Bioinformatics, 2021, 1–3
doi: 10.1093/bioinformatics/btab022
Advance Access Publication Date: 20 January 2021
Applications Note

Systems biology

PrInCE: an R/Bioconductor package for protein–protein interaction network inference from co-fractionation mass spectrometry data

Michael A. Skinnider[1], Charley Cai[1], R. Greg Stacey [1] and Leonard J. Foster [1,2,*]

[1]Michael Smith Laboratories, University of British Columbia, Vancouver, BC V6T 1Z4, Canada and [2]Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, BC V6T 1Z3, Canada

*To whom correspondence should be addressed.

Associate Editor: Lenore Cowen

Received on October 2, 2020; revised on January 3, 2021; editorial decision on January 5, 2021; accepted on January 8, 2021

Abstract

Summary: We present PrInCE, an R/Bioconductor package that employs a machine-learning approach to infer protein–protein interaction networks from co-fractionation mass spectrometry (CF-MS) data. Previously distributed as a collection of Matlab scripts, our ground-up rewrite of this software package in an open-source language dramatically improves runtime and memory requirements. We describe several new features in the R implementation, including a test for the detection of co-eluting protein complexes and a method for differential network analysis. PrInCE is extensively documented and fully compatible with Bioconductor classes, ensuring it can fit seamlessly into existing proteomics workflows.

Availability and implementation: PrInCE is available from Bioconductor (https://www.bioconductor.org/packages/devel/bioc/html/PrInCE.html). Source code is freely available from GitHub under the MIT license (https://github.com/fosterlab/PrInCE). Support is provided via the GitHub issues tracker (https://github.com/fosterlab/PrInCE/issues).

Contact: foster@msl.ubc.ca

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

1 Introduction

Many biological functions are carried out by complex and dynamic networks of interacting proteins. Mapping the complete network of protein–protein interactions (PPIs)—the 'interactome'—has been a central objective of high-throughput biology. Traditionally, efforts to this end relied on labor-intensive techniques such as yeast two-hybrid (Y2H) and affinity purification-mass spectrometry (AP-MS). More recently, CF-MS has emerged as a powerful strategy for interactome mapping based on fractionation of protein complexes according to their biophysical properties, followed by quantitative proteomic analysis of each fraction. Key advantages of CF-MS include its high throughput, its power to map interactomes in their native cellular or physiological contexts, and its ability to monitor interactome rearrangements in response to stimulation.

Over the past eight years, we have distributed and maintained collections of freely available Matlab scripts for CF-MS data analysis (Kristensen *et al.*, 2012; Scott *et al.*, 2015). In 2017, we refined and expanded these scripts into a supervised machine-learning pipeline for PPI prediction from CF-MS data, forming the first release of PrInCE (Stacey *et al.*, 2017). However, the distribution of this workflow in Matlab—a closed-source, commercial language—represented a barrier to wider uptake. Here, we present an open-source implementation of PrInCE in the R programming language, distributed through the Bioconductor software project. This new release of PrInCE is substantially faster and more lightweight, and includes several new functionalities that enable new avenues of CF-MS data analysis.

2 The PrInCE R package

PrInCE employs a supervised classification approach to infer PPIs from CF-MS data. Briefly, after fitting a mixture of Gaussians to each chromatogram and discarding low-quality profiles, a set of five features is calculated for each potential interacting protein pair (Fig. 1a). These features are provided as input to a machine-learning classifier, alongside a set of reference protein complexes. Protein pairs are ranked by their mean classifier score in ten-fold cross-validation, to avoid data leakage for complexes in the training set, and a precision–recall curve is calculated. The complete ranked list of all protein pairs can be subset to a user-specified precision threshold. A complete description of the PrInCE workflow is included in the Supplementary Information.

A comparison of networks inferred from four CF-MS datasets (Scott *et al.*, 2017) confirmed that the R implementation of PrInCE

