BIOINFORMATICS

Vol. 00 no. 00 2012
Pages 1–3

Pathway-based visualization of cross-platform microarray datasets

Clemens Wrzodek [1]*, Johannes Eichner [1] and Andreas Zell [1]

[1]Center for Bioinformatics Tuebingen (ZBIT),
University of Tuebingen, 72076 Tübingen, Germany

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Results: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text

Contact: clemens.wrzodek@uni-tuebingen.de

1 INTRODUCTION

The first generation of microarray platforms was developed as a high-throughput technique for profiling the transcriptome of diverse biological systems (i.e., cells, organs or organisms) under various experimental conditions (TODO: Refs. Golub, etc.). As these traditional gene-centered arrays were mostly limited to mRNA transcripts, the vast majority of visualization tools are still focused on mRNA datasets (Refs.). To date, a plethora of different microarray platforms are readily available. These include gene-centered platforms which rely on current genome annotations as well as unbiased tiling arrays which interrogate large non-repetitive regions of the genome. Diverse types of platforms have been specificly designed for the interrogation of different genomic features, ranging from mRNA or miRNA transcripts, through proteins or protein modifications, to relevant functional elements such as exons, SNPs or promoters (TODO: Refs). In addition to arrays serving for the quantification of global gene expression on the RNA or protein level, also epigenetic modifications such as DNA methylation (DNAm) can be monitored on a genome-wide level using microarray technology (Hoheisel, 2006).

Several tools exist for the visual inspection of datasets from individual platforms (Refs.). However, the current inventory of publicly available tools, which are capable of integrating and jointly visualizing data from multiple microarray platforms, is still very limited. Here, we introduce a method, for integrated pathway-centered visualization of datasets generated from the same biological samples using different microarray platforms, which interrogate complementary genomic and epigenomic features.

In contrast to commonly used region-based visualization methods (e.g., (see Kent *et al.*, 2002), TODO: weitere Beispiele und Zitate), we propose to visualize the microarray data in the context of specific signaling or metabolic pathways, which can in many cases be more easily related to the biological problem under study than individual genes or genomic regions.

There are other tools, specialized in pathway analysis (e.g., Ingenuity,), or in pathway visualization (Cytoscape, KEGG Atlas). Some even offer visualizing data in a pathway (GenMAPP, KEGG Array, Symons programm). No method today for high-dimensional, heterogeneous cross-platform datasets.

which are relevant for the conducted experiment can be deduced from the differentially expressed genes using pathway enrichment analysis. For this purpose, candidate pathways which are putatively involved in the studied biological phenomenon are typically ranked according to p-values resulting from a hypergeometric test for overrepresentation. The results are usually presented to the user as a sorted table or barplot which does not show any superordinate relations of the pathways detected as enriched with differentially expressed genes. In addition to this traditional approach, we also implemented an alternative method, which provides the user with a more structured view of the metabolic pathways linked to a certain microarray experiment. Owing to the hierarchical structure of the KEGG PATHWAY database, InCroMAP can visualize the enrichments computed for each individual metabolic pathway in the context of the higher-order overall metabolic pathway map compiled by KEGG (http://www.genome.jp/kegg/pathway/map/map01100.html). Starting from either a ranked pathway table or a colored meta-pathway map, individual pathways can be visualized in InCroMAP and overlaid with microarray data from multiple platforms, to facilitate thorough visual inspection of the measured pathway alterations. In contrast to previous work (TODO: Refs zu Cytoscape, Explain, etc.), InCroMAP offers convenient functions to overlay a pathway plot with sample-matched microarray data from platforms measuring gene regulation on mRNA, miRNA, DNAm and protein level.

2 METHODS

Before use with InCroMAP the microarray datasets of interest have to be preprocessed and annotated using platform-specific workflows (Refs zu Ringo, MEDME, AgiMicroRna, Limma, Annotate). These workflows usually involve (1) the quality control of the raw data

*to whom correspondence should be addressed

(TODO: cite arrayQualityMetrics, simpleaffy, etc.), (2) the data normalization to correct for background noise and experimental variation (TODO: cite RMA and GCRMA paper or review paper about normalization), and (3) the mapping of the probes to genes or genomic regions. After these preprocessing steps the microarray data has to be exported in tabular format (e.g., CSV or Excel). These tables have to contain two types of columns: (1) annotation columns, containing probe or probeset IDs (e.g., Affymetrix IDs) and database IDs of the corresponding genes (e.g., Ensembl or Entrez IDs), and (2) data columns containing either fold-changes and/or p-values resulting from basic statistical analysis of the microarray data.

