Literature-Based Enrichment of Experimental Results

Marvin Meeng[1], Arno Knobbe[1], Peter-Bram 't Hoen[2], and Ko Willems van Dijk[2]

[1] LIACS, Leiden University, the Netherlands, meeng@liacs.nl
[2] LUMC, Leiden University, the Netherlands

In this abstract, we outline a method for enriching results from high-throughput genomics, proteomics and metabolomics experiments using a number of sources of background knowledge from the biological and medical domain. The method brings together a number of recent developments in the fields of Data Mining and Text Mining. From the former field, we employ state-of-the-art Subgroup Discovery techniques that form a generalisation of standard enrichment techniques (such as SEGS [8]). These SD techniques not only allow testing of enriched binary concepts, such as functional annotations from Gene Ontology (GO) [1], but also real-valued associations between the entities under examination and biological concepts, and additionally, more complex combinations thereof. From the Text Mining field, we employ recent techniques [2, 3] for obtaining large thesauri of common concepts in different biological domains, and derive the strength of association between such concepts and entities such as genes and metabolites. Each association between a concept (essentially a common phrase or synonym thereof) and an entity captures similarities in the literature contexts in PubMed abstracts. As such, we do not only consider strict functional annotations such as from GO or KEGG, but also include the possibility of finding new relationships between entities that have not been formally recognized. This not only allows us to test for non-standard or detailed domain-specific collections of concepts (for example, those related to neoplastic processes), but also to consider entities other than the usual genes, such as metabolites.

The process is outlined in Figure 1. The figure shows an integrated relational database that holds various sources of background knowledge. The sources include standard background information such as functional annotations from GO, UniProtKB/Swiss-Prot, protein-protein interaction networks and so on (middle stream). More importantly, the database includes the mentioned association matrices obtained by text mining, one for each thesaurus related to a specific domain (upper stream). Association matrices are typically fairly large, associating several thousand concepts with some 23k genes. Matrices are available for thesauri describing e.g. diseases, tissues or neoplastic processes, but also for concepts taken from the GO domains, such as *biological process*. Note that these GO-derived thesauri should not be confused with the actual annotations available from GO.

The lower stream of Figure 1 describes how primary experimental data, e.g. microarray data, are analysed for differential expression, resulting in a ranking of genes (or alternative entities). Using a common identifier scheme based on so-called concept unique identifiers (CUIs), the entity ranking and the background knowledge are connected in the integrated database, after which the actual enrichment can take place. The enrichment is performed using a generic data mining algorithm, called Subgroup Discovery. Our implementation of this algorithm, that is able to deal with the challenges of this data (large, heterogeneous and partly numeric), is available through the Safarii system [4]. How Safarii implements enrichment is outlined in [6].

To summarize, the novelty of our approach lies in the integration of two established techniques, from the fields of Text Mining and Data Mining respectively. As a result of this integration, a variety of entities (not just genes) can now be enriched, using a potentially richer and more up-to-date source of background knowledge (the literature), compared to standard GO-based enrichment. The integrated database enables the mining of a heterogeneous collection of background sources (for example both concepts and GO annotations), an unlimited range of association-levels, and retrieval of multiple (and mixed) conditions, e.g. ($\textit{'steroid catabolism'} \geq 0.0052 \wedge \textit{GO:0016337 cell-cell adhesion}$).

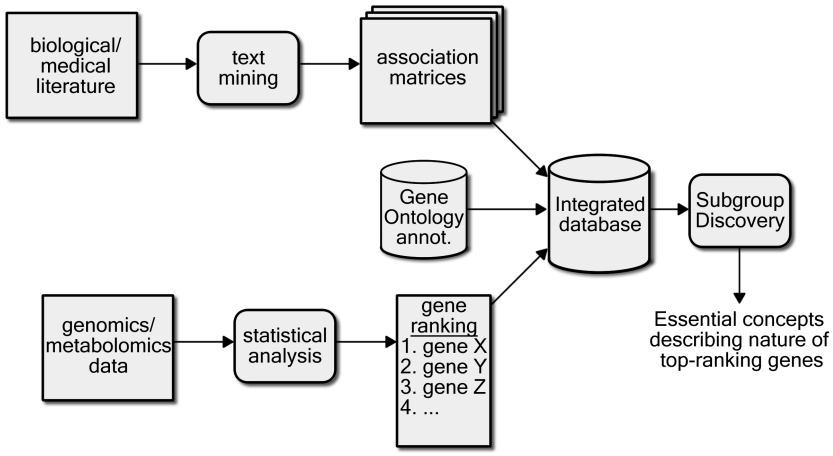

Fig. 1. Overview of the entire process.

