PIE the search: searching PubMed literature for protein interaction information

Sun Kim1,*, Dongseop Kwon2, Soo-Yong Shin3 and W. John Wilbur1

1National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA, 2Department of Computer Engineering, Myongji University, Gyeonggi-do 449-728 and 3Department of Clinical Epidemiology and Biostatistics, Asan Medical Center; University of Ulsan College of Medicine, Seoul 135-798, South Korea

Received on August 15, 2011; revised on December 2, 2011; accepted on December 17, 2011

1 INTRODUCTION

Researchers keep track of protein-protein interaction (PPI) information by searching literature online or using PPI database services. When using PPI databases, well-summarized information can be obtained. But, newly discovered evidence may be missed due to the rapid growth of the biomedical literature, it has become more difficult for biologists and curators to locate PPI information quickly. Therefore, a tool for prioritizing PPI informative articles can be a useful assistant for finding this PPI-relevant information.

Results: PIE (Protein Interaction information Extraction) is a web service implementing a competition-winning approach utilizing word and syntactic analyses by machine learning techniques. For easy user access, PIE the search provides a PubMed-like search environment, but the output is the list of articles prioritized by PPI confidence scores. By obtaining PPI-related articles at high rank, researchers can more easily find the up-to-date PPI information, which cannot be found in manually curated PPI databases.


Supplementary information: Supplementary data are available at Bioinformatics online.

2 SYSTEM AND FUNCTIONALITY

Figure 1 shows the overall architecture of the PIE the search system. The web interface manages the whole process of PPI article prediction for users. For user queries, PubMed IDs are first retrieved through online PubMed services. PPI confidence scores are calculated for retrieved articles, and articles are re-ranked based on scores. For protein name queries, this process does not guarantee highly ranked articles that contain query-specific PPIs. However, it is still likely to have useful PPI information related to protein names. The prediction module learns and classifies PubMed articles from PubMed®. To effectively capture PPI patterns from biomedical literature, our approach utilizes both word and syntactic features in a machine learning framework. Dependency parsing, gene mention tagging and term-based features are utilized along with a Huber classifier.

Since PIE the search is designed to provide only compact, but necessary features for PPI article search, its use is very straightforward, especially for PubMed users. It accepts PubMed input formats including All Fields, Abstract, Journal, MeSH Terms, Publication Date, Title and Title/Abstract with Boolean operations (AND, OR and NOT). However, the output is the list of articles prioritized by PPI confidence scores. Search results can be sorted by either PPI scores or dates. With the date sorting, only articles with PPI scores > 0.1 will appear. For convenient use, there are no page changes.
The performance of the PIE system was evaluated on the BC3 test set. The prediction module in PIE is a web-based ranking system, the search engine.

3 RESULTS AND DISCUSSION

The prediction module in PIE the search is trained by all available BC datasets except for the BC3 test set [Kim and Wilbur 2010]. The performance of the PIE system was evaluated on the BC3 test set in terms of F1, MCC, and AUC iP/R1 measures [Krallinger et al. 2010]. This test set contains 910 PPI and 5090 non-PPI articles, which is unbalanced reflecting a real-world situation. The proposed system provides 0.6258 F1, 0.5610 MCC and 0.6834 AUC iP/R, whereas the median of BC3 participant results are 0.5353 F1, 0.4563 MCC and 0.5367 AUC iP/R. The PIE system significantly outperformed the other approaches on all measures.

Since PIE the search is a web-based ranking system, the performance at top-ranked articles is more important than overall classification performance. At rank 10 (P@10), 100 (P@100) and 200 (P@200), our system achieves 100, 94 and 91.50% precision, respectively. Our system also produces over 95% precision at 10% recall. Even though the ratio between PPI and non-PPI articles in the PubMed database is being updated monthly. PIE the search is already practical as a ranking system since it provides high classification performance at top-ranked articles. The web service is freely accessible and the local PubMed database in PIE is being updated monthly.

ACKNOWLEDGEMENTS

The authors would like to thank all the participants for their contribution to the manual evaluation. The authors also would like to thank Natalie Xie for valuable comments on web implementation of the system.

Funding: Intramural Research Program of the National Institutes of Health, National Library of Medicine (to S.K. and W.J.W.); Basic Science Research Program through the National Research Foundation of Korea (NRF-2011-0002437) (to D.K.).

Conflict of Interest: none declared.

REFERENCES


