Published in final edited form as:

Proceedings (IEEE Int Conf Bioinformatics Biomed). 2018 December ; 2018: 2709–2712. doi:10.1109/
BIBM.2018.8621426.

INDEED: R package for network based differential expression analysis

Zhenzhi Li, Yiming Zuo, Chaohui Xu, Rency S. Varghese, and Habtom W. Ressom*

Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University,
Washington DC, USA

Abstract

With recent advancement of omics technologies, fueled by decreased cost and increased number of available datasets, computational methods for differential expression analysis are sought to identify disease-associated biomolecules. Conventional differential expression analysis methods (e.g. student's t-test, ANOVA) focus on assessing mean and variance of biomolecules in each biological group. On the other hand, network-based approaches take into account the interactions between biomolecules in choosing differentially expressed ones. These interactions are typically evaluated by correlation methods that tend to generate over-complicated networks due to many seemingly indirect associations. In this paper, we introduce a new R/Bioconductor package INDEED that allows users to construct a sparse network based on partial correlation, and to identify biomolecules that have significant changes both at individual expression and pairwise interaction levels. We applied INDEED for analysis of two omic datasets acquired in a cancer biomarker discovery study to help rank disease-associated biomolecules. We believe biomolecules selected by INDEED lead to improved sensitivity and specificity in detecting disease status compared to those selected by conventional statistical methods. Also, INDEED's framework is amenable to further expansion to integrate networks from multi-omic studies, thereby allowing selection of reliable disease-associated biomolecules or disease biomarkers.

Keywords

network-based differential expression analysis; metabolomics; partial correlation; proteomics; biomarker discovery

I. INTRODUCTION

Differential expression analysis is typically performed by statistical methods (e.g., *t*-test, SAM, etc.) to identify biomolecules with significant changes between biologically distinct groups. However, independent studies for the same clinical types of patients often lead to different sets of significant biomolecules with only few in common [1]. This might be attributed to the fact that differential expression analysis ignores the interactions among the biomolecules of interest. Network-based approaches are increasingly applied in omic studies to gain insights into these interactions. It has been demonstrated in cancer studies that

*Corresponding author: hwr@georgetown.edu.

combining individual biomarkers through sub-networks or network motifs can improve the diagnostic accuracy and biomarker robustness [2]-[4]. Recently, there has been a growing interest in differential network analysis, where the connection in the network represents a statistically significant change in the pairwise interaction between two biomolecules in biologically distinct groups. From the rewiring interactions shown in differential networks, biomolecules that have strongly altered connectivity between diverse groups can be identified.

We previously demonstrated the use of partial correlation to construct sparse biological networks based on omic data [5], [6]. Also, we introduced an INtegrated DifferEntial Expression and Differential network analysis (INDEED) that takes advantages of both conventional differential expression analysis (statistical methods) and a network-based approach to find disease-associated biomolecules. In a differential network constructed by INDEED, nodes denote biomolecules and the edges represent statistically significant changes in pairwise associations between pairs of biomolecules. Correlation is used to assess the changes in pairwise associations. However, conventional correlation methods tend to generate overcomplicated networks by inferring changes to many seemingly indirect associations; thus making it unsuitable for datasets with large dimensions [5], [6].

In this paper, we introduce a new R/Bioconductor package that uses partial correlation in the INDEED framework to distinguish direct and indirect associations and generate a sparse network of biomolecules based on omic data [7]. INDEED calculates an activity score for each biomolecule by combining the node degree of the biomolecule in the differential network with the statistical significance level of the biomolecule in distinguishing two disease groups. The score is used to rank the biomolecules and identify disease-associated biomolecules. The package allows users to choose the correlation method and associated parameters suitable to their omic datasets. Both partial or conventional correlation methods are implemented in the package. We tested INDEED using metabolomic and proteomic datasets from a cancer biomarker discovery study [8], [9].

The paper is organized as follows: Section II describes INDEED's workflow. Section III presents two omic datasets that we used to evaluate the performance of INDEED in identifying disease-associated biomolecules. Section IV discusses the results obtained using INDEED. Section V concludes the paper and discusses future plans.

