BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 13 2011, pages 1878–1879
doi:10.1093/bioinformatics/btr278

Systems biology

Advance Access publication May 5, 2011

MPEA—metabolite pathway enrichment analysis

Matti Kankainen[1,*], Peddinti Gopalacharyulu[1], Liisa Holm[2] and Matej Orešič[1]

[1]VTT Technical Research Centre of Finland, Espoo and [2]Institute of Biotechnology and Department of Biological Sciences, University of Helsinki, Helsinki, Finland

Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: We present metabolite pathway enrichment analysis (MPEA) for the visualization and biological interpretation of metabolite data at the system level. Our tool follows the concept of gene set enrichment analysis (GSEA) and tests whether metabolites involved in some predefined pathway occur towards the top (or bottom) of a ranked query compound list. In particular, MPEA is designed to handle many-to-many relationships that may occur between the query compounds and metabolite annotations. For a demonstration, we analysed metabolite profiles of 14 twin pairs with differing body weights. MPEA found significant pathways from data that had no significant individual query compounds, its results were congruent with those discovered from transcriptomics data and it detected more pathways than the competing metabolic pathway method did.

Availability: The web server and source code of MPEA are available at http://ekhidna.biocenter.helsinki.fi/poxo/mpea/.

Contact: matti.kankainen@helsinki.fi

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

Received on December 23, 2010; revised on April 22, 2011; accepted on April 25, 2011

1 INTRODUCTION

Metabolite profiling refers to a set of techniques by which a wide range of metabolites can be detected and quantified from biological extracts (Fernie *et al*., 2004). Such experiments are often conducted to discover biomarkers and data resulting from these studies are analyzed using complex multivariate statistical methods (Cevallos-Cevallosa *et al.*, 2009; Fernie *et al*., 2004). Recently, tools for analysing metabolite data in the context of predefined biological metabolite sets have, however, been reported (Aggio *et al.*, 2010; Xia and Wishart, 2010). Instead of discovering specific metabolites that discriminate sample groups, the aim of these methods is to discover predefined metabolic pathways or biological networks that are co-ordinately altered in the experiment. Such changes are often assessed by employing the gene set enrichment analysis (GSEA) procedure (Subramanian *et al.*, 2005) to test whether the elements of some predefined biological group are enriched towards the top or bottom of a ranked list.

Here, we describe the metabolite pathway enrichment analysis (MPEA) for the visualization and functional analysis of metabolite data at the system level. MPEA tests whether metabolites in predefined pathways occur towards the top or bottom of a ranked list. It can be applied either to pre-annotated compounds or gas chromatography coupled with mass spectrometry (GC-MS) data consisting of mass spectral tags (MSTs). If MSTs are provided, they are characterized using GMD (Kopka *et al.*, 2005). This feature was included to facilitate the analysis of data generated by GC-MS—a popular analytical tool for metabolite profiling (Fernie *et al*., 2004). Furthermore, as it is unlikely that all query compounds are unambiguously resolved, MPEA has been developed to handle ambiguously identified compounds.

We tested MPEA on a study in which adipose tissue transcriptome and metabolome were studied in 14 twin pairs discordant for body weight (Pietiläinen *et al.*, 2008). MPEA revealed more significantly altered pathways in acquired obesity based on metabolite data than the competing method did and its results corresponded with the pathways that GSEA found from transcriptomics data.

2 METHODS

The idea of MPEA is to identify coordinately changed Kyoto Encyclopedia of Genes and Genomes (KEGG; Ogata *et al.*, 1999) and Small Molecule Pathway Database (SMPDB; Frolkis *et al.*, 2010) pathways using metabolite data. It requires a list of pre-annotated compounds or GC-MS-based MSTs and has two modes of enrichment analysis: single set and iterative. Single set mode is used for unranked lists whereas the iterative mode is for entries ranked, for example, by performing *t*-tests on estimated concentrations. User-given annotations should include KEGG-identifiers (KIDs). If required, an identifier set with multiple KIDs can be created for a compound. The KIDs within the identifier set represent interchangeable and alternative mappings for that compound. It is also possible to assign multiple identifier sets to a compound that has been identified ambiguously. Alternatively, MPEA supports GC-MS-based MSTs consisting of mass fragmentation pattern and retention index (RI). MSTs are identified and annotated using GMD. Pathway inconsistencies are resolved and the statistical enrichment of metabolic pathways is estimated using the hypergeometric distribution coupled with a permutation test.

Identification of MSTs is based on the spectral comparison against reference spectra and is achieved using the MS analysis tool at GMD (Kopka *et al.*, 2005). The relatedness of mass-to-charge ratios is measured with several metrics and analytes below a given similarity threshold are discarded. Spectral matching can be further restricted by setting a RI window for matches or by naming MSTs. If the name is detected among the matches, only analytes having that name are accepted. An identifier set is formed over metabolites associated with a particular analyte and each matching analyte generates a new identifier set for the query.

To obtain reliable results, ambiguous identifications need to be resolved and the number of KIDs assigned to the pathway requires to be established. We tackle this problem using a novel approach (Fig. 1). For each pathway, an entity relationship matrix is created to resolve the many-to-many relationships between the query compounds and KIDs assigned to the pathway. The number of KIDs found for the pathway is estimated

*To whom correspondence should be addressed.