The pathway data is automatically imported from KEGG. In the KEGG PATHWAY database each pathway map is available for download as a KGML document which is internally converted into a GraphML document by InCroMAP. These GraphML files can be interactively visualized by InCroMAP using the yWorks graph viewer (TODO: Ref.). In order to overcome limitations of the KGML format, one can create an overlay graph that shows the original KEGG pathway image in the background, which may provide the user with additional information about cellular structure and compartmentalization.

After uploading the data, the KEGG pathway nodes, which correspond to genes or gene families, can be overlaid with fold-changes measured on mRNA or protein expression level. Furthermore, DNA methylation changes observed in the proximal promoter regions can be visualized. In the final step, additional nodes corresponding to miRNAs can be added to the pathway and colored according to the expression changes measured in the underlying experiment. See Figure 1 for an illustration of all the above-mentioned visualization steps.

2.1 Pathway visualization

The basic prerequisite for generating a pathway-based visualization is visualizing the pathway itself. For this purpose, we are using KEGGtranslator (see Wrzodek *et al.*, 2011), which performs a basic conversion of the KEGG KGML documents to GraphML and to annotates all nodes with EntrezGene identifiers (Ref. zu NCBI EntrezGene). In short, KEGGtranslator converts all KGML entries to nodes and all relations to edges. Some basic errors are corrected automatically and appropriate shapes, colors and labels are inferred. Then, all nodes are annotated with diverse identifiers, descriptions, and further information. The resulting document provides the basis for the subsequently generated visualizations.

At least for some pathways the KGML document available at KEGG does not contain all information displayed in the corresponding pathway map. Thus, an overlay graph can be generated which contains the original pathway map as a static transparent image in the background of the interactive graph plot (see Figure 1b). Owing to this feature, additional information on cellular structure (e.g., schematic drawings of receptors involved in cell signaling) can be maintained. This feature also enhances the visualization of the global maps in KEGG PATHWAY, as additional information on the organisation of subordinate pathway groups is provided by the pathway image.

2.2 Visualization of messenger-RNA expression data

As mRNA expression data is typically available for the whole genome, and thus also for the majority of nodes in a particular KEGG pathway, these data are displayed in the background color of the nodes.

As input our method requires preprocessed mRNA datasets with annotation columns (e.g., probeset and gene identifiers) and data columns, which are referred to as *observations*. In this context, observations can be any statistical significance (e.g., p-values) or comparative measure (e.g., fold-changes or log-ratios).

Next, these data have to be broken down to a single value for each pathway node, which then corresponds to the color of the node. As single nodes can represent multiple genes in KEGG, the intensities measured by probes, corresponding to the same node, have to be summarized. To this end, the mean or median is calculated across probes, which were mapped to the same node, or the probe with the strongest or most significant signal (i.e., $\min p - value$ or $\max |fold - change|$) is adopted for coloring the node.

For fold-changes (which are usually log_2 values), we color every node with a fold-change ≥ 2 red and all fold-changes ≤ -2 blue. Fold-changes of zero are defined to have a white color and colors between ± 2 are faded from blue or red to white, depending on the actual fold-change. The same procedure can be used for p-values, except that just one minimum threshold and one minimum color must be defined. Furthermore, the color for p-values should not be changed on a linear, but on a log-scale. See Figure 1c) for an example of visualized mRNA data.

2.3 Visualization of protein and protein modification expression data

Visualization of protein datasets is performed by adding small boxes below pathway nodes and changing the color of the boxes according to the corresponding protein expression data. Protein datasets usually have identifiers, like Entrez Gene IDs, UniProt IDs, etc. which allows to make a straightforward mapping to pathway nodes. Then, all values must be collected and a color must be calculated for each node in the same way as already described for mRNA datasets.

Protein modification datasets must be treated differently. They usually not only contain one expression values for the basic form of the protein, but also for some phosphorylated or likewise modified form. Therefore, separate boxes are created below each pathway node for all modifications. These boxes are labeled according to the modification. Furthermore, the color for each box must only be calculated on probes that match the pathway node and the modification of each box.

