BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 23 2014, pages 3419–3420
doi:10.1093/bioinformatics/btu517

Gene expression

Advance Access publication August 18, 2014

PEPPER: cytoscape app for protein complex expansion using protein–protein interaction networks

C. Winterhalter[1,2,†], R. Nicolle[1,3,†], A. Louis[1], C. To[1], F. Radvanyi[3] and M. Elati[1,*]

[1]iSSB, CNRS, University of Evry, Genopole, 5 rue H. Desbruères, 91030 Evry Cedex, France, [2]School of Computing Science, Newcastle University, Newcastle NE1 7RU, UK and [3]UMR 144 CNRS/Institut Curie, 26 rue d'Ulm, Paris, 75248 cedex 05, France

Associate Editor: Janet Kelso

ABSTRACT

We introduce PEPPER (Protein complex Expansion using Protein–Protein intERactions), a Cytoscape app designed to identify protein complexes as densely connected subnetworks from seed lists of proteins derived from proteomic studies. PEPPER identifies connected subgraph by using multi-objective optimization involving two functions: (i) the coverage, a solution must contain as many proteins from the seed as possible, (ii) the density, the proteins of a solution must be as connected as possible, using only interactions from a proteome-wide interaction network. Comparisons based on gold standard yeast and human datasets showed Pepper's integrative approach as superior to standard protein complex discovery methods. The visualization and interpretation of the results are facilitated by an automated post-processing pipeline based on topological analysis and data integration about the predicted complex proteins. PEPPER is a user-friendly tool that can be used to analyse any list of proteins.

Availability: PEPPER is available from the Cytoscape plug-in manager or online (http://apps.cytoscape.org/apps/pepper) and released under GNU General Public License version 3.

Contact: mohamed.elati@issb.genopole.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 20, 2014; revised on July 1, 2014; accepted on July 24, 2014

1 INTRODUCTION

Most cellular processes require a large number of proteins to assemble into functional complexes to perform their activity. Therefore, describing functional protein complexes taking part in given processes is critical to the underlying molecular mechanism understanding. Experimental protocols such as Affinity Purification followed by Mass-Spectrometry (AP-MS) have been devised to *pull down* a protein of interest (*bait*) together with all the interacting proteins within the same protein complex (*preys*). However, these sets of *preys* may contain both false positives, proteins detected despite not actually interacting with the *bait*, and omit false negatives (Gingras *et al.*, 2007), proteins interacting in the cellular context studied but not detected. Effective control experiments and usage of contaminants repositories can remove some false positives. However, false negative interacting partners identification, thereby the definition of the entire protein complex, remains challenging. Protein–Protein Interaction (PPI) data represents abundant information that can be used for this purpose.

Protein complexes extraction from PPI networks is a very active area of research and many methodologies have been developed to tackle this problem. These computational methods generally model protein complexes as dense subnetworks within the complete set of PPIs and thus try to solve a graph clustering problem or to identify dense regions. Clustering approaches were shown to be efficient either on large PPI networks or with large-scale experimental settings in which big numbers of *baits* result in context-specific PPI networks (Bader and Hogue, 2003; Nepusz *et al.*, 2012). However, these algorithms were not developed for use in small-scale AP-MS experiments (*e.g.* using only a single *bait* protein) and are unable to integrate experimental data with repositories of PPI.

We reasoned that although not all the protein partners may be detected in a given AP-MS experiment, these proteins may have been previously identified as interacting with either the *bait* or some of the *preys* of the experiment. Based on this hypothesis, we developed PEPPER, which addresses the problem of finding protein complexes by combining the experimental results of a single AP-MS assay with the available information from protein interactions in a global PPI network. PEPPER solves this non-trivial problem by using a multi-objective evolutionary algorithm (Elati *et al.*, 2013), which was tested to demonstrate the relevance of our integrative approach. To do so, we used publicly available AP-MS datasets for yeast and human species and compared PEPPER's results with those of state-of-the-art protein complex discovery methods. Our findings highlight the relevance of integrating PPI repositories to the analysis of AP-MS experiments. We propose PEPPER as a Cytoscape application to further refine protein complex predictions through functional and topological analyses.

2 METHODS AND IMPLEMENTATION

In the context of a single AP-MS experiment, PEPPER aims to identify a dense subnetwork within the PPI network connecting as many of the proteins identified in this experiment as possible, referred to hereafter as the list of *seed* proteins. PEPPER solves this problem by maximizing two objective functions: (i) coverage, a solution must contain as many proteins from the *seed* protein list as possible; (ii) density, a solution must

