Journal of Theoretical Biology 403 (2016) 85–96

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Integration of multiple biological features yields high confidence human protein interactome

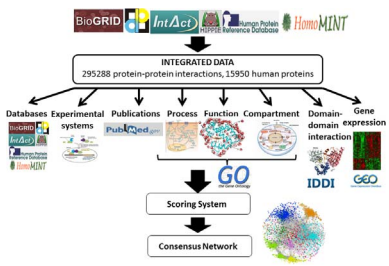
Kubra Karagoz, Tuba Sevimoglu, Kazim Yalcin Arga *

Department of Bioengineering, Marmara University, 34722 Istanbul, Turkey

HIGHLIGHTS

- The noisiness of the existing PPI data is evident, and it is essential that a more reliable data is generated.
- A highly reliable, almost complete PPI network of *Homo sapiens* is not proposed yet.
- The quality and coverage of human protein interactome need to be improved to be used in biomedicine.
- An unsupervised statistical approach was proposed to assign confidence scores to PPIs of *Homo sapiens*.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 7 February 2016
Received in revised form
1 May 2016
Accepted 11 May 2016
Available online 16 May 2016

Keywords:
Protein interaction network
Protein–protein interaction
Network topology
Cancer subnetworks

ABSTRACT

The biological function of a protein is usually determined by its physical interaction with other proteins. Protein–protein interactions (PPIs) are identified through various experimental methods and are stored in curated databases. The noisiness of the existing PPI data is evident, and it is essential that a more reliable data is generated. Furthermore, the selection of a set of PPIs at different confidence levels might be necessary for many studies. Although different methodologies were introduced to evaluate the confidence scores for binary interactions, a highly reliable, almost complete PPI network of *Homo sapiens* is not proposed yet. The quality and coverage of human protein interactome need to be improved to be used in various disciplines, especially in biomedicine. In the present work, we propose an unsupervised statistical approach to assign confidence scores to PPIs of *H. sapiens*. To achieve this goal PPI data from six different databases were collected and a total of 295,288 non-redundant interactions between 15,950 proteins were acquired. The present scoring system included the context information that was assigned to PPIs derived from eight biological attributes. A high confidence network, which included 147,923 binary interactions between 13,213 proteins, had scores greater than the cutoff value of 0.80, for which sensitivity, specificity, and coverage were 94.5%, 80.9%, and 82.8%, respectively. We compared the present scoring method with others for evaluation. Reducing the noise inherent in experimental PPIs via our scoring scheme increased the accuracy significantly. As it was demonstrated through the assessment of process and cancer subnetworks, this study allows researchers to construct and analyze context-specific networks via valid PPI sets and one can easily achieve subnetworks around proteins of interest at a specified confidence level.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

The multi-functional nature of proteins enables them to play different roles in a variety of biological processes, thus creating numerous interactions with diverse partners in distinct biological

* Correspondence to: Faculty of Engineering, Department of Bioengineering, Marmara University, Building D, Office: 404, 34722 Kadikoy, Istanbul, Turkey.
E-mail addresses: kubra.karagoz@marun.edu.tr (K. Karagoz), tubasevim@hotmail.com (T. Sevimoglu), kazim.arga@marmara.edu.tr (K.Y. Arga).

http://dx.doi.org/10.1016/j.jtbi.2016.05.020
0022-5193/© 2016 Elsevier Ltd. All rights reserved.

scenarios (Chen et al., 2014). Since the 1980s great effort has been put into understanding proteins and their interactions (Chou and Cai, 2006; Hu et al., 2011; Braun and Gingras, 2012; Jia et al., 2015a, 2015b). Protein–protein interactions (PPIs) are a fundamental integral component in the road to the comprehension of the very much tangled biological systems. Systems biomedicine makes use of interactome networks to better our understanding towards the "chaotically" organized proteins as to make inferences about disease networks that will ideally motivate us in the interpretation of their biological functions. Determination of disease genes and proteins, disease networks and subnetworks and designation of function to new proteins are some of the benefits of a well-established and reliable human interactome (Huang and Chen, 2011; Xiao et al., 2013, 2015). Sevimoglu and Arga (2014) have compiled a list of studies regarding the role of PPI networks in systems biomedicine.