1 Experiments

In this section, results from two different case studies will be presented. The first deals with data obtained in a *metabolic syndrome* study. This study demonstrates that our method is not limited to genes, but can be applied to any biological entity, metabolites in this case, as long as background information is available for it. The second study deals with a particular type of cancer known as *neuroblastoma*. Results will be presented that contrast the results of enrichment using traditional GO-terms and the new text-mined matrices based on the literature, for a well-annotated domain such as oncology.

1.1 Case study 1: Metabolic Syndrome

In the first study, metabolite and mRNA expression profiles in mice on two distinct high-fat diets were contrasted with a control group on a normal diet. The high-fat HFL and HFW diet consisted of animal fat (beef tallow), and plant fat (palm oil), respectively. The CHOW control diet is a standardized mouse diet. In our analyses, each high-fat diet is separately contrasted with the CHOW diet, as the course of development of the metabolic syndrome differs between the two high-fat groups [7].

Liver mRNA expression profiles were measured on microarrays. Two sources of background information were used to find biological processes associated with differential mRNA expression between the mice on the different diets. The first source comprises the standard GO annotations for Biological

Table 1. *CHOW vs. HFL* expression data, enriched using either concepts or GO-annotations.

rank	coverage	score	concept (CUI)	coverage	score	annotation (GO)
1	191	0.123	'*steroid catabolism*'	23	0.064	lipid biosynthetic process
2	807	0.123	'*anti-inflamm. response*'	12	0.061	induction of apoptosis
3	509	0.121	'*Drug Transport*'	100	0.052	oxidation reduction
4	50	0.121	'*progesterone catabolism*'	5	0.048	mammary gland development
5	123	0.120	'*canal. bile acid transport*'	5	0.048	antigen processing & presentation
6	617	0.117	'*spermine catabolism*'	11	0.047	cell-cell adhesion
7	187	0.117	'*sterol catabolism*'	5	0.045	cellular amino acid biosyn. process
8	57	0.115	'*protein retention in ER*'	14	0.044	cell communication
9	718	0.115	'*xenobiotic transport*'	9	0.043	fatty acid biosyn. process
10	311	0.114	'*vitamin K metabolism*'	5	0.041	endoplasmic reticulum unfolded protein response

Table 2. Results for *CHOW vs. HFL.*

rank	ranking of metabolites
1	2-(Methylamino)isobutyric acid
2	Carnitine
3	Glycine
4	Ornithine
5	Oxoproline
6	Glucose
7	4-aminobenzoyl-glutamate
...	...
32	Lysine
33	Proline

rank	ranking of concepts
1	'*bile acid transport*'
2	'*regulation of metabolism*'
3	'*regulation of cholesterol absorption*'
4	'*phospholipid translocation*'
5	'*lipid digestion*'
6	'*hexose transport*'
7	'*glucose import*'
8	'*aminophospholipid transport*'
9	'*sodium-independent organic anion transport*'
10	'*organic acid transport*'

Processes associated with the differentially expressed genes, while the second source are the subset of literature concepts referring to the same Biological Processes. This allows a direct comparison between the results obtained. Table 1 shows the top scoring biological processes for the two types of approaches. Both the coverage (the number of genes associated with the biological process) and enrichment-score are much higher in the literature-based setting, which may be attributed to the fact that more domain-specific knowledge can be retrieved from the literature, while GO annotations should be highly generic.

By means of mass-spectrometry, the levels of 33 metabolites in blood serum of the same animals were measured. Unlike genes, metabolites are not annotated in GO and only the literature can be used to associate the metabolites with the biological processes associated with differential mRNA expression. Table 2 (left) shows the primary ranking of these metabolites, as produced by a t-test comparing the animals on CHOW and HFL. After integration and enrichment using Safarii, we obtained the literature-derived biological processes associated with the metabolites, shown in table 2 (right) (only top 10 displayed). Among the top concepts found, there is a significant number related to *metabolic syndrome* [7]. Bile acid transport and steroid (cholesterol) metabolism were found in the analysis of both metabolite and expression data, demonstrating the power of our approach to integrate genomics and metabolomics data.

1.2 Case study 2: Neuroblastoma

In the second case study, the data analysed were obtained from the *European Embryonal Tumour Pipeline (EETP)* project [5], which aimed to validate targets for novel cancer therapies and diagnostic development. The primary data set contains gene expression data of 251 neuroblastoma patients, and was analysed for two different targets that reflect the tumor status of the patients. Because of disseminated resistant tumour cells, stage 4 tumours frequently relapse, resulting in poor overall survival rates (<35%). Also, the frequency of spontaneous regression is much higher in the other stages. Results are shown for only one target, *Stage 4 vs. other*.

