DOI 10.18699/WIBSB-2018-35

Knowledge mining from large scale protein-protein interaction datasets at the era of big data

D. Li
State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Life Omics, Beijing, China
e-mail: lidong.bprc@foxmail.com

Key words: proteomics, protein-protein interaction, big data, data mining, knowledge discovery

Motivation and Aim: Throughout the history of natural science, it is definite that our knowledge and disciplines come from accumulated discoveries, which are triggered by the unprecedented scale and speed of big data and achieved by efficient mathematical strategies, taking the discovery of Kepler's laws as an example. In our laboratories, simple machine learning strategies, such as Naïve Bayesian network, have been used to find the instinct features of proteome organization, especially the protein interactions [1–4].
Methods and Algorithms: By using naïve Bayesian network, reliability was assigned to the human protein-protein interactions identified by high throughput experiments by combining multiple heterogeneous biological evidences. Then domain enrichment ratio was introduced to measure the direction between interacting proteins, resulting in an integrated human directional protein interaction network. Next, logistic regression was taken to integrate six representative features, to develop a proteome-wide prediction model of self-interacting proteins.
Results and Conclusion: Recently, we developed a naïve Bayesian classifier to combine multiple heterogeneous biological evidences to investigate the human E3-substrate interactions which determine the high specificity of ubiquitination. UbiBrowser is now provided as an integrated bioinformatics platform to predict and present the proteome-wide human E3-substrate interaction network.
It is believed that the era of big data will bring in new insights in life sciences and present new opportunities in research. Artificial intelligence strategy will play dominant roles in the coming knowledge discovery. Our colleagues are now engaged to develop an automatic knowledge discovery highway (ProDiGy), integrating the OMICS datasets and literature information into a biomedical knowledge graph (a heterogamous information network) followed by the feature extraction and deep learning for grand knowledge.

References

1. Li Y., Xie P., Lu L., Wang J., Diao L., Liu Z., Guo F., He Y., Liu Y., Huang Q., Liang H., Li D., He F. (2017) An integrated bioinformatics platform for investigating the human E3 ubiquitin ligase-substrate interaction network. Nat. Commun. 8:347.
2. Li D., Liu W., Liu Z., Wang J., Liu Q., Zhu Y., He F. (2008) PRINCESS: a PRotein INteraction Confidence Evaluation System with multiple data Sources. Mol. Cell. Proteomics. 7:1043-1052.
3. Liu Z., Guo F., Zhang J., Wang J., Lu L., Li D., He F. (2013) Proteome-wide prediction of self-interacting proteins based on multiple properties. Mol. Cell. Proteomics. 12(6):1689-1700.
4. Liu W., Li D., Wang J., Xie H., Zhu Y., He F. (2009) Proteome-wide prediction of signal flow direction in protein interaction network based on interacting domains. Mol. Cell. Proteomics. 8(9):2063-2070.