning classifier, Random Forest, is trained using the above features. The proposed approach achieve an averaged micro-F of 71.28% and an averaged macro-F of 64.03% over five PPI corpora, an improvement over the results of using only the conventional co-occurrence feature (averaged micro-F of 68.79% and an averaged macro-F of 60.49%). A relation-word reinforcement further improves the averaged micro-F to 71.3% and averaged macro-F to 65.12%. Comparing the results of the current work with other studies on the AIMed corpus (ranging from 77.58% to 85.1% in micro-F, 62.18% to 76.27% in macro-F), we show that the proposed approach achieves micro-F of 81.88% and macro-F of 64.01% without the use of sophisticated feature extraction. Finally, we manually examine the newly discovered PPI pairs based on a literature review, and the results suggest that our approach could extract novel protein-protein interactions.