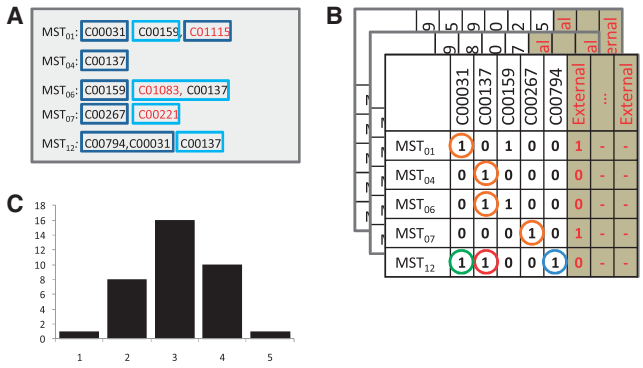

Fig. 1. An example of five MSTs associated with galactose metabolism. (**A**) Listing shows KIDs assigned to compounds and identifier sets (boxes). KIDs not part of the pathway are shown in red. (**B**) The relationship matrix is the Cartesian product of KID and compound groups for the pathway and includes columns for the external mapping variables. KIDs by which the compounds are describable are highlighted with ones. Selecting the KIDs marked with orange circles and the KID marked with green, blue or red circle would create a compound–KID combination with 3, 3 or 4 distinct KIDs, respectively. (**C**) The number of distinct KIDs is recorded giving the minimum (1), median (3) or maximum (5) numbers.

by enumerating compound-KID combinations from the matrix. For each combination, one of the alternative KIDs of the compound is chosen for the compound, the selected KIDs are joined and the number of distinct KIDs is recorded.

Alternatively, if a query compound has other identifier sets that do not link the compound to the pathway, external mapping variables are added for the compound in the entity relationship matrix. As the external mapping variable is not part of the pathway, it creates a set of combinations in which the count of distinct KIDs is not increased. This approach allows us to decrease the probability that a compound with distinct identifier sets is added to the current pathway. The upper bound for the number of external mapping variables can be set by the user.

The statistical significance of pathways is tested using the hypergeometric distribution. The minimum, median or maximum number of distinct KIDs represents the number of successfully drawn elements from the input list whereas the background distribution is calculated using the number of KIDs associated with pathways in total and KIDs part of the tested pathway. Alternatively, background distribution specific to given organism(s) can be constructed. User may also supply the list of metabolites for calculating the background distributions. In the iterative mode, the statistic is calculated repetitively increasing the size of the list by one. The list position yielding the most significant result is recorded. Corrected *P*-values are calculated by permuting the list and repeating the statistical procedure many times. In the single set mode, the whole list is used to compute the enrichment statistic.

The MPEA web interface and source code are available at: http://ekhidna.biocenter.helsinki.fi/poxo/mpea/. See Supplementary Material for illustrative analyses.

3 EXAMPLE ANALYSIS USING MPEA

For an evaluation, metabolomics and transcriptomics data obtained from adipose tissue biopsies in 14 twin pairs discordant for body weight was analysed. The transcriptomics data were previously reported (Pietiläinen *et al.*, 2008) and was here studied using GSEA. Metabolomic analysis in the same tissue samples was performed using 2D GC coupled to time-of-flight MS, as described previously (Mattila *et al.*, 2008). *T*-test was used to compare study groups and for ranking MSTs. MPEA and MSEA were used to find altered pathways (see Supplementary Material).

In our test, MPEA assigned query compounds to 4–15 pathways. Spearman's (ρ_S) correlations were calculated over the complete observation pairs to compare the lists of significant pathways derived from GSEA and MPEA (Supplementary Material). The result shows that the highest correlation values ($\rho_S = 0.94$, $\rho_S = 0.93$ and $\rho_S = 0.90$, $n = 6$ in each comparison) were obtained using median number of distinct KIDs and a maximum of three, one and two external mapping variables, respectively. Thus, the use of external mapping variables to resolve many-to-many relationships makes the results of MPEA more similar to that of GSEA. Of the listed parameter settings, the one with one external mapping variable gave two significant pathways ($P \leq 0.05$). Metabolites belonging to amino sugar and nucleotide sugar, and ascorbate and aldarate metabolism were all coordinately more abundant in lean twins. Neither of the pathways was however detected by GSEA.

In comparison with MSEA, MPEA detected all but two of the SMPDB pathways detected by MSEA. MPEA associated more metabolites to most pathways and produced more significant *P*-values. For example, MPEA assigned four compounds to fatty acid biosynthesis and galactose metabolism, whereas the MSEA was at best able to assign a single metabolite to any pathway, illustrating the advantage of GMD-based compound identification and the better sensitivity of an enrichment analysis done to ranked metabolite lists.

ACKNOWLEDGEMENTS

We thank QBIX group at VTT for helpful discussion.

Funding: European Union-funded project ETHERPATHS (FP7-KBBE-222639, http://www.etherpaths.org/).

Conflict of Interest: none declared.

REFERENCES

Aggio,R.B. *et al.* (2010) Pathway Activity Profiling (PAPi): from the metabolite profile to the metabolic pathway activity. *Bioinformatics*, **26**, 2969–2976.

Cevallos-Cevallosa,J.M. *et al.* (2009) Metabolomic analysis in food science: a review. *Trends Food Sci. Technol.*, **20**, 557–566.

Fernie,A.R. *et al.* (2004) Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.*, **5**, 763–769.

Frolkis,A. *et al.* (2010) SMPDB: the small molecule pathway database. *Nucleic Acids Res.*, **38**, D480–D487.

Kopka,J. *et al.* (2005) GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics*, **21**, 635–638.

Mattila,I. *et al.* (2008) Application of lipidomics and metabolomics to the study of adipose tissue. *Methods Mol. Biol.*, **456**, 123–130.

Ogata,H. *et al.* (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **27**, 29–34.

Pietiläinen,K.H. *et al.* (2008) Global transcript profiles of fat in monozygotic twins discordant for BMI: pathways behind acquired obesity. *PLoS Med.*, **11**, e51.

Subramanian,A. *et al.* (2010) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Xia,J. and Wishart,D.S. (2010) MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res.*, **38**, W71–W77.