BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 5 2014, pages 737–739
doi:10.1093/bioinformatics/btt576

Systems biology

Advance Access publication October 3, 2013

iPEAP: integrating multiple omics and genetic data for pathway enrichment analysis

Haoqi Sun[1,†], Haiping Wang[2,†], Ruixin Zhu[1], Kailin Tang[1], Qin Gong[1], Juan Cui[3], Zhiwei Cao[1,*] and Qi Liu[1,*]

[1]Department of Bioinformatics, School of Life Science and Technology, Tongji University, Siping Rd. No. 1239, Shanghai 200092, [2]Department of Computer Science, Hefei University of Technology, Tunxi Rd. No. 193, Hefei 230009, China and [3]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30602-7229, USA

Associate Editor: Martin Bishop

ABSTRACT

Summary: A challenge in biodata analysis is to understand the underlying phenomena among many interactions in signaling pathways. Such study is formulated as the pathway enrichment analysis, which identifies relevant pathways functional enriched in high-throughput data. The question faced here is how to analyze different data types in a unified and integrative way by characterizing pathways that these data simultaneously reveal. To this end, we developed integrative Pathway Enrichment Analysis Platform, *iPEAP*, which handles transcriptomics, proteomics, metabolomics and GWAS data under a unified aggregation schema. *iPEAP* emphasizes on the ability to aggregate various pathway enrichment results generated in different high-throughput experiments, as well as the quantitative measurements of different ranking results, thus providing the first benchmark platform for integration, comparison and evaluation of multiple types of data and enrichment methods.

Availability and implementation: *iPEAP* is freely available at http://www.tongji.edu.cn/~qiliu/ipeap.html.

Contact: qiliu@tongji.edu.cn or zwcao@tongji.edu.cn

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

Received on June 21, 2013; revised on August 27, 2013; accepted on September 29, 2013

1 INTRODUCTION

As the advent of post-genomic era, omics science and high-throughput technology have generated massive data. Quantitative methods are developed to better relate the underline pattern of the data to biologic concepts, like Gene Ontology terms, biological functions or pathways and so forth. Among them, biochemical pathways are the primary focus, and they are extensively used to interpret biological data in a network viewpoint (Kamburov *et al.*, 2011). In particular, pathway overrepresentation and enrichment analyses have become important approaches for the interpretation of data from various high-throughput experiments, which aim at projecting a set of genes, proteins or metabolites onto predefined groups, calculating significance of correlation for each group and then ranking them according to the statistic. Correlation statistics can be defined in many ways, and various pathway enrichment analysis algorithms were proposed, like overrepresentation analysis (Draghici *et al.*, 2003), gene set enrichment analysis (Subramanian *et al.*, 2005), network topology-based approaches SPIA (Draghici *et al.*, 2007), DEAP (Haynes *et al.*, 2013) and so forth.

With rapid accumulation of massive and various types of omic and genetic data, the major challenge for pathway analysis now becomes how to analyze different types of data for a given experiment in a unified way and interactively mining the underline functions under these data simultaneously. For a given set of samples, multiple types of high-throughput data, such as the transcriptomics, proteomics, metabolomics and genome-wide association study (GWAS) data can be generated at the same time. These data describe the samples from different perspectives, and each of them can be complementary to each other to obtain the unbiased enrichment pathway list. There are a few works existed on the integrative pathway analysis like integrative GWAS and gene expression analysis in prostate cancer (Jia *et al.*, 2012), IMPaLA (Kamburov *et al.*, 2011), MAPE (Shen and Tseng, 2010) and so forth. But traditionally only one pathway enrichment algorithm or one specific integration method was provided. To the best of our knowledge, integrative benchmark platforms allowing multiple enrichment and integration algorithms to handle multiple omic and genetic data, together with easy-to-use user interface, are absent.

