Robustness of modeling-based experiment retrieval to differences in measurement and preprocessing techniques

With the rapid progress in high-throughput measurement technologies, public repositories such as ArrayExpress and Gene Expression Omnibus continue to grow enormously with freely available gene expression datasets. This rapid growth poses a challenge in the retrieval of datasets relevant to a researcher for biological research. To overcome problems related to inaccurate or missing metadata, the task of querying a database of experiments using measurement data, instead of metadata, has recently received increased attention in the literature. Earlier studies of such methods have been conducted with a subset of gene expression experiments generated using one array design or generated with a dedicated preprocessed technique. However, public repositories continue to store experiments which have been conducted on various array designs and generated with various preprocessing techniques.

In our recent study, we evaluated a recently introduced model-distance-based method [1] together with two other content-based retrieval methods available in the literature to differences in preprocessing techniques (RMA, MAS5) and array designs (Affymetrix's Human Gene Array and Human Genome Arrays). The methods were (i) a model-distance-based method which represents each experiment as a probabilistic model and uses the distance between the models as a measure of relevance, (ii) a likelihood-based method [3] which is similar to the previous one but uses the likelihood of each model as the measure of relevance, and (iii) a non-probabilistic method [2] based on correlations between differential expression profiles. The experimental evaluations demonstrate that the model-distance-based method is tolerant to differences in preprocessing techniques and outperforms the remaining two methods. Furthermore, the model-distance-based method is tolerant to differences in Human Genome arrays and the results to differences between Human Genome arrays and Human Gene Array is inconclusive.

The model-distance-based method facilitates the retrieval of gene expression datasets, exhibiting similar expression patterns in different conditions, across the platforms and preprocessing techniques as long as sufficient amounts of data are available. While the method has currently only been tested on microarray gene expression experiments, there is ongoing work to extend it as a general purpose retrieval scheme for other experiment types. Thus, it leads to making maximal usage of data available in the public repositories to discover hidden patterns in biological mechanisms.

References

[1] Blomstedt, P., Dutta, R., Seth, S., Brazma, A., and Kaski, S. Modelling-based experiment retrieval: a case study with gene expression clustering. *Bioinformatics 32* (2016), 1388–1394.

[2] Engreitz, J. M., Morgan, A. A., Dudley, J. T., Chen, R., Thathoo, R., and Altman, Russ B.and Butte, A. J. Content-based microarray search using differential expression profiles. *BMC Bioinformatics 11* (2010), 603.

[3] Faisal, A., Peltonen, J., Georgii, E., Rung, J., and Kaski, S. Toward computational cumulative biology by combining models of biological datasets. *PLOS ONE 9*, 11 (2014), 1–17.