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

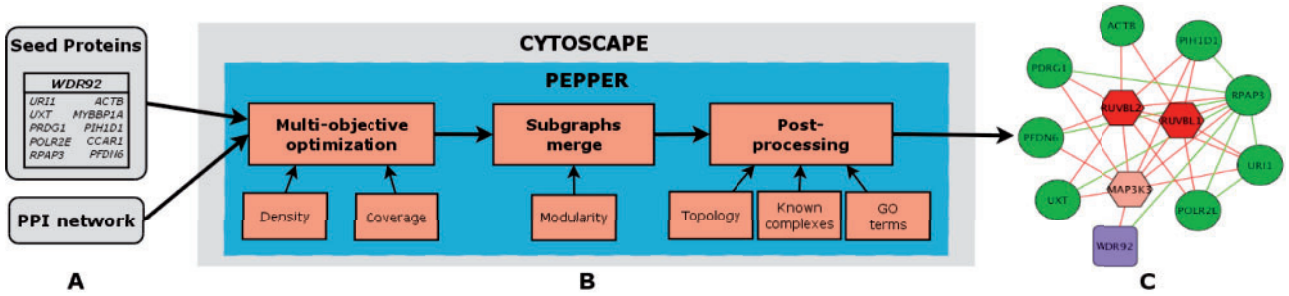

Fig. 1. Schematic representation of the plug-in. (**A**) Example of input data, a large-scale PPI and the results of an AP-MS experiment with the *bait* and a list of *prey* proteins. (**B**) Context-specific protein complex extraction pipeline. (**C**) Output subnetwork representing a putative protein complex using only interactions from the input PPI network: example of *WDR92*. Purple squares and green circles correspond to *bait* and *prey* proteins, respectively. Hexagons indicate the expansions proposed by PEPPER and are shown in various shades of red, according to their post-processing score. Dark red indicates a high predicted relevance to the solution. The edges shown in the graph are exclusively those found in the input PPI network. Green edges are set between seed proteins. All edges involving an expansion protein are red

contain as many interactions as possible. These objectives are often conflicting, and thus, no single solution can be considered to dominate over the others. Instead, the optimal solution is a Pareto optimal set with multiple solutions. SPEA2 (Zitzler *et al.*, 2001), a popular Multi-Objective Evolutionary Algorithm, is used for the simultaneous optimization of the two objective functions and to identify solutions approximating the set of pareto-optimal solutions. These solutions are merged into a final predicted protein complex by maximizing the modularity with a greedy search (see *SI algorithm* section).

PEPPER was developed as a Cytoscape application, which uses a *seed* list of proteins and a large-scale PPI network as inputs (Fig. 1A). In addition to the aforementioned subnetwork extraction procedure, PEPPER includes a topological and function-based post-processing pipeline for ranking the added proteins (*expansions*) according to their relevance (Fig. 1B). The predicted complex and each of the proteins are annotated based on their cellular localization or function annotation specificity. Enrichment analysis is complemented by matching the solutions to a collection of reference protein complexes, and *expansions* are scored according to their co-occurrence with the *seeds* in these complexes. Topological scoring is based on the impact of the *expansions* on the overall connectivity of the subnetwork (see *SI post-processing* section). PEPPER uses these scores to rank *expansions* and to facilitate results visualization and interpretation (Fig. 1C).

3 CASE STUDY

We assessed the performance of PEPPER and two network clustering algorithms for protein complex discovery—MCODE (Bader and Hogue, 2003) and ClusterONE (Nepusz *et al.*, 2012)—on a benchmark dataset of 135 yeast and 9 human single-bait AP-MS experiments and using a set of hand-curated protein complexes as gold standards. For network clustering methods, performance was assessed for each AP-MS experiment by selecting the predicted complex which best matched the *seed* (details in *SI performance comparison* section). For each experiment, the reference complex from the gold standard best matching the *seed* was used as the ground truth in a binary classification task. Compared with both of the clustering methods tested, the complexes predicted by PEPPER scored higher in all of the performance measures for both organisms (details in *SI performance comparison* section) with notably an average increase of 16% of the geometric accuracy in human and 12% in yeast.

As an example, we describe here the results obtained for the human WDR92 protein. In the initial list of *preys*, WDR92 was identified as interacting with only one protein. PEPPER expanded the *seed* with three new proteins (Fig. 1C) and greatly increased the overall density of the original solution (22 to 47%). The new *expansion* proteins were ordered on the basis of post-processing score. The first two proteins, RUVBL1 and RUVBL2, have both a high topological and Gene Ontology score. The lower scored protein, MAP3K3, still remains relevant according to its high topological score (connected to >90% of the predicted complex proteins). AP-MS experiments using RUVBL1 or RUVBL2 as *baits* both identified WDR92 as a *prey* protein (Choi *et al.*, 2010). Moreover, in the raw WDR92 experimental data, the set of *preys* with lower processing scores (based on peptide counts) than the threshold contains RUVBL1 (see *SI Case study* section). Thus, the application of PEPPER to this experiment led to the recovery of proteins that would not have been identified otherwise (potential false negatives).

Overall, these results demonstrate the feasibility of expanding the protein complexes identified in an AP-MS experiment through the use of PPI networks and the value of PEPPER for this purpose.

Funding: This work was supported by the French National Cancer Institute (INCa_2960: PLBIO10) and the European Union/Framework Programme 7/2009 ("SYSCILIA" consortium, grant 241955). Funding for open access charge: SYSCILIA.

Conflict of interest: none declared.

REFERENCES

Bader,G.D. and Hogue,C.W.V. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.

Choi,H. *et al.* (2010) Saint: probabilistic scoring of affinity purification-mass spectrometry data. *Nat. Methods*, **8**, 70–73.

Elati,M. *et al.* (2013) Multi-objective optimization for relevant sub-graph extraction. In: Nicosia,G. and Pardalos,P. (eds) *Learning and Intelligent Optimization (LION'7)*, LNCS. Vol. 7997, Springer, Berlin Heidelberg, pp. 104–109.

Gingras,A.-C. *et al.* (2007) Analysis of protein complexes using mass spectrometry. *Nat. Rev. Mol. Cell Biol.*, **8**, 645–654.

Nepusz,T. *et al.* (2012) Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods*, **9**, 471–472.

Zitzler,E. *et al.* (2001) SPEA2: Improving the strength Pareto evolutionary algorithm. *Technical report*, Athens, Greece.