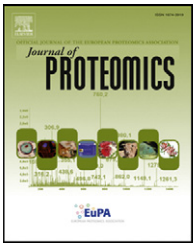
Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot



Extracting high confidence protein interactions from affinity purification data: At the crossroads☆

Shuye Pu[a,*], *James Vlasblom*[a,c], *Andrei Turinsky*[a], *Edyta Marcon*[d], *Sadhna Phanse*[d], *Sandra Smiley Trimble*[b], *Jonathan Olsen*[b,d], *Jack Greenblatt*[b,d], *Andrew Emili*[b,d], *Shoshana J. Wodak*[a,b,c,**]

[a] *Hospital for Sick Children, 555 University Avenue, Toronto, Ontario M4K 1X8, Canada*
[b] *Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada*
[c] *Department of Biochemistry, University of Toronto, Toronto, Ontario, Canada*
[d] *Banting and Best Department of Medical Research, University of Toronto, Donnelly Centre for Cellular and Biomolecular Research, 160 College Street, Toronto, Ontario M5S 3E1, Canada*

ARTICLE INFO

Available online 14 March 2015

Keywords:
Protein–protein interaction
Scoring methods
Affinity purification
Mass spectrometry

ABSTRACT

Deriving protein–protein interactions from data generated by affinity-purification and mass spectrometry (AP–MS) techniques requires application of scoring methods to measure the reliability of detected putative interactions. Choosing the appropriate scoring method has become a major challenge. Here we apply six popular scoring methods to the same AP–MS dataset and compare their performance. The comparison was carried out for six distinct datasets from human, fly and yeast, which focus on different biological processes and differ in their coverage of the proteome. Results show that the performance of a given scoring method may vary substantially depending on the dataset. Disturbingly, we find that the high confidence (HC) PPI networks built by applying the six scoring methods to the same raw AP–MS dataset display very poor overlap, with only 1.7–4.1% of the HC interactions present in all the networks built, respectively, from the proteome-wide human, fly or yeast datasets. Various properties of the shared versus unique interactions in each network, including biases in protein abundance, suggest that current scoring methods are able to eliminate only the most obvious contaminants, but still fail to reliably single out specific interactions from the large body of spurious associations detected in the AP–MS experiments.

Biological significance

The fast progress in AP–MS techniques has prompted the development of a multitude of scoring methods, which are relied upon to remove contaminants and non-specific binders. Choosing the appropriate scoring scheme for a given AP–MS dataset has become a major challenge. The comparative analysis of 6 of the most popular scoring methods, presented here, reveals that overall these methods do not perform as expected. Evidence is provided that this is due to 3 closely related issues: the high 'noise' levels of the raw AP–MS data, the

☆ This article is part of a Special Issue entitled: Protein dynamics in health and disease. Guest Editors: Pierre Thibault and Anne-Claude Gingras.
* *Corresponding author.*
** *Correspondence to:* S.J. Wodak, VIB Structural Biology Centre, VUB, Building E, Pleinlaan 2, 1050 Brussels, Belgium.
E-mail addresses: shuye2009@gmail.com (S. Pu), shoshana@sickkids.ca (S.J. Wodak).

http://dx.doi.org/10.1016/j.jprot.2015.03.009

limited capacity of current scoring methods to deal with such high noise levels, and the biases introduced using Gold Standard datasets to benchmark the scoring functions and threshold the networks. For the field to move forward, all three issues will have to be addressed.

This article is part of a Special Issue entitled: Protein dynamics in health and disease. Guest Editors: Pierre Thibault and Anne-Claude Gingras

1. Introduction

Affinity purification–mass spectrometry (AP–MS) has become one of the dominant experimental approaches for high-throughput analyses of protein–protein interactions (PPIs) and protein complexes [1–5]. With the improved detection sensitivity of MS instruments, the number of hit proteins (preys) that co-purify with the target proteins (baits) and can be detected has increased significantly. However, a sizeable fraction of these preys represent spurious binders that engage in non-specific interactions [6,7]. In order to filter out such spurious interactions, scoring methods are used to estimate the reliability of individual associations, a quantity often considered as related to their specificity. These estimates are then benchmarked against a reference set of reliable known interactions (the so-called 'Gold Standard') and used to derive the final high confidence (HC) network that contains only PPI of an acceptable reliability level.