In figure 2, we show a comparison of the scores achieved using various sources of background knowledge. The main goal was to evaluate how the new literature-based enrichment would compare to a traditional GO-based enrichment in a well-annotated domain such as cancer. On the left, a comparison is shown of the various available literature domains. Using neuroblastoma-related domains like genes, neoplastic process, disease and biological process leads to much higher scores those for the tissue and cellular component domains. On the right, concepts (CUI) and annotations (GO) are compared. Although enrichment-scores are initially better using GO annotations, the scores drop much more rapidly, implying that the literature-based enrichment resulted in the finding of many more relevant literature concepts. Furthermore, the top-ranking GO-terms were very generic (e.g. *Cell Cycle*), and therefore less informative than the literature-derived concepts.

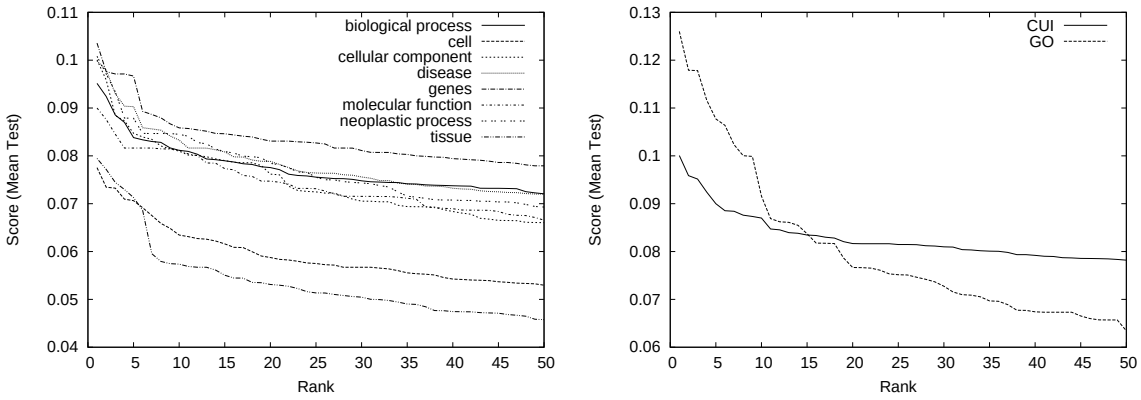

Fig. 2. Score comparison of CUI domains (left) and CUI vs. GO (right).

2 Conclusion

The results presented in Section 1 are relevant to the metabolic syndrome and neuroblastoma domains, highlighting that our new approach helps the interpretation of high dimensional datasets. Furthermore, it was shown that the new method is capable of producing relevant results for biological entities other than genes, which is a clear advantage over other enrichment methods. It is not only possible to analyse a wider variety of input data, using a broader set of background knowledge, but also to this uniform way, using the same tools, which would be helpful for data integration. Our findings seem to indicate that, at least for the domain of metabolic syndrome, the use of concept profiles extracted from (recent) scientific literature yields a better enrichment result in terms of relevant concepts found, as opposed to enrichment using annotation-based background knowledge. This may be attributed to the fact that biological processes in this domain are less well annotated in public databases. For the well-annotated cancer domain, GO annotations scored better initially, but the performance drops much more rapidly than with the literature-based enrichment. We will start exploring the obvious benefits of the use of multiple background resources simultaneously in our integrated system.

Acknowledgements

We acknowledge financial support from the European Community's Seventh Framework Programme (FP7/2007-2013)-funded ENGAGE project (grant agreement HEALTH-F4-2007-201413) and the Netherlands Consortium for Systems Biology.

References

1. The Gene Ontology, 2009. http://www.geneontology.org.
2. Jelier, Schuemie, Veldhoven, et al., Anni 2.0: a multipurpose text-mining tool for the life sciences, Genome Biol 2008;9(6):R96.
3. Jelier, Goeman, Hettne, et al. *Literature-aided interpretation of gene expression data with the weighted global test*, Briefings in Bioinformatics, doi:10.1093/bib/bbq082, 2010.
4. Knobbe, *Safarii Multi-Relational Data Mining system*, www.kiminkii.com/safarii.html.
5. van de Koppel, Slavkov, Astrahantseff, Schramm, Schulte, Vandesompele, de Jong, Dzeroski & Knobbe, *Knowledge Discovery in Neuroblastoma-related Biological Data*, Data Mining in Functional Genomics and Proteomics workshop at ECML PKDD 2007.
6. Pieters, Knobbe & Džeroski, *Subgroup Discovery in Ranked Data, with an Application to Gene Set Enrichment*, Preference Learning workshop at ECML PKDD 2010.

7. Radonjic, de Haan, van Erk, et.al, *Genome-Wide mRNA Expression Analysis of Hepatic Adaptation to High-Fat Diets Reveals Switch from an Inflammatory to Steatotic Transcriptional Program*, PLoS ONE 4(8): e6646. doi:10.1371/journal.pone.0006646, 2009.
8. Subramanian, et al. *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.* PNAS October 25, 102:43, 2005.