To tackle the multiple data integration problem, we presented a software *iPEAP*, namely integrative Pathway Enrichment Analysis Platform, to perform integrative pathway enrichment analysis. It is a Java based, user-friendly, graphical tool with the aim of integrating transcriptomics, proteomics, metabolomics and GWAS data for pathway level analysis. Furthermore, various state-of-art pathway enrichment analysis algorithms for the single data type as well as the quantitative evaluation measurements and tools for pathway ranking were incorporated into *iPEAP*, which are useful for the access, evaluation and comparison of different approaches in one platform.

2 METHODS

iPEAP guides users to follow the following pipeline, i.e. data type selection, data input, enrichment analysis, ranking aggregation and evaluation to perform the integrative pathway enrichment analysis (Fig. 1). The calculated pathway ranking is displayed as a table including pathway

*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.

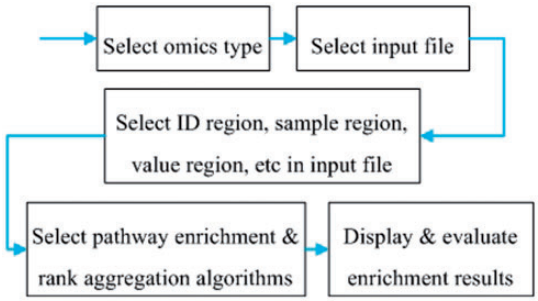

Fig. 1. Main workflow of *iPEAP*

names, categories, enrichment scores and so forth. Official KEGG pathway graphs are accessible by double clicking the pathway ID or title. The hit genes/metabolites in pathway are shown in an adjacent window by clicking the corresponding pathway record, together with their expression values, which can be the reads per kilobase per million (RPKM)/fragments per kilobase of exon per million fragments mapped (FPKM) values or the calculated differential expression values. Linkouts to other network sources like STRING (Franceschini *et al*., 2013), MetPA (Xia and Wishart, 2010), IMPaLA (Kamburov *et al*., 2011) and KOBAS (Wu *et al*., 2006) are also provided. Users can also compare and evaluate the pathway rankings using two tools *ListComparer* and *RankEvaluator*, respectively, which were developed in *iPEAP* (See Supplementary Material). For the first time three quantitative measurements, i.e. *NDCG* (normalized discounted cumulative gain) (Jarvelin *et al*., 2002) *ERR* (expected reciprocal rank) (Chapelle *et al*., 2009) and *P* (proportion) (Tsai *et al*., 2007) were included in *iPEAP* to evaluate ranking results (See Supplementary Methods). In addition, various state-of-art pathway enrichment algorithms were also incorporated for handling transcriptomics, proteomics, metabolomics and GWAS data, respectively (Supplementary Table S1). Two integration schemas, i.e. *RobustRankAggreg* (Kolder *et al*., 2012) and *RankAggreg* (Pihur *et al*., 2007) can be selected to integrate different pathway rankings into one unbiased overall ranking. Some simple aggregation methods such as min, median and mean are also provided (Willett, 2013) (See Supplementary Methods).

3 RESULTS AND DISCUSSION

3.1 Application to multiple types of data

To illustrate the use of *iPEAP* as an integrative tool, we analyzed two cases with multiple types of data integration from two established studies. One case is focused on non-genotoxic carcinogenesis where the human hepatocarcinoma cell line HepG2 was exposed to the environmental carcinogen 2, 3, 7, 8-tetrachlorodibenzo-*p*-dioxin. Both the transcriptomics and metabolomics profiles were integrated with *iPEAP*, helping to understand the molecular mechanisms induced by toxic compounds *in vitro* in human cells (Jennen *et al*., 2011). Another one is a non-targeted metabolomics study associated with the single nucleotide polymorphism loci on two cohorts, namely the German KORA F4 Study ($n = 1768$) and the British Twins UK study ($n = 1052$) (Suhre *et al*., 2011). In this case, the metabolomics and GWAS profile were integrated to study the functional pathways using *iPEAP*. The integrated pathway enrichment results for two cases are listed in Supplementary Table S4 and S7, respectively. Both of them obtained more reasonable and novel biological results compared with that only one single type of data was used. Detailed analysis can be referred in Supplementary Materials.

