Gene 518 (2013) 78–83

Contents lists available at SciVerse ScienceDirect

Gene

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

Methods Paper

# PRASA: An integrated web server that analyzes protein interaction types

Chen-Yu Fan, Yi-Han Bai, Cheng-Yi Huang, Tsung-Ju Yao, Wen-Hao Chiang, Darby Tien-Hao Chang *

*Department of Electrical Engineering, National Cheng Kung University, Tainan 70101, Taiwan*

ARTICLE INFO

Available online 28 December 2012

*Keywords:*
Interaction type
Machine learning
Protein–protein interaction
Web server

ABSTRACT

This work presents the Protein Association Analyzer (PRASA) (http://zoro.ee.ncku.edu.tw/prasa/) that predicts protein interactions as well as interaction types. Protein interactions are essential to most biological functions. The existence of diverse interaction types, such as physically contacted or functionally related interactions, makes protein interactions complex. Different interaction types are distinct and should not be confused. However, most existing tools focus on a specific interaction type or mix different interaction types. This work collected 7234058 associations with experimentally verified interaction types from five databases and compiled individual probabilistic models for different interaction types. The PRASA result page shows predicted associations and their related references by interaction type. Experimental results demonstrate the performance difference when distinguishing between different interaction types. The PRASA provides a centralized and organized platform for easy browsing, downloading and comparing of interaction types, which helps reveal insights into the complex roles that proteins play in organisms.



## 1. Introduction

Various protein interactions determine the outcome of most biological processes in living cells. Identifying and characterizing these protein–protein interactions (PPIs) and their interaction networks helps us understand the mechanisms driving biological processes at the molecular level (Shoemaker and Panchenko, 2007). Considerable effort has been expended to construct comprehensive PPI collections and to predict novel PPIs. Many databases, such as the Database of Interacting Proteins (DIP) (Salwinski et al., 2004), Molecular Interaction (MINT) (Ceol et al., 2010), IntAct (Aranda et al., 2010), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2012) and BioGRID (Stark et al., 2011), have been developed to collect experimentally reported PPIs. Some databases integrate different sources, such as Agile Protein Interaction DataAnalyzer (APID) (Prieto and De Las Rivas, 2006) and Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (Szklarczyk et al., 2011), into a single platform. Furthermore, the low coverage of experimentally verified interactions in the complete interactome calls for the development of computational methods that generate complementary data for experimental techniques (Han et al., 2005; Hart et al., 2006). Various PPI prediction methods based on localization data (Pellegrini et al., 1999), expression data (Marcotte et al., 1999; Soong et al., 2008) or orthologous data (Espadaler et al., 2005; Huang et al., 2004) have been developed. Recently, *de novo* methods have garnered significant attention because they require only protein primary sequences (Chang et al., 2010; Shen et al., 2007; Yu et al., 2010). In addition to protein sequence, some studies showed that promoter sequences could be potentially good features in developing *de novo* methods (Chang et al., 2011, 2012a).

However, the existence of diverse types of protein interactions makes PPI complex. The most common definition for PPI is two proteins in physical contact (De Las Rivas and Fontanillo, 2010). Such PPI can be identified by analyzing protein complexes (Chang et al., 2012b). Genetic associations – simultaneous mutations of two genes results in a phenotype different from that of individual mutations – have also been well studied (Mani et al., 2008). Two proteins participating in the same biological process have been commonly seen in co-evolution studies (Enault et al., 2003; Snitkin et al., 2006). These interaction types are quite distinct and should not be confused with one another (De Las Rivas and Fontanillo, 2010).

To predict protein interactions as well as interaction types, this work presents the Protein Association Analyzer (PRASA). The term *association*, which stands for the existence of interactions along with interaction type in the PRASA, is used to prevent any confusion with a conventional definition of a single interaction type. The PRASA accommodates three interaction types: two proteins i) are in physical contact, ii) whose simultaneous mutations result in a phenotype that differs from that of the individual mutations and iii) participate in the same pathway. This work collected 7234058 associations with experimentally verified interaction types from five databases. These known associations were used to train a *de novo* PPI predictor previously developed by the authors (Yu et al., 2010). This predictor performs well for human physical interactions. This work refined the prediction workflow of the *de novo* PPI predictor for interaction type prediction.

*Abbreviations:* PPI, protein–protein interaction; PSI-BLAST, Position-Specific Iterative Basic Local Alignment Search Tool; RVKDE, relaxed variable kernel density estimation.

* Corresponding author. Tel.: +886 6 2757575x62421; fax: +886 6 2345482.
*E-mail address:* darby@mail.ncku.edu.tw (D.T.-H. Chang).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2012.11.083

Several experiments were conducted to investigate the prediction performance for different interaction types.

Given a set of query protein pairs, the PRASA generates predictions based on the *de novo* PPI predictor and performs a literature search of experimentally verified data. The web server of the PRASA has been carefully designed, such that users can easily identify the differences between interaction types as many important observations are revealed by considering only these differences. For example, a comparison between the canonical pathway and an interaction network derived from physical interactions helps to understand the complex functional roles that proteins play in biological systems (De Las Rivas and Fontanillo, 2010). In this regard, the PRASA provides a centralized and organized platform for easy browsing, downloading and comparing interaction types, which may reveal additional insights into the complex roles that proteins play in organisms.

## 2. Methods

Analysis by the PRASA is based on a collection of protein associations that are experimentally verified. This section describes the definition and collection of different interaction types by the PRASA, followed by processes of analyzing query protein pairs based on experimental data.