II. METHOD

The main goal of INDEED is to integrate results from statistical significance analysis with those from differential network analysis to find disease-associated biomolecules. Given an omic dataset, differential expression analysis is first performed using statistical methods to obtain p-values, which assess the significance of changes in the levels of individual biomolecules between distinct biological groups. Then, a differential network is built based on partial correlation, which can distinguish between direct and indirect associations when evaluating the change of pairwise association on a biomolecular pair between distinct biological groups. Activity scores are computed based on p-values and the topology of the differential network. The activity scores are used to prioritize or rank biomolecules in

accordance with their ability to detect disease status. The top ranked biomolecules are expected to be more reproducible and reliable biomarkers than those identified by typical statistical analysis such as Student's *t*-test.

A. INDEED Workflow

An R/Bioconductor package INDEED can be found in GitHub (<https://github.com/ressomlab/INDEED.html>) or via Bioconductor (<https://www.bioconductor.org/packages/devel/bioc/html/INDEED.html>). In addition to providing this package with the omic dataset, class label, and biomolecule IDs (e.g., KEDD, UniProt, etc.), the user chooses the desired correlation method (partial, Pearson, or Spearman correlation methods) and corresponding parameters prior to constructing a differential network. P-values for each biomolecule can be provided by the user, but it is optional.

Activity scores are computed based on p-values and the topology of the differential network. Figures 1a and 1b summarize the workflows of the INDEED package for partial and conventional correlation methods, respectively. In the following, we describe partial and conventional correlation methods implemented in INDEED.

B. Partial Correlation

If partial correlation method is selected, INDEED applies a z-score transformation to the omic data prior to network construction. Then a parameter ρ is determined based on cross validation using `select_rho_partial` function as described in ref. [7]. This parameter controls the sparsity of the network. An optional parameter of this function is `error_curve`; user can choose whether to display the error curve plot in the window to assist visualization and ρ value selection. The output of `select_rho_partial` is a list of preprocessed data from users input data, and ρ values for each biological or disease group (e.g., case or control).

In the second step of partial correlation methods, user needs to input the preprocessed data list from `select_rho_partial` function. User will also need to select ρ values for case and control groups, option for ρ value selection can be “min” as ρ value from minimum rule; “ste” as ρ value from one standard error rule; user can also enter a ρ value of their interest, based on error curve from `select_rho_partial` function or outside knowledge. In building differential networks, number of permutation and the threshold are pre-defined to be 1000 and 0.025, respectively, and subject to changes based on user preference. P-values of all biomolecules can be supplied by the user in the function argument, but is not required. If p-values are not provided by the user, INDEED's `partial_cor` will calculate a p-value for each biomolecule. The output of `partial_cor` consists of two dataframes. The first dataframe “activity_score” contains list of biomolecules with p-values, node degrees, and activity scores. Detailed method on activity score calculation can be found in ref. [7]. The second data frame “diff_network” contains differential network information consisting of node to node connections with weights and binary values of each connection.

C. Conventional Correlation

INDEED allows users the option to construct a differential network based on the Pearson or Spearman correlation methods using the “non_partial_cor” function. In addition to proving

the omic dataset, class labels, and biomolecule IDs, the user may select either “pearson” or “spearman” for correlation analysis. Optional parameters include the desired number of permutations and the permutation threshold; defaults for these parameters are 1000 and 0.025, respectively. The output of `non_partial_cor` is similar to `partial_cor`, a list of two dataframes, “activity_score” and “diff_network”.

III. DATASETS

A. Metabolomics

To evaluate the performance of INDEED, we considered metabolomic data acquired by measuring the levels of 39 metabolites in sera from 60 hepatocellular carcinoma (HCC) cases and 60 patients with liver cirrhosis (CIRR). The subjects were recruited at MedStar Georgetown University (GU) Hospital, Washington, DC through a protocol approved by Georgetown University Institutional Review Board (URB). Table I shows the characteristics of the subjects in this cohort (referred to as GU cohort).