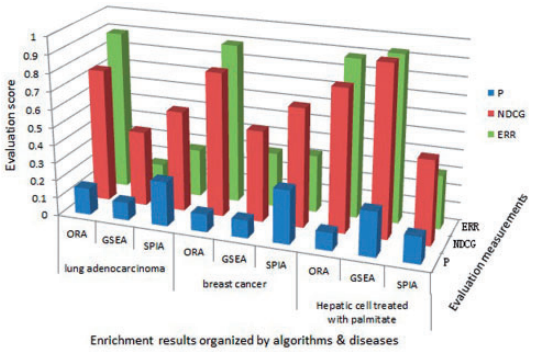

Fig. 2. The ranking performance of various pathway enrichment algorithms on three datasets

3.2 *iPEAP* for pathway ranking evaluation and comparison

Using the built-in tools and quantitatively ranking evaluation measurements in *iPEAP*, ranking comparison and evaluation can be easily carried out to further interpret the pathway ranking results. In our study, three transcriptomics datasets, namely lung adenocarcinoma, breast cancer and hepatic cell treated with palmitate, initially used by (Draghici *et al*., 2007), were applied as the benchmark data to evaluate three classical pathway enrichment algorithms, namely overrepresentation analysis, gene set enrichment analysis and SPIA, using the built-in tool RankEvaluator. We evaluated the top 20 pathways of the nine ranking results, using three measurements as NDCG, ERR and P, respectively, provided the first time a quantitative evaluation of distinct pathway enrichment algorithms, as shown in Figure 2.

In summary, *iPEAP* provides a powerful platform to analyze different types of high-throughput data and investigate their pathway-level mechanism in an integrative fashion.

Funding: National Natural Science Foundation of China (31100956 and 61173117) and National 863 program (2012AA020405 and 2012AA011005).

Conflict of Interest: none declared.

REFERENCES

Chapelle,O. *et al*. (2009) Expected reciprocal rank for graded relevance. In: *CIKM '09 Proceedings of the 18th ACM Conference on Information and Knowledge Management*. pp. 2–6, 621–630.

Draghici,S. *et al*. (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.

Draghici,S. *et al*. (2007) A systems biology approach for pathway level analysis. *Genome Res*., **17**, 1537–1545.

Franceschini,A. *et al*. (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*., **41**, D808–D815.

Haynes,WA. *et al*. (2013) Differential expression analysis for pathways. *PLoS Comput. Biol*., **9**, e1002967.

Jarvelin,K. and Kekalainen,J. (2002) Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst*., **20**, 422–446.

Jennen,D. *et al*. (2011) Integrating transcriptomics and metabonomics to unravel modes-of-action of 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) in HepG2 cells. *BMC Bioinformatics*, **5**, 139–151.

Jia,P. *et al*. (2012) Integrative pathway analysis of genome-wide association studies and gene expression data in prostate cancer. *BMC Syst. Biol*., **6**, S13.

Kamburov,A. *et al.* (2011) Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics*, **27**, 2917–2918.
Kolder,R. *et al.* (2012) Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, **28**, 573–580.
Pihur,V. *et al.* (2007) Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach. *Bioinformatics*, **23**, 1607–1615.
Shen,K. and Tseng,GC. (2010) Meta-analysis for pathway enrichment analysis when combining multiple genomics studies. *Bioinformatics*, **26**, 1316–1323.
Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
Suhre,K. *et al.* (2011) Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, **477**, 54–62.
Tsai,MF. *et al.* (2007) FRank: a ranking method with fidelity loss. In: *SIGIR '07 Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 23–27, 383–390.
Willett,P. (2013) Combination of similarity rankings using data fusion. *J. Chem. Inf. Model.*, **53**, 1–10.
Wu,J. *et al.* (2006) KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res.*, **34** (**Suppl. 2**), W720–W724.
Xia,J. and Wishart,D. (2010) MetPA: a web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics*, **26**, 2342–2344.