BMC Bioinformatics

Poster presentation

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]

Address: [1]Helsinki Institute for Information Technology, Department of Information and Computer Science, Helsinki University of Technology, P.O. Box 5400, Helsinki, FI-02015 HUT, Finland, [2]European Bioinformatics Institute, Cambridge, CB10 1SD, UK and [3]Graduate School of Life Sciences, University of Cambridge, Cambridge, CB2 1RX, UK

E-mail: José Caldas* - jose.caldas@tkk.fi; Nils Gehlenborg - nils@ebi.ac.uk; Ali Faisal - ali.faisal@tkk.fi; Alvis Brazma - brazma@ebi.ac.uk; Samuel Kaski - samuel.kaski@tkk.fi
*Corresponding author

from Fifth International Society for Computational Biology (ISCB) Student Council Symposium
Stockholm, Sweden 27 June 2009

Published: 19 October 2009

BMC Bioinformatics 2009, **10**(Suppl 13):P1 doi: 10.1186/1471-2105-10-S13-P1

This article is available from: http://www.biomedcentral.com/1471-2105/10/S13/P1

Background

Repositories of genome-wide expression studies such as ArrayExpress [1] have been growing rapidly over the last few years and continue to do so. The more experimental data are deposited into these repositories, the more likely it becomes that some of them can provide a meaningful biological context to aid in the planning and analysis of new studies. Retrieval of experiments based on their textual description and experimental design has several shortcomings. First of all, textual description of an experiment or its results is not as information-rich as the actual data itself. Secondly, information about the experimental design alone is only of limited use in retrieving biologically relevant data because it does not reflect the results, which contain the bulk of the information and may reveal unexpected relationships. 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 [2].

Methods

We developed a two-stage procedure, first identifying differentially active gene sets in each experiment using a recent nonparametric statistical method [3], and then combining gene set activation patterns into higher-level structures, so-called biological topics, using a state-of-the-art probabilistic model [4]. The probabilistic formulation enables the use of a natural and rigorous metric for assessing the similarity between two experiments. For interpreting and exploring retrieval results, we have developed visualization methods that also provide insight into the model used to perform the retrieval.

Results

We show that gene sets corresponding to each biological topic form highly coherent and holistic components. Several case studies performed on a subset of ArrayExpress show that our method can retrieve experiments relevant to a biological question, as long as sufficient amounts of data are available, and highlight relations between experiments, either because the same biological questions were targeted, or because of unexpected relationships that were confirmed in the literature. The visualization methods allow us to both efficiently interpret the model and put retrieval results in the context of the whole set of experiments (see Figure 1 for an example).

Conclusion

Using a combination of existing and novel methods for modeling and visualizing a heterogeneous collection of gene expression experiments, we were able to

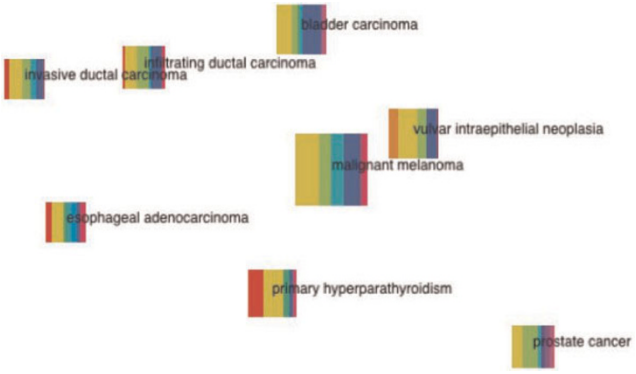

Figure 1
2D NeRV projection of retrieval results when the model is queried with a malignant melanoma experiment. Each experiment is represented as a striped glyph. Colors indicate biological topics. Stripe widths indicate the predominance of each biological topic in each experiment. Glyph size indicates relevance to the malignant melanoma query experiment [5].

decompose and relate experiments via biologically meaningful components. Our approach allows search within a gene expression database to be driven by actual measurement data.

Acknowledgements

This work was supported by TEKES (grant no. 40101/07). JC, AF and SK are additionally partially supported by PASCAL 2 Network of Excellence, ICT 216886. JC is additionally supported by a doctoral grant from the Portuguese Foundation for Science and Technology (FCT). NG is supported by a PhD fellowship of the European Molecular Biology Laboratory (EMBL).

References

1. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Brube H, Dylab M, Emam I and Farne A: **ArrayExpress update – from an archive of functional genomics experiments to the atlas of gene expression.** *Nucleic Acids Res* 2007, **37(Database Issue):**D868–D872.
2. Caldas J, Gehlenborg N, Faisal A, Brazma A and Kaski K: **Probabilistic retrieval and visualization of biologically relevant microarray experiments. .** *Bioinformatics* 2009, **25:** i145–i153.
3. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gilette MA, Paulovich A, Pomeroy SL and Golub TR, *et al*: **Gene set enrichment analysis – a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102:**15545–15550.
4. Blei D, Ng A and Jordan MI: **Latent Dirichlet Allocation.** *J Mach Learn Res* 2003, **3:**993–1022.
5. Venna J, Kaski S and Jordan MI: **Nonlinear dimensionality reduction as information retrieval.** *AISTATS'07* 2007.