B. Proteomics

Another dataset from a proteomic study that consists of 100 proteins was used to evaluate INDEED. The data were acquired by analysis of plasma samples from 89 participants (40 HCC cases and 49 patients with liver cirrhosis) recruited at Tanta University (TU) Hospital, Tanta, Egypt through a protocol approved by Tanta University Ethical Committee. Table II presents the characteristics of this TU cohort.

IV. RESULTS AND DISCUSSION

A. Metabolomics

Differential network analysis of the metabolomic data was performed via INDEED. Tables III and IV show the top 10 metabolites obtained by INDEED using partial and the Pearson correlation method, respectively. The metabolites are ranked by activity score, which combines p-values and node degrees. Among the top 10 candidates ranked by partial correlation, 50% were also found by Pearson correlation. Overall, those found by partial correlation tend to have higher activity scores. Figure 2 shows the differential network obtained by INDEED using partial correlation. Each node represents a metabolite and each edge represents a connection between two metabolites. Node size is proportional to the node degree of the metabolite.

B. Proteomics

We performed differential network analysis of the proteomic data. Tables V presents the top 10 proteins obtained by INDEED using partial correlation and ranked by activity score. Figure 3 depicts the corresponding differential network.

V. CONCLUSION

In this paper, we introduce a new R/Bioconductor package INDEED that can be used to construct to help build differential networks by evaluating the partial correlation between

biomolecules based on omic data. The package computes activity scores that can be used to prioritize or rank biomolecules in accordance with their level of association with disease status. The top ranked biomolecules are expected to be more reproducible and reliable biomarkers than those identified by typical statistical analysis such as Student's t-test. We used metabolomic and proteomic datasets for a cancer biomarker discovery study to test INDEED. Future work will focus on performing a thorough evaluation of INDEED using data that consist of ground-truth information to determine the biological significance of the markers discovered by INDEED. Also, we will investigate the potential use of INDEED for differential expression analysis by integration of networks derived from multi-omic data.

ACKNOWLEDGMENT

This work is in part supported by NIH grants U01CA185188 and R01GM123766 awarded to HWR.

REFERENCES

- [1]. Ein-Dor L, Zuk O and Domany E, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, (15), pp. 5923–5928, 2006. [PubMed: 16585533]
- [2]. Chuang HY et al., "Network-based classification of breast cancer metastasis," *Mol Syst Biol*, vol. 3, pp. 140, 2007. [PubMed: 17940530]
- [3]. Zhang Y et al., "Network motif-based identification of transcription factor-target gene relationships by integrating multi-source biological data," *BMC Bioinformatics*, vol. 9, pp. 203–2105–9–203, 2008. [PubMed: 18426580]
- [4]. Kim Youngsoon et al., "DiffGRN: differential gene regulatory network analysis," *International Journal of Data Mining and Bioinformatics (IJDMB)*, vol. 20 (4), pp. 362–379, 2018. [PubMed: 31114627]
- [5]. Zuo Y et al., "Biological network inference using low order partial correlation," *Methods*, vol. 69, (3), pp. 266–273, 2014. [PubMed: 25003577]
- [6]. Zuo Y et al., "Incorporating prior biological knowledge for network-based differential gene expression analysis using differentially weighted graphical LASSO," *BMC Bioinformatics*, vol. 18, (1), pp. 99–017–1515–1, 2017. [PubMed: 28187708]
- [7]. Zuo Y et al., "INDEED: Integrated differential expression and differential network analysis of omic data for biomarker discovery," *Methods*, vol. 111, pp. 12–20, 2016. [PubMed: 27592383]
- [8]. Nezami Ranjbar MR et al., "GC-MS based plasma metabolomics for identification of candidate biomarkers for hepatocellular carcinoma in Egyptian cohort," *PloS One*, vol. 10, (6), pp. e0127299, 2015. [PubMed: 26030804]
- [9]. Di Poto C et al., "Metabolomic characterization of hepatocellular carcinoma in patients with liver cirrhosis for biomarker discovery," *Cancer Epidemiol. Biomarkers Prev*, 2017; 26(5):675–683. [PubMed: 27913395]

Fig. 1. INDEED workflows using partial (A) and non-partial (B) correlation methods.