### 2.1. Data collection

The PRASA addresses three protein interaction types: i) physical association, which indicates that two proteins are in physical contact; ii) genetic association, which indicates that two proteins co-work for a phenotypic effect; and iii) pathway association, which indicates that two proteins participate in the same biological process. The first two association types are in accordance with the Proteomics Standards Initiative-Molecular Interactions (PSI-MI) standard (Hermjakob et al., 2004). Physical associations are identified by specific experimental systems, such as the yeast two-hybrid system (Fields and Song, 1989), in which the bait must physically capture the prey. Genetic associations occur when two genetic variations, such as mutations or over-expressions, result in a composite phenotypic effect that is not activated by either individual variation. The pathway associations represent a relatively distant relationship, namely, proteins with such associations are necessary to a biological process, but may not be required at the same time. Pathway association is not conventionally considered a protein interaction. The PRASA includes such an association because comparisons between pathways and interaction networks helps to understand biological systems (De Las Rivas and Fontanillo, 2010). The physical and genetic associations in the PRASA were collected from DIP (Salwinski et al., 2004), MINT (Ceol et al., 2010), IntAct (Aranda et al., 2010) and BioGRID (Stark et al., 2011). The pathway associations of the PRASA were collected from KEGG (Kanehisa et al., 2012). Any pair of proteins in the same pathway is defined as having a pathway association. Two proteins may have multiple pathway associations when they jointly participate in more than one pathway. The protein primary sequences in the FASTA format, which are required for analyzing processes, were collected from UniProt (Apweiler et al., 2010), a high-quality database of protein sequences and functional annotations. The UniProt identifier (i.e., entry name) was adopted as the protein identifier of the PRASA to integrate DIP, MINT, IntAct, BioGRID, KEGG and UniProt data. Associations involving proteins without entry names were excluded.

### 2.2. Association analysis

Association prediction by the PRASA is based on the recently proposed *de novo* PPI predictor (Yu et al., 2010). This predictor encodes each protein pair as a feature vector based on a probability-based mechanism. A training set of protein pairs, including positive (interacting protein pairs) and negative (non-interacting protein pairs), is submitted to the relaxed variable kernel density estimator (RVKDE) to construct the prediction model (Oyang et al., 2005). The prediction model of the RVKDE contains two probability density functions, say, $f_{pos}$ and $f_{neg}$, over the vector space of the encoded feature. A query protein pair, encoded as feature vector $\mathbf{v}$, is predicted as interacting when $f_{pos}(\mathbf{v}) > f_{neg}(\mathbf{v})$. The details of feature encoding and the RVKDE can be found in (Yu et al., 2010). The following descriptions characterize the implementation modifications applied on the predictor to fit the aim of the PRASA—to predict interaction types.

Instead of preparing only a single prediction model of only positive and negative probability density functions, the PRASA uses the RVKDE to construct three individual prediction models for physical, genetic and pathway association prediction. Namely, the PRASA constructed six probability density functions, $f_{phy}$, $f_{!phy}$, $f_{gen}$, $f_{!gen}$, $f_{pat}$, and $f_{!pat}$, for physical association, no physical association, genetic association, no genetic association, pathway association and no pathway association. The symbol "!" is borrowed from C language and stands for "not." A query protein pair $\mathbf{v}$ is reported as having a physical association when $f_{phy}(\mathbf{v}) > f_{!phy}(\mathbf{v})$, without considering the other four probability density functions. The same rule is applied to predict genetic and pathway associations. The RVKDE can handle multi-class prediction. However, as will be shown in the Performance section, using three independent binary predictors has some advantages over using a single multi-class prediction model.

## 3. Web interface

Users can specify a set of protein pairs in the PRASA. Proteins can be specified by name or by their sequences in the FASTA format. When the organism is selected as "Auto-detection," the PRASA uses PSI-BLAST to align the first query protein to all collected proteins and chooses the protein organism with the highest bit score. The PRASA outputs two pages (Fig. 1). Users are first directed to the "Result" page, which includes query information specified by the user (Fig. 1a), an interactive network view (Fig. 1b) and a table view (Fig. 1c) of predicted associations. The network view (Fig. 1b) uses Cytoscape Web version 1.0 (Lopes et al., 2010), which provides basic functions such as dragging and zooming. The nodes represent proteins. By clicking the nodes, a user is directed to the description page of the UniProt database. The edges, which represent associations, have three styles: i) blue solid lines indicate associations predicted by the PRASA and with literature support; ii) blue dotted lines indicate associations predicted by the PRASA without literature support; and iii) red solid lines indicate associations with literature support but not predicted by the PRASA. The number over an edge indicates the number of references related to a corresponding association. Associations with literature support are those that have at least one reference. Clicking an edge directs a user to the "Pair" page, which provides the basic information of the corresponding protein pair (Fig. 1d) and related references (Fig. 1e). Users can choose specific associations, such as a predicted physical interaction or genetic interactions with literature support, to be shown in the network view. The table view (Fig. 1c) of the "Result" page provides the same information as the network view, allowing a user to see both the prediction and number of related references of different interaction types. Clicking protein names directs users to UniProt, while clicking the number of related references directs users to the "Pair" page. Finally, the bottom of the table view has a "Download" link. Clicking this link directs users to a text representation of the table (Fig. 1g). This allows users to download the corresponding information for further manipulation.

## 4. Performance

### 4.1. Experimental design

Prediction performance of the PRASA was evaluated using human and yeast protein interactions (Table 1). Based on collected associations