2.4 Visualization of DNA-methylation data

Ein Wert nur als Hinweis, hier geht etwas [click gibt details?]. fold-change wird zu box von -2 bis +2, p-value im grunde ein bar-blot von 1 bis 0.00005 oder so...

Einzelner Wert mit binning und $\frac{\sum_{i=1}^{n} \log_2 x}{n}$, fr fold-changes oder so peak detection mglich und max. peak anzeigen.

2.5 Visualization of micro-RNA expression data

Visualizing micro RNA (miRNA) datasets is not straightforward, because pathways usually do not contain miRNAs. Pathway mainly consist of small molecules and enzymes, which are products of protein coding genes. Therefore, to add miRNAs to a pathway, a connection must be established from miRNAs to to protein coding genes.

Biologically, miRNAs are small non-coding RNAs that regulate gene expression by binding to mRNA targets and somehow inducing a degradation of the targeted mRNA (Bartel, 2004). The targets for each miRNA must be known and there are several databases that contain information about experimentally verified miRNA targets (e.g., miRecords (see Xiao *et al.*, 2009), miRTarBase (see Hsu *et al.*, 2011), TarBase (see Papadopoulos *et al.*, 2009)) or predicted miRNA targets (Alexiou *et al.*, 2009). We use these miRNA target databases to perform the linkage between miRNAs and pathways.

Pathway-based visualization of miRNA datasets is done by annotating all known targets to every miRNA in the input dataset. Then all miRNAs that have targets in any pathway that should be visualized are added to the pathway as small triangular nodes. The relation to the pathway is then established by creating an edge from every miRNA to every target in the pathway. The triangular miRNA nodes themselves are colored according to their expression, as described for mRNA. This leads to an integrated visualization that contains miRNA expression, miRNA target relation information and (if also mRNA data is visualized) the expression of the targeted mRNA. Figure 1f) shows an example result of the described procedure.

3 RESULTS AND DISCUSSION

4 CONCLUSION

ACKNOWLEDGEMENT

We gratefully acknowledge contributions from Andreas Dräger and Finja Büchel, as well as the whole MACRCAR consortium.

Funding: The research leading to these results has received funding from the Innovative Medicine Initiative Joint Undertaking (IMI JU) under grant agreement nr. 115001 (MARCAR project).

REFERENCES

Alexiou, P., Maragkakis, M., Papadopoulos, G. L., Reczko, M., and Hatzigeorgiou, A. G. (2009). Lost in translation: an assessment and perspective for computational microrna target identification. *Bioinformatics*, **25**(23), 3049–3055.

Bartel, D. P. (2004). Micrornas: genomics, biogenesis, mechanism, and function. *Cell*, **116**(2), 281–297.

Hoheisel, J. D. (2006). Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet*, **7**(3), 200–210.