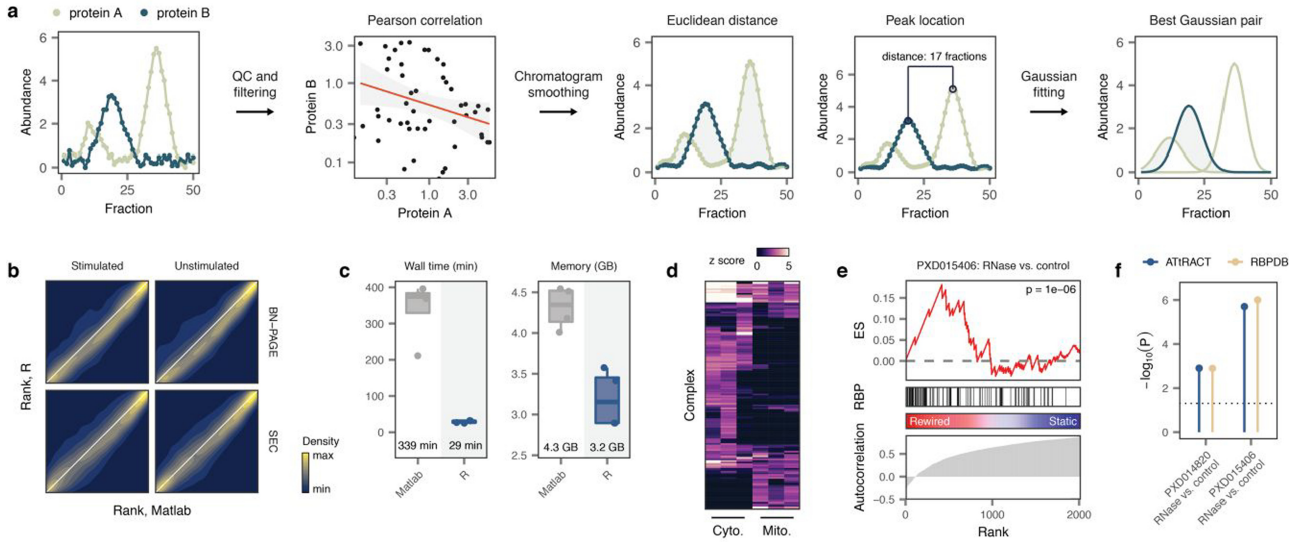

Fig. 1. Functionality of the PrInCE R package. (**a**) Schematic overview of the features calculated by PrInCE for each possible pair of interacting proteins. (**b**) Comparison of ranks assigned by the R and Matlab implementations of PrInCE to 10 157 866 candidate protein–protein interactions in cytoplasmic and mitochondrial membrane CF-MS data from Jurkat T cells before and after Fas-mediated apoptosis (Scott *et al.*, 2017). The CORUM database of protein complexes (Giurgiu *et al.*, 2019) was used to train both sets of classifiers. (**c**) Time and memory requirements of PrInCE analysis in R and Matlab for the four networks shown in panel (b). (**d**) Protein complexes from the CORUM database detected in cytoplasmic versus mitochondrial CF-MS data using the 'detect_complexes' function in PrInCE. Z-scores quantify the strength of protein complex co-elution signatures compared to 100 randomly shuffled sets of complexes. Only complexes with a z-score ≥ 1.96 in at least one replicate are shown. (**e**) Gene set enrichment analysis (GSEA) of RNA-binding proteins (RBPs) from the RBPDB database (Cook *et al.*, 2011), applied to autocorrelation scores computed fromcomparative CF-MS data before and after RNAse treatment (Mallam *et al.*, 2019). (**f**) GSEA *P*-values for two CF-MS datasets and two RBP databases after RNAse treatment, as calculated using the 'fgsea' package

yielded similar results to the previous Matlab version (Fig. 1b) and substantially outperformed a random classifier (Supplementary Fig. S1). However, this new version of PrInCE was substantially more efficient, displaying a 91% increase in speed and a 26% decrease in peak memory use (Fig. 1c). This increased efficiency could be attributed primarily to an increase in the efficiency of the Gaussian fitting (Supplementary Fig. S2), and opens up the possibility of larger-scale analyses using only a laptop computer. For example, we used PrInCE to re-analyze a human CF-MS dataset, with a total of 11 replicates and 1198 fractions, from a large-scale study (Wan *et al.*, 2015) in only 4.5 h, using 11.1 GB of RAM. The same analysis could not be completed in Matlab with 32 GB of RAM.

Beyond improvements in computational efficiency, the PrInCE R package also includes new functionality for CF-MS data analysis. First, we adapted a test previously described for thermal proximity co-aggregation data (Tan *et al.*, 2018) to identify protein complexes with statistically significant co-elution signatures, implemented in the 'detect_complexes' function. For protein complexes with at least three subunits, the median correlation between all subunits is computed, and compared to 100 shuffled complexes of equal size. Applying this test to cytoplasmic and mitochondrial CF-MS data (Scott *et al.*, 2017) clearly distinguished these two compartments on the basis of their protein complexes (Fig. 1d).