Fig. 2.
Differential network of metabolites in case study. Node size is proportional to the node degree.

Fig. 3. Differential network of proteins. Node size is proportional to the node degree.

TABLE I.

CHARACTERISTICS OF THE GU COHORT.

		HCC (N=60)	CIRR (N=60)	p-value
Age	Mean (SD)	59.4 (6.7)	58.9 (7.1)	0.7303
Gender	Male	73.3%	73.3%	1
	Caucasian	56.7%	65.0%	
Ethnicity	African American	28.3%	25.0%	0.7228
	Others	15.0%	10.0%	
	Current	23.3%	21.7%	
Smoker	Former	48.3%	48.3%	1
	None	28.3%	26.7%	
	Current	13.3%	15.0%	
Alcohol	Former	51.7%	51.7%	0.9353
	None	33.3%	30.0%	
HCV Serology	HCV Ab+	63.3%	38.3%	0.0189
	HCV RNA+	55.0%	38.3%	0.7916
HBV Serology	Anti HBC+	41.7%	28.3%	0.2498
	HBsAg+	11.7%	5.0%	0.3212
MELD	Mean (SD)	11.1 (3.7)	16.4 (14.3)	7.10E-03
	MELD < 10	48.3%	16.7%	6.91E-05
AFP	Median (IQR)	28.8 (83.2)	4.4 (9.7)	3.22E-05
HCC Stage	Stage I	60.0%		
	Stage II	38.3%		
	Stage III	5.0%		
	Unknown	13.3%		

TABLE II.**CHARACTERISTICS OF THE TU COHORT.**

		HCC (N= 40)	CIRR (N= 49)	p-value
Age	Mean (SD)	53.2 (3.9)	53.8 (7.6)	0.353
Gender	Male	77.5%	67.3%	0.3474
HCV Serology	HCV Ab+	100.0%	100.0%	1
HBV Serology	HBsAg+	0.0%	6.1%	0.2492
MELD	Mean (SD)	18.6 (7.7)	18.9 (7.1)	0.1328
	MELD < 10	20.0%	12.2%	0.3863
AFP	Median	275.9		
HCC Stage	Stage I	72.5%		
	Stage II	15.0%		
	Stage III	5.0%		
	Unknown	7.5%		

TABLE III**TOP 10 METABOLITES FROM PARTIAL CORRELATION METHOD**

	KEGG ID	P-value	Node Degree	Activity Score
1	C03046	0.077	8	9.3
2	C00065	0.162	7	8.4
2	C02497	0.725	5	8.1
4	C00097	0.552	6	7.7
5	C00188	0.423	4	6.7
6	C05145	0.079	3	6.2
7	C00022	0.385	5	6.1
8	C00064	0.033	3	5.7
9	C00183	0.058	3	5.4
10	C00124	0.58	3	4.8

TABLE IV

TOP 10 METABOLITES FROM PEARSON CORRELATION METHOD

	KEED ID	P-value	Node Degree	Activity Score
1	C00581	0.55	6	8.8
2	C00186	0.318	5	6.6
2	C03046	0.077	4	6.6
4	C00025	0.511	4	5.8
5	C00183	0.058	4	5.7
6	C05145	0.079	4	5.7
7	C00064	0.033	3	4.6
8	C00124	0.58	5	4.6
9	C00187	0.123	1	4.6
10	C00383	0.003	1	4.6

TABLE V**TOP 10 PROTEINS FROM PARTIAL CORRELATION METHOD**

	UniProt ID	P-value	Node Degree	Activity Score
1	P29622	0.001	9	25
2	P01042	0.001	8	21.8
3	P04004	0.002	8	20.3
4	Q6EMK4	0.01	6	18.2
5	P03952	0.132	6	17.6
6	P00747	0	5	17.3
7	Q13201	0.003	5	16.8
8	P51884	0.009	5	16.6
9	P24592	0.228	6	16.5
10	P07333	0.004	5	16.4