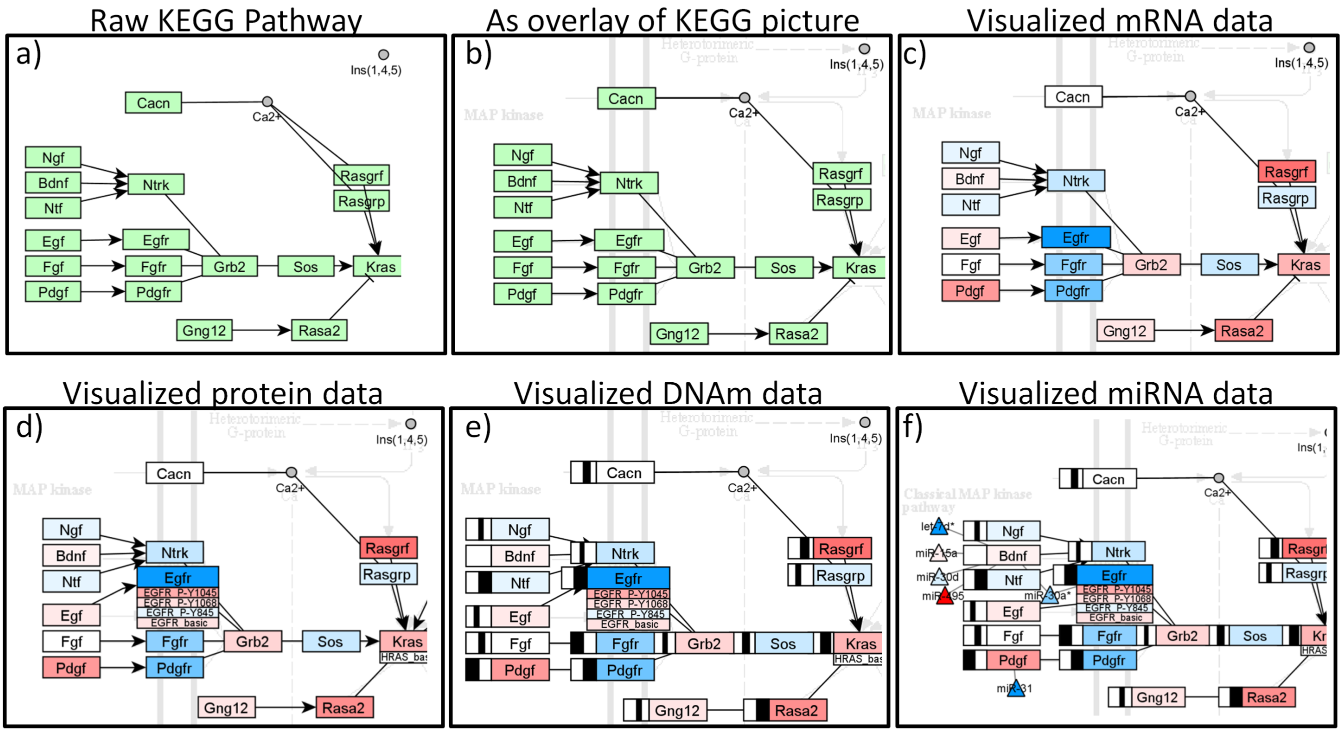
Hsu, S. D., Lin, F. M., Wu, W. Y., Liang, C., Huang, W. C., Chan, W. L., Tsai, W. T., Chen, G. Z., Lee, C. J., Chiu, C. M., Chien, C. H., Wu, M. C., Huang, C. Y., Tsou, A. P., and Huang, H. D. (2011). miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res*, **39**(Database issue), D163–9. Hsu, Sheng-Da Lin, Feng-Mao Wu, Wei-Yun Liang, Chao Huang, Wei-Chih Chan, Wen-Ling Tsai, Wen-Ting Chen, Goun-Zhou Lee, Chia-Jung Chiu, Chih-Min Chien, Chia-Hung Wu, Ming-Chia Huang, Chi-Ying Tsou, Ann-Ping Huang, Hsien-Da England Nucleic Acids Res. 2011 Jan;39(Database issue):D163-9. Epub 2010 Nov 10.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res*, **12**(6), 996–1006.

Papadopoulos, G. L., Reczko, M., Simossis, V. A., Sethupathy, P., and Hatzigeorgiou, A. G. (2009). The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res*, **37**(Database issue), D155–8. Papadopoulos, Giorgos L Reczko, Martin Simossis, Victor A Sethupathy, Praveen Hatzigeorgiou, Artemis G England Nucleic Acids Res. 2009 Jan;37(Database issue):D155-8. Epub 2008 Oct 27.

Wrzodek, C., Drger, A., and Zell, A. (2011). Keggtranslator: visualizing and converting the kegg pathway database to various formats. *Bioinformatics*, **27**(16), 2314–2315.

Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., and Li, T. (2009). miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res*, **37**(Database issue), D105–10. Xiao, Feifei Zuo, Zhixiang Cai, Guoshuai Kang, Shuli Gao, Xiaolian Li, Tongbin 1R21CA126209/CA/NCI NIH HHS/ R43 GM076941/GM/NIGMS NIH HHS/ England Nucleic Acids Res. 2009 Jan;37(Database issue):D105-10. Epub 2008 Nov 7.



Raw KEGG Pathway

As overlay of KEGG picture

Visualized mRNA data

Visualized protein data

Visualized DNAm data

Visualized miRNA data

Fig. 1. TODO: Bild zeigt AUSSCHNITT des MAPK signaling pathway. Alle Schritte von a) bis f) kurz erlauetern und kurz sagen, dass f) quasi dem finalen Bild dieser Methode entspricht.