BMC Bioinformatics

BioMed Central

Oral presentation

Public microarray repository semantic annotation with ontologies employing text mining and expression profile correlation

David Ruau*[1], Corinna Kolárik[2,3], Heinz-Theodor Mevissen[2], Emmanuel Müller[4], Ira Assent[4], Ralph Krieger[4], Thomas Seidl[4], Martin Hofmann-Apitius[2,3] and Martin Zenke[1,5]

Address: [1]Department of Cell Biology, Institute for Biomedical Engineering, RWTH Aachen, University Medical School, 52074 Aachen, Germany, [2]Fraunhofer Institute SCAI, Schloss Birlinghoven, 53754 Sankt Augustin, Germany, [3]Bonn-Aachen International Center for Information Technology (B-IT), Department of Applied Life Science Informatics, 53113 Bonn, Germany, [4]Data management and data exploration group, RWTH Aachen University, Germany and [5]Helmholtz Institute for Biomedical Engineering, RWTH Aachen University, 52074 Aachen, Germany

Email: David Ruau* - David.Ruau@rwth-aachen.de

* Corresponding author

from Fourth International Society for Computational Biology (ISCB) Student Council Symposium
Toronto, Canada. 18 July 2008

Published: 30 October 2008

BMC Bioinformatics 2008, **9**(Suppl 10):O5 doi:10.1186/1471-2105-9-S10-O5

This abstract is available from: http://www.biomedcentral.com/1471-2105/9/S10/O5

Public microarray repository annotation

Gene Expression Omnibus (GEO) [1] is the largest public web repository of microarray experiments. GEO, like ArrayExpress and Stanford MicroArray Database, provides descriptions of microarray experiments in free text making it difficult to search and comprehensively link those data to other knowledge resources. Text mining techniques applied to microarray experiment annotation are challenged by poor and/or ambiguous free text description and consequently leave some objects unlabelled. Previous work organized GEO entries at the level of series (GSE) and data sets (GDS) [2] using the Unified Medical Language System (UMLS) [3]. GSE and GDS description are often too broad and a better quality of annotation can be achieved if the GEO samples (GSM) are considered directly. Here we report on a novel approach for annotating GSM objects by employing a combination of text mining and global gene expression similarity. We hypothesize that the biological material analyzed on microarrays is related if unlabeled and labeled objects are highly similar in expression values and hence the class/annotation of one object can help annotate an unlabeled object. Our new method allows us to achieve a higher percentage of semantic annotation by combining both types of information stored in microarray databases.

Results

The GSM free text description (downloaded from GEO in November 2007) was mined using ProMiner [4], a software for Named Entity Recognition based on dictionaries of cell, tissue and disease ontologies from OBO [5] plus cell line resources. This resulted in 73.5–97.6% class labeling of the GSM objects (Table 1). Next the labeled objects were used to annotate the unlabeled objects. We computed the correlation matrix for all the objects where the raw data were available and followed the nearest neighbor approach [6] to identify the nearest labeled object within a δ range. The δ value is an input parameter determined empirically and limits the propagation of too dissimilar annotations. In this study we selected a delta value of 0.04 and observed an increase of the annotation percentage up to 4.9%, depending on the platform. The class labeling overall percentage after annotation propagation reached 78.4–99.4% (Table 1). The results were then stored into a relational database allowing to semantically search for microarray experiments.

Conclusions and perspectives

The class/annotation propagation from a labeled object to an unlabeled object works only if there is one labeled object within the δ range. Thus the chances of class prop-

Table 1: GSM object annotation coverage.

Platform	GSM object number (a)	Labeled objects by ProMiner (b)	Raw data available/ProMiner labeled (c)	Propagated annotation (d)
mouse4302 (GPL1261)	4288	87.2%	3362/2852	145 [90.6%]
mouse430a2 (GPL339)	3521	91.1%	321/288	25 [91.8%]
moe430b (GPL340)	795	97.6%	545/524	14 [99.4%]
mgu74av2 (GPL81)	4676	86.2%	2390/2120	183 [89.5%]
Hgu95av2 (GPL91)	4512	73.5%	1865/1325	221 [78.4%]

We performed our study on a subset of platforms from Affymetrix. (a) Total number of GSM objects present in GEO per platform (November 2007). (b) Percentage of objects labeled by ProMiner. (c) Objects with raw data stored in GEO and number of labeled objects with raw data. (d) Number of propagated annotations and final percentage of labeled objects

agation increase with the number of available objects. We plan to improve on this by merging different types of microarray platforms by using tools like AILUN [7] as well as adding a confidence score to the propagated annotations. Ultimately, the annotation process will be automatized and the resulting database made freely available.

References

1. **Gene Expression Omnibus** [http://www.ncbi.nlm.nih.gov/geo/]
2. Butte AJ, Kohane IS: **Creation and implications of a phenome-genome network.** *Nat Biotechnol* 2006, **24:**55-62.
3. **Unified Medical Language System** [http://www.nlm.nih.gov/research/umls/]
4. Hanisch D, Fundel K, Mevissen HT, Zimmer R, Fluck J: **ProMiner: rule-based protein and gene entity recognition.** *BMC Bioinformatics* 2005, **6(Suppl 1):**S14.
5. **The Open Biomedical Ontologies** [http://obofoundry.org/]
6. Dasarathy BV: *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques IEEE Computer Society Press*; 1991.
7. Chen R, Li L, Butte AJ: **AILUN: reannotating gene expression data automatically.** *Nature Methods* 2007, **4:**879.