The massive amount of interaction data obtained from high-throughput experiments (e.g. yeast two-hybrid (Y2H) and mass spectrometry (MS)) stemming from the approximately 30,000 genes identified by the Human Genome Project (www.genome.gov) is collected, integrated and stored in various PPI databases. BIND (Biomolecular Interaction Network Database) acquires information about interactions, molecular complexes and pathways and accepts interactions with as little as one publication reference to PubMed (Bader et al., 2001). HomoMINT is an online database for inferred network of protein interactions where interactions discovered in model organisms are mapped onto the corresponding human orthologs. The InParanoid algorithm in conjunction with a string matching algorithm is used to construct the HomoMINT network, and the inferred network is compared to an experimental interactome as well as a network formed by literature survey for validation (Persico et al., 2005). BioGRID (the Biological General Repository for Interaction Datasets) includes interaction data from high-throughput experiments such as Y2H for binary interactions and MS analysis for purified complexes. To overcome the problem of false positives and negatives they have also included literature-derived interactions from PubMed (Chatr-Aryamontri et al., 2015). HitPredict database calculates a comprehensive reliability score using experimental data together with features of interacting proteins (Lopez et al., 2015). Another online database, HINT (High-quality INTeractomes) uses a filter system to eliminate the low-quality interactions by considering the number of experimental studies carried out to map these interactions as well as the number of interactions reported by each unique publication. They rank the publications in decreasing order of interactions reported per study (Das and Yu, 2012). HIPPIE (Human Integrated Protein–Protein Interaction Reference) is a human PPI dataset that uses a normalized scoring scheme integrating various experimental datasets (Schaefer et al., 2012). STRING is also an online database that uses confidence scores from three genomic associations (gene fusion, co-occurrence of proteins in a metabolic map and neighborhood evidence) to rank the reliability of protein interactome. This database maps interactions based on experimental studies as well as prediction algorithms (Szklarczyk et al., 2015). Other publicly available PPI databases that collect and store interaction data include HPRD (the Human Protein Reference Database) (Prasad et al., 2009), DIP (the Database of Interacting Proteins) (Salwinski et al., 2004), and IntAct (the IntAct molecular interaction database) (Orchard et al., 2014).

Research efforts towards achieving reliable PPI data have focused on combining multiple heterogeneous evidence to assign reliability to the interactions. This evidence mostly included biological features such as functional annotation and gene expression, and text-mining results. On the other hand, various graphic approaches have been successfully used to analyze many important biological problems, such as enzyme-catalyzed reactions (Jiang et al., 1979; Chou, 1989; Zhou and Deng, 1984), protein folding kinetics and folding rates (Chou, 1990), inhibition of HIV-1 reverse transcriptase (Althaus et al., 1993; Gonzales et al., 1993), non-steady drug metabolisms systems (Chou, 2010), evolution of biological sequences (Wu and Xiao, 2010a), and protein–protein interactions using wenxiang diagrams (Lin and Xiao, 2011; Zhou, 2011). Therefore, graph theoretical network topology analyzes were also employed in several interaction scoring schemes. Patil and Nakamura (2005) used a combination of three genomic features (domain–domain interactions, Gene Ontology annotations, and sequence homology). Chen et al. (2006) used network topological metrics to increase the confidence of protein interactions in yeast, fruit fly, and worm. Wu et al. (2010b) proposed a two-step method to integrate biological and computational sources to achieve reliable PPIs. First, they grouped PPIs to estimate the confidence scores for each biological or computational evidence source then classified these interactions to identify reliable interactions. Their proposed method eliminated false positives in detected PPIs and identified false negatives by predicting novel PPIs. In another study, a functional protein interaction network for humans was constructed by integrating biological and computational sources of information, such as gene co-expression, protein–domain interactions, and text-mined protein interactions (Wu et al., 2010c). Chen et al. (2014) adapted a phylogenic grouping method in conjunction with force-directed graph simulation to dissect the human interactome that enabled them to associate network properties with evolutionary and biological properties. User-friendly and publicly available web tools, such as PRINCESS (Li et al., 2008), PSISCOREweb (Aranda et al., 2011), INTSCORE (Kamburov et al., 2012) and HitPredict (Lopez et al., 2015) were also presented, which combine model organism protein–protein interaction, interaction domain, functional annotation, gene expression, genome context, and network topology structure, to assign reliability to the user-specified sets of interactions.

Even though the PPIs mapped by the above mentioned and other databases are increasing rapidly, there are still problems regarding coverage, sensitivity, and specificity that are essential for completeness and reliability of the PPI data. False positives and negatives due to laboratory bias such as contaminating nucleic acids (Nguyen and Goodrich, 2006) and prey proteins that appear to co-purify specifically with the bait (Banks et al., 2015) are still an issue. Computational PPI data integration methods should include several criteria to measure the contributions of such biases (Gillis et al., 2014). Not all the interactome databases use a scoring system to map their interactions hence the specificity and sensitivity as well as the coverage of the data are unknown. Moreover, the amount of data and content also differs from one database to another. Additionally, the interactome data from different sources are usually non-comparable.

To get a more comprehensive and reliable interactome one should combine the available databases and also use a scoring scheme that includes a wide variety of biological features (Karagoz and Arga, 2013). A more reliable human protein interactome is needed for a better understanding of the global human PPIs as well as disease networks such as different types of cancer and functional pathways (Sinha et al., 2016; Calimlioglu et al., 2015; Sevimoglu and Arga, 2015; Karagoz et al., 2016). In the present study, a high confidence human interactome is achieved by incorporating eight biological features (e.g. experimental system and domain interactions) in a scoring scheme with a large group of PPIs obtained by compiling data from various human protein interaction databases. To reveal the performance of the reconstructed network, comparison with other protein–protein interaction networks under a comparative analysis based on Receiver Operating Characteristic (ROC) curve as well as enrichment analysis and construction of subnetworks of cancer-specific PPIs are performed.

Download English Version:

https://daneshyari.com/en/article/6369146

Download Persian Version:

https://daneshyari.com/article/6369146

Daneshyari.com