Probabilistic Retrieval and Visualization of Biologically Relevant Microarray Experiments

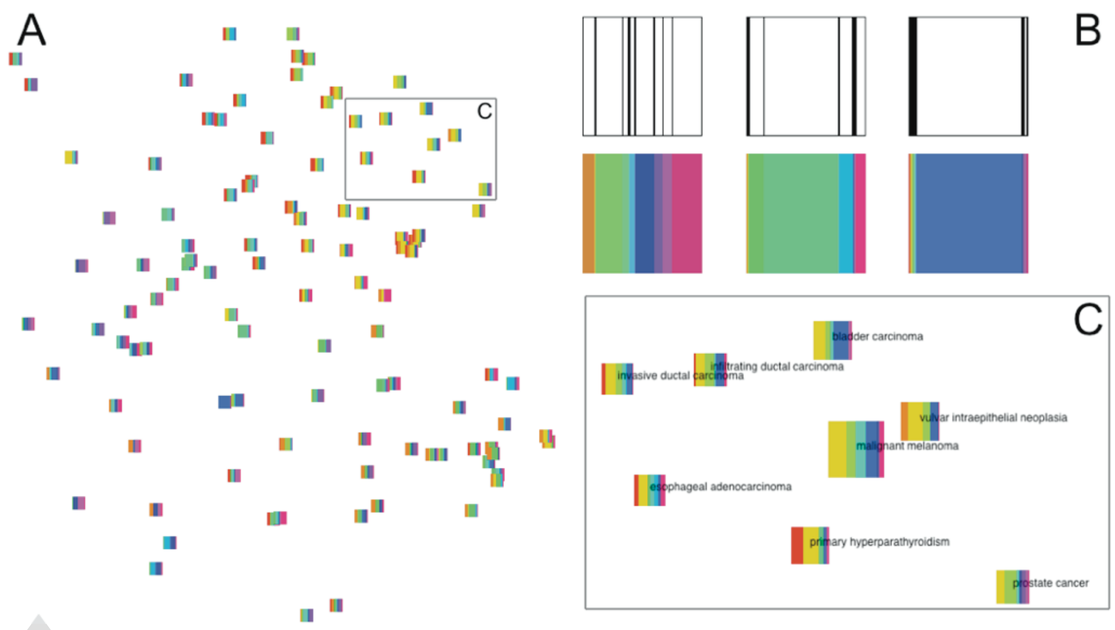
José Caldas[1], Nils Gehlenborg[2,3], Ali Faisal[1], Alvis Brazma[2] and Samuel Kaski[1]

[1] Helsinki Institute for Information Technology, Department of Information and Computer Science, Helsinki University of Technology, Finland; [2] Microarray Team, European Bioinformatics Institute, Cambridge, UK; [3] Graduate School of Life Sciences, University of Cambridge, Cambridge, UK

Contact: jose.caldas@tkk.fi

Presentation: Proceedings Track PT21, June 30th, 11:45 am - 12:10 pm, Victoria Hall.

In a Nutshell: We propose methods for retrieving and visualizing related experiments in large microarray gene expression repositories such as ArrayExpress [1], taking into account similarity in actual measurement data instead of textual annotations. Our approach is based upon the probabilistic clustering [2] of gene set differential expression patterns [3]. Case studies highlight the method's capability of retrieving studies related in non-trivial ways to query experiments.

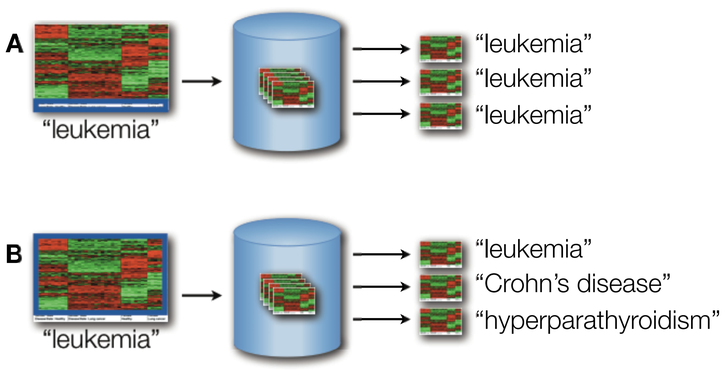


A NeRV projection [4] of 105 experiments, each shown as a glyph. **B** The slices of each glyph show the distribution of topics in three example experiments. **C** Enlarged region from A where glyphs have additionally been scaled according to their relevance to the query with the 'malignant melanoma' experiment shown in the center.