Recently, several computational methods have been proposed for assigning reliability or confidence scores to associations detected in proteomic studies [8–13]. These scoring methods vary in many aspects (see [14] for an extensive review). Some methods consider only bait–prey interactions (spoke model), others take into account both bait–prey and prey–prey interactions (matrix model). These methods have been developed in studies that employ diverse experimental protocols and probe association landscapes that vary in coverage of the proteome, in the binding propensities of the corresponding proteins, and in the overall quality of the raw datasets. A new scoring method is usually developed on a single AP–MS dataset, and the same dataset is commonly used to compare its performance to those of extant methods. It is often unclear, therefore, if a given method shown to outperform others on a specific dataset also performs well on other datasets. Faced with a newly derived dataset, experimentalists are therefore often unable to make an informed choice about the scoring methods that is best suited for processing their data.

To address this gap, we compare the performance of six of the most popular scoring methods, with an emphasis on recently devised methods that incorporate spectral counts. We analyze 3 such methods, including the Comparative Proteomic Analysis Software Suite (ComPASS) [12], the Significant Analysis of Interactome (SAINT) [8] and the Hypergeometric Spectral Counts score (HGSCore) [10]. Spectral counts are a semi-quantitative measure of protein abundance in samples [15] and their incorporation may therefore improve the reliability measure encoded in the score. On the other hand, spectral counts can be affected by inefficient protein digestion and peptide ionization [16]. Their incorporation might thus also have deleterious effects on the scoring method. To obtain a more general picture we also analyzed 3 popular scoring methods that do not utilize spectral counts: the Purification Enrichment (PE) [9], Dice Coefficient (Dice) [13], and the Hart score (Hart) [11].

We did not evaluate several other published scoring methods, such as Mass Spectrometry Interaction Statistics (MiST) [17], Decontaminator [18], Socio-affinity score (SA) [19], Improved Socio-affinity score (ISA) [20] and Interaction Detection Based on Shuffling (IDBOS) [21]. This choice was due either to the lack of proper data (e.g., MiST requires protein intensity data and large number of replicates, Decontaminator requires multiple control purifications to build a model of contaminants and uses Mascot scores as input) or to the similarity of the methods to one of those we chose to evaluate (e.g., SA is a simpler variant of the PE score, ISA and IDBOS are similar to the more widely used SAINT and ComPASS in their use of the Spoke model).

For the purpose of the present study all 6 methods were applied respectively to 6 different published raw AP–MS datasets, all of which had available spectral count data (Table 1). For the majority of the datasets (5 out of 6) a single scoring method was developed or applied by the authors to produce the final HC protein–protein interaction network. Here, all 6 methods were applied to each of the 6 datasets. Receiver Operating Characteristic (ROC) curve analysis was employed to benchmark the interactions scored by each method against literature-curated high confidence PPIs used as a reference (Gold Standard). These reference sets were retrieved from iRefWeb, a web resource for consolidated protein–protein interactions [22]. Since interacting protein pairs frequently share functions and/or cellular localizations, the similarity of the Gene Ontology (GO) [23] annotations of interacting pairs was used as an additional validation criterion.

Our study describes the most thorough comparison to date of HC confidence networks derived by applying different scoring methods to the same dataset. This comparison also evaluates the extent to which the performance of different methods changes with the dataset to which they are applied. It demonstrates the importance of choosing a scoring method that is appropriate for the dataset at hand, and confirms the determinant role that features of the raw experimental AP–MS data themselves play in shaping the end result. By far the most important observation we make is that HC interactions derived by different scoring methods from the exact same raw dataset display very limited overlap. The poor overlap of HC interaction networks derived for the same organism by different experimental techniques has been previously documented [24–26] Concerns have been raised that this may stem from the fact that the derived HC networks still incorporate a non-negligible fraction of spurious (non-specific) interactions, possibly because scoring methods may be less effective than expected [27,28], especially when they are applied to noisy datasets. Our