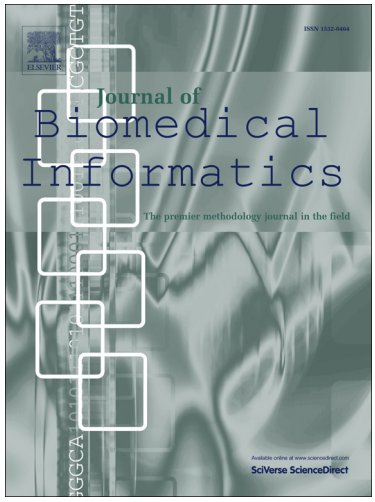

Prioritization of potential candidate disease genes by topological similarity of protein-protein interaction network and phenotype data

Jiawei Luo, Shiyu Liang

PII: S1532-0464(14)00234-2
DOI: http://dx.doi.org/10.1016/j.jbi.2014.11.004
Reference: YJBIN 2250

To appear in: *Journal of Biomedical Informatics*

Received Date: 5 May 2014
Accepted Date: 7 November 2014

Please cite this article as: Luo, J., Liang, S., Prioritization of potential candidate disease genes by topological similarity of protein-protein interaction network and phenotype data, *Journal of Biomedical Informatics* (2014), doi: http://dx.doi.org/10.1016/j.jbi.2014.11.004

Prioritization of Potential Candidate Disease Genes by Topological Similarity of Protein-Protein Interaction Network and Phenotype Data

Jiawei Luo[a*], Shiyu Liang[a]

[a] School of Information Science and Engineering, Hunan University, Changsha, China

* Corresponding author. Permanent address: School of Information Science and Engineering, Hunan University, Changsha, China 410082. Tel: 86-0731-88821971. E-mail address: luojiawei@hnu.edu.cn

Abstract: Identifying candidate disease genes is important to improve medical care. However, this task is challenging in the post-genomic era. Several computational approaches have been proposed to prioritize potential candidate genes relying on protein-protein interaction (PPI) networks. However, the experimental PPI network is usually liable to contain a number of spurious interactions. In this paper, we construct a reliable heterogeneous network by fusing multiple networks, a PPI network reconstructed by topological similarity, a phenotype similarity network and known associations between diseases and genes. We then devise a random walk-based algorithm on the reliable heterogeneous network called RWRHN to prioritize potential candidate genes for inherited diseases. The results of leave-one-out cross-validation experiments show that the RWRHN algorithm has better performance than the RWRH and CIPHER methods in inferring disease genes. Furthermore, RWRHN is used to predict novel causal genes for 16 diseases, including breast cancer, diabetes mellitus type 2, and prostate cancer, as well as to detect disease-related protein complexes. The top predictions are supported by literature evidence.

Keywords: disease genes, random walk, topological similarity, protein-protein interaction networks, phenotype

1. Introduction

Elucidating the underlying molecular mechanisms basis of diseases has become increasingly important in disease prevention, diagnosis, and drug design. Understanding the relationship