Background

As genome-wide expression studies in repositories such as ArrayExpress [1] accumulate, it becomes more important to develop methods that search for relevant experiments given a particular study. Textual description and experimental design are not as informative as actual measurement data, and are limited and biased by the scope of each experiment.

We introduce novel retrieval methods that incorporate the actual gene expression measurements into the search process, along with visualization tools for interpreting and exploring the results.



Difference between **A** Annotation-based queries and **B** Measurement data-based queries.

Methods

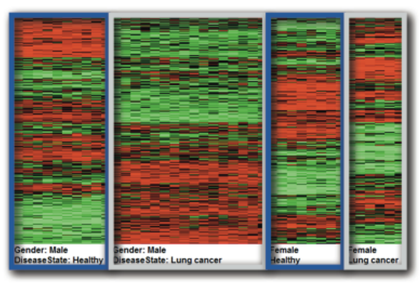
In our proposed approach we decompose each study into a collection of comparisons between phenotypes, e.g. healthy vs disease.

1 Gene Set Differential Expression: First, encode each experiment as a vector of differentially expressed gene sets, using a recent nonparametric statistical method [3].

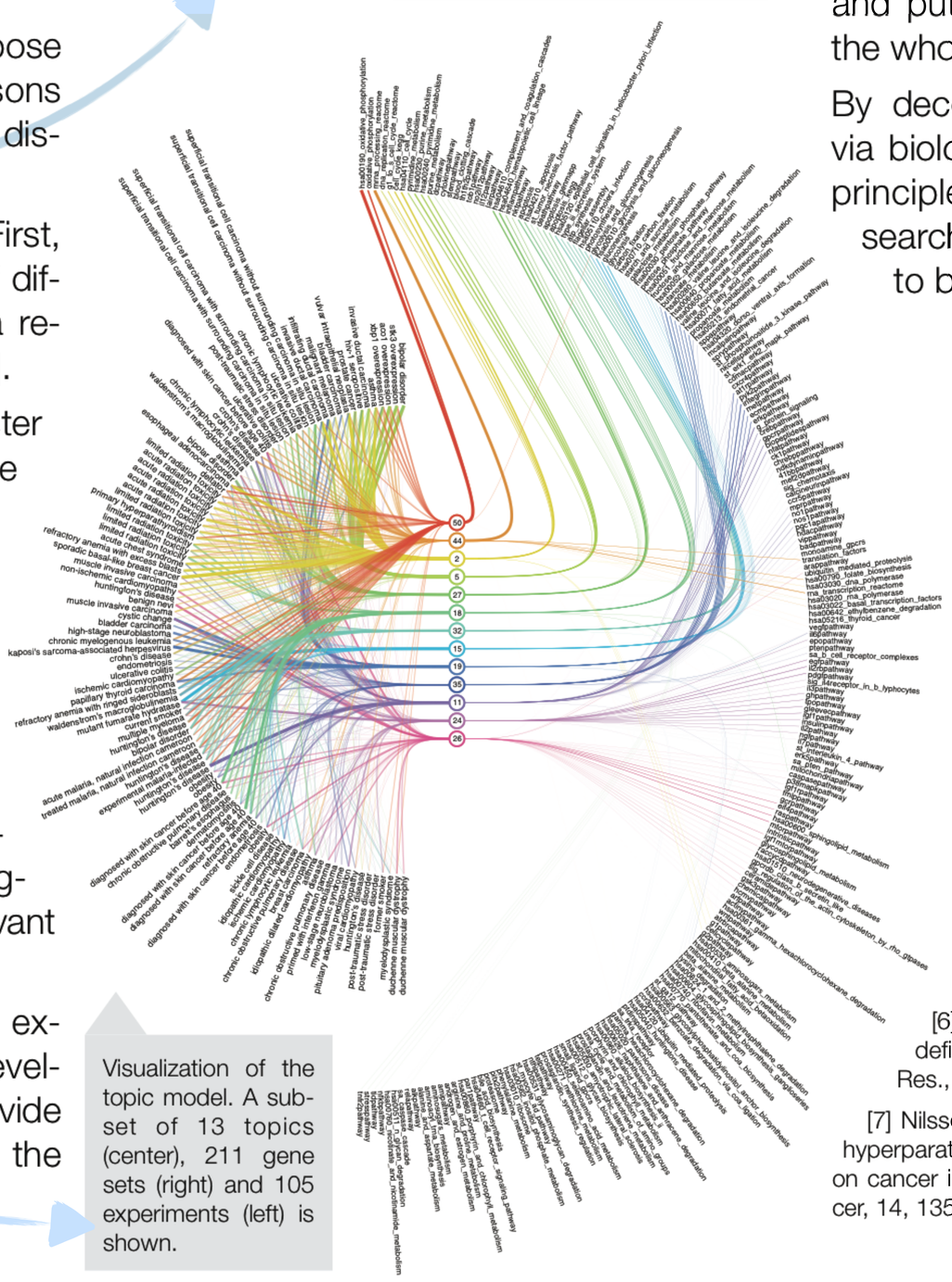
2 Topic Model: Probabilistically cluster encoded experiments with a generative model known as Latent Dirichlet Allocation (LDA) [2]. The output of LDA is a collection of so-called topics, probability distributions over gene sets that represent differential expression patterns, along with soft assignments of experiments to topics.