PrInCE also implements an autocorrelation-based method (Kerr *et al.*, 2020) to identify proteins whose interactions are 'rewired' in response to stimulation, in comparative CF-MS datasets. Briefly, for a given protein, the Pearson correlation to all other proteins in the dataset is calculated in each condition separately, yielding two vectors of correlation coefficients. These two vectors are compared to one another to produce the autocorrelation. Low autocorrelation values are indicative of proteins whose interaction profiles are rewired between conditions, whereas high autocorrelation values reflect consistent elution profiles. To demonstrate this test, implemented in the 'calculate_autocorrelation' function, we applied it to CF-MS data collected before and after RNAse treatment (Mallam *et al.*, 2019), and confirmed that known RNA-binding proteins were significantly enriched among proteins with a low autocorrelation (Fig. 1e and f).

Last, PrInCE implements several new classifiers in addition to the previously described naive Bayes model, including random forests, support vector machines and logistic regression, as well as the option to aggregate results from an ensemble of different classifiers. While the optimal choice of classifier may vary from one dataset to another, the ensemble option has the advantage that false positive interactions specific to a particular classifier will tend to be down-weighted in the aggregate rankings.

3 Conclusions

Through a ground-up rewrite of its Matlab predecessor, we have developed a fully open-source implementation of PrInCE that interfaces seamlessly with existing proteomics workflows in the Bioconductor project. Extensive documentation and tutorial vignettes are included in PrInCE to guide users through its major functionalities. Importantly, unlike other tools for CF-MS data analysis (Hu *et al.*, 2019), PrInCE does not consider any external data sources (e.g. gene coexpression or coevolution) in network inference, increasing its ability to discover novel interactions (Skinnider *et al.*, 2018). However, the performance of PrInCE depends on both the amount and quality of CF-MS training data, and the number of known protein complexes used to train the classifier. We hope that PrInCE will provide a useful resource for the systems biology and computational proteomics communities.

Acknowledgements

M.A.S. acknowledges support from a CIHR Vanier Canada Graduate Scholarship, an Izaak Walton Killam Memorial Pre-Doctoral Fellowship, a UBC Four Year Fellowship and a Vancouver Coastal Health–CIHR–UBC MD/PhD Studentship.

Funding

This work was supported, in part, by funding from Genome Canada/Genome BC (264PRO), and enabled in part by the support provided by WestGrid and Compute Canada (to L.J.F.), and through computational resources and

services provided by Advanced Research Computing at the University of British Columbia (to L.J.F.).

Conflict of Interest

none declared.

References

Cook,K.B. *et al.* (2011) RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.*, **39**, D301–D308.

Giurgiu,M. *et al.* (2019) CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Res.*, **47**, D559–D563.

Hu,L.Z. *et al.* (2019) EPIC: software toolkit for elution profile-based inference of protein complexes. *Nat. Methods*, **16**, 737–742.

Kerr,C.H. *et al.* (2020) Dynamic rewiring of the human interactome by interferon signaling. *Genome Biol.*, **21**, 140.

Kristensen,A.R. *et al.* (2012) A high-throughput approach for measuring temporal changes in the interactome. *Nat. Methods*, **9**, 907–909.

Mallam,A.L. *et al.* (2019) Systematic discovery of endogenous human ribonucleoprotein complexes. *Cell. Rep.*, **29**, 1351–1368.e5.

Scott,N.E. *et al.* (2015) Development of a computational framework for the analysis of protein correlation profiling and spatial proteomics experiments. *J. Proteomics*, **118**, 112–129.

Scott,N.E. *et al.* (2017) Interactome disassembly during apoptosis occurs independent of caspase cleavage. *Mol. Syst. Biol.*, **13**, 906.

Skinnider,M.A. *et al.* (2018) Genomic data integration systematically biases interactome mapping. *PLoS Comput. Biol.*, **14**, e1006474.

Stacey,R.G. *et al.* (2017) A rapid and accurate approach for prediction of interactomes from co-elution data (PrInCE). *BMC Bioinformatics*, **18**, 457.

Tan,C.S.H. *et al.* (2018) Thermal proximity coaggregation for system-wide profiling of protein complex dynamics in cells. *Science*, **359**, 1170–1177.

Wan,C. *et al.* (2015) Panorama of ancient metazoan macromolecular complexes. *Nature*, **525**, 339–344.