3 Retrieval: The probabilistic formulation enables the use of a natural and rigorous metric for assessing how relevant each experiment is to a query study.

4 Visualization: For interpretation and exploration of retrieval results we have developed visualization methods that also provide insight into the model used to perform the retrieval [4].



Decomposition of a gene expression study into comparisons between phenotypes.



Visualization of the topic model. A subset of 13 topics (center), 211 gene sets (right) and 105 experiments (left) is shown.

Results

Gene sets from each biological topic form coherent and holistic components. Case studies on ArrayExpress show that the method mostly retrieves cancer experiments when queried with a cancer study, and relates cancer to pathological entities such as Crohn's disease [5,6] and hyperparathyroidism [7]. Information retrieval measures such as average precision and precision-recall curves show that the method's performance is significant. The visualization methods allow us to both efficiently interpret the model and put retrieval results into the context of the whole set of experiments .

By decomposing and relating experiments via biologically meaningful components in a principled manner, our approach allows search within a gene expression database to be driven by the measured data.

References

[1] Parkinson, H. *et al.* (2009) ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res., 37, D868–D872.

[2] Blei, D. *et al*. (2003) Latent Dirichlet allocation. J. Mach. Learn. Res., 3, 993–1022.

[3] 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.

[4] Venna, J. and Kaski, S. (2007) Nonlinear dimensionality reduction as information retrieval. In Meila,M. and Shen,X. (eds), AISTATS'07. Omnipress, San Juan, Puerto Rico.

[5] Hoffbrand, A. V. *et al.* (1968) Folate deficiency in Crohn's disease: incidence,pathogenesis, and treatment. Br. Med. J., 2, 71–75.

[6] Au, W. Y. *et al.* (2009) Cough mixture abuse, folate deficiency and acute lymphoblastic leukemia, Leukemia Res., 33, 508–509.

[7] Nilsson, I.-L. *et al.* (2007) The association between primary hyperparathyroidism and malignancy: nationwide cohort analysis on cancer incidence after parathyroidectomy. Endocr. Relat. Cancer, 14, 135–140.