Toward computational cumulative biology by combining models of biological datasets

Ali Faisal[1,2], **Jaakko Peltonen**[1,2,3], **Elisabeth Georgii**[1,2], **Johan Rung**[4], **Samuel Kaski**[1,2,5]
1 Helsinki Institute for Information Technology HIIT
2 Department of Information and Computer Science, Aalto University
3 School of Information Sciences, University of Tampere
4 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI)
5 Department of Computer Science, University of Helsinki

Abstract

A challenge of data-driven sciences is how to make maximal use of growing databases of experimental datasets to keep research cumulative. We introduce a modeling-based dataset retrieval engine for relating a researcher's experimental dataset to earlier work in the field. The search is (i) data-driven to enable new findings, going beyond keyword searches in annotations, (ii) modeling-driven, to include both biological knowledge and insights learned from data, and (iii) scalable, as it is done without building one unified grand model of all data. We apply a rapidly computable and optimizable combination model to decompose a new dataset into contributions from earlier relevant models. We thus identify a network of interrelated datasets from a human gene expression atlas. While tissue type and disease were major driving forces, found relationships between datasets were richer, and model-based search was more accurate than keyword search; moreover, it recovered biologically meaningful relationships not straightforwardly visible from annotations. Data-driven links and citations matched to a large extent; the data-driven links even uncovered corrections to the publication data. This is a short version of our accepted PLOS ONE paper, arXiv version at [1].

1 Introduction

Molecular biology has been transformed into a data-driven science with as much importance given to computational and statistical analysis as to experimental design and assay technology. Challenges include processing of massive sequencing data and statistical challenges from having few samples and many variables. Many successful methods rely on increasing the effective number of samples by combining with similar experiments in a large meta-analysis [3], but this is not straightforward. Public repositories largely rely on annotation and meta-data from the submitter. Database curators and ontologies help in harmonizing and standardizing annotation, but the user who wants to find datasets combinable with her own most often must resort to searches in free text or in controlled vocabularies, needing much downstream curation and data analysis before any meta-analysis [4].

Instead of searching for similar dataset descriptions only, we wish to search in a data-driven way, querying with the dataset itself or its statistical description. This is implicitly done in multi-task learning which builds a unified model of datasets. But as the number of datasets and the amount of quantitative biological knowledge grow, building a unified model becomes computationally prohibitive. We consider the scenario where future researchers increasingly develop hypotheses in terms of (probabilistic) models of their data. A similar trend exists for sequence motif data, often published as Hidden Markov models [5]. We ask *what could be done with these models towards cumulatively building knowledge from data in molecular biology*? We propose *a modeling-driven dataset retrieval engine*, which a researcher can use for positioning her data into context of earlier

biology. Retrieval will be based on data, instead of keywords and ontologies, enabling unexpected novel findings. The retrieval will use the models of the datasets, thus utilizing their built-in knowledge, but will be more scalable than building a unified grand model of all data. Instead of matching single observations [2] whole datasets, incorporating the experimental designs, will be matched.

We explain a new dataset by a combination of models of earlier datasets and a novelty term. This mixture modeling scales well to large numbers of datasets, and modeling speed does not depend on sizes of earlier datasets. The largest weights point at the most relevant earlier datasets. Mixture components are stored models of each dataset, bringing their built-in knowledge. We apply the method to an atlas from ArrayExpress [6]. Earlier work is restricted to pairwise dataset comparisons: representing each dataset by pairwise correlations [7] needs many samples for good estimates and expensive computation in dataset comparisons; others assume specific case-control designs or known biological processes [8]. Ours is the first approach that allows data-driven retrieval of relevant datasets by decomposing a query dataset into contributions from several earlier datasets, without needing specific designs for earlier datasets or their models. Our approach is scalable and not limited to available dataset annotation; unlike Pfam [5] we use models in retrieval; we match whole datasets, not only individual observations; we fully decompose datasets instead of only computing pairwise similarities; and we allow arbitrary models without restrictive assumptions.

2 Combination of stored models for dataset retrieval

Our goal is to infer data-driven relationships between a new "query" dataset q and earlier datasets. The query is a dataset of N_q samples $\{x_i^q\}_{i=1}^{N_q}$; in our ArrayExpress study, samples are gene expression profiles, element x_{ij}^q being expression of gene set j in sample i of query q, but the setup applies to other data as well. Assume a dataset repository of N_S earlier datasets, where each dataset s_j, $j = 1, \ldots, N_S$, has already been modeled with a *base model* denoted by M^{s_j}. The base models are probabilistic generative models, capturing prior knowledge and data-driven discoveries. Base models for different datasets may come from different model families, as chosen by the dataset authors. We build a *combination model* for the query dataset as a mixture model of the base distributions $p(x|M^{s_j})$, parameterized by $\mathbf{\Theta}^q = \{\theta_j^q\}_{j=1}^{N_S+1}$. The likelihood of observing the query is

$$p(\{x_i^q\}_{i=1}^{N_q}; \mathbf{\Theta}^q) = \prod_{i=1}^{N_q} \Big[\Big(\sum_{j=1}^{N_S} \theta_j^q p(x_i^q|M^{s_j}) \Big) + \theta_{N_S+1}^q p(x_i^q|\psi) \Big] \tag{1}$$

where θ_j^q is the mixture proportion or *weight* of the jth base distribution (model of dataset s_j), and $\theta_{N_S+1}^q$ is the weight for the novelty term. The novelty is modeled by a background model ψ, a broad distribution covering overall gene-set activity across the repository. Weights are non-negative and $\sum_{j=1}^{N_S+1} \theta_j^q = 1$. This representation approximately explains biological activity in the query dataset as a combination of earlier datasets and a novelty term. For each query q, given the known models M^{s_j} of datasets in the repository, we infer a maximum a posteriori (MAP) estimate of the combination weights $\{\theta_j^q\}_{j=1}^{N_S+1}$. Alternatively, we could sample over the posterior, but MAP inference already yielded good results. We optimize the weights to maximize their posterior probability, proportional to $p(\{x_i^q\}_{i=1}^{N_q}; \mathbf{\Theta}^q) \cdot p(\{\theta_j^q\})$ where $p(\{\theta_j^q\}) = \mathcal{N}(0, \lambda^{-1}\boldsymbol{I})$ is a naturally non-sparse L_2 prior for the weights with regularization term λ. The cost is strictly concave and weights can be optimized by standard constrained convex optimization; details and and convergence proof for the Frank-Wolfe algorithm are given in [1]. After computing the MAP estimate, we rank the datasets for retrieval by decreasing weights. Advantages of our approach: 1) approximations become more accurate as more datasets enter the repository; 2) it is fast and scalable— computation time is linear in N_S and an approximate variant can run in sublinear time [1]; 3) any model types can be included, as long as likelihoods can be computed; 4) relevant datasets are not assumed naïvely similar to the query, they only need explain part of it; 5) relevance scores have natural meaning as mixture weights.

3 Results

We used the human gene expression atlas [6], ArrayExpress accession number E-MTAB-62. Data were preprocessed by gene set enrichment analysis using pathway collection C2-CP from the Molecular Signatures Database. Each sample was represented by top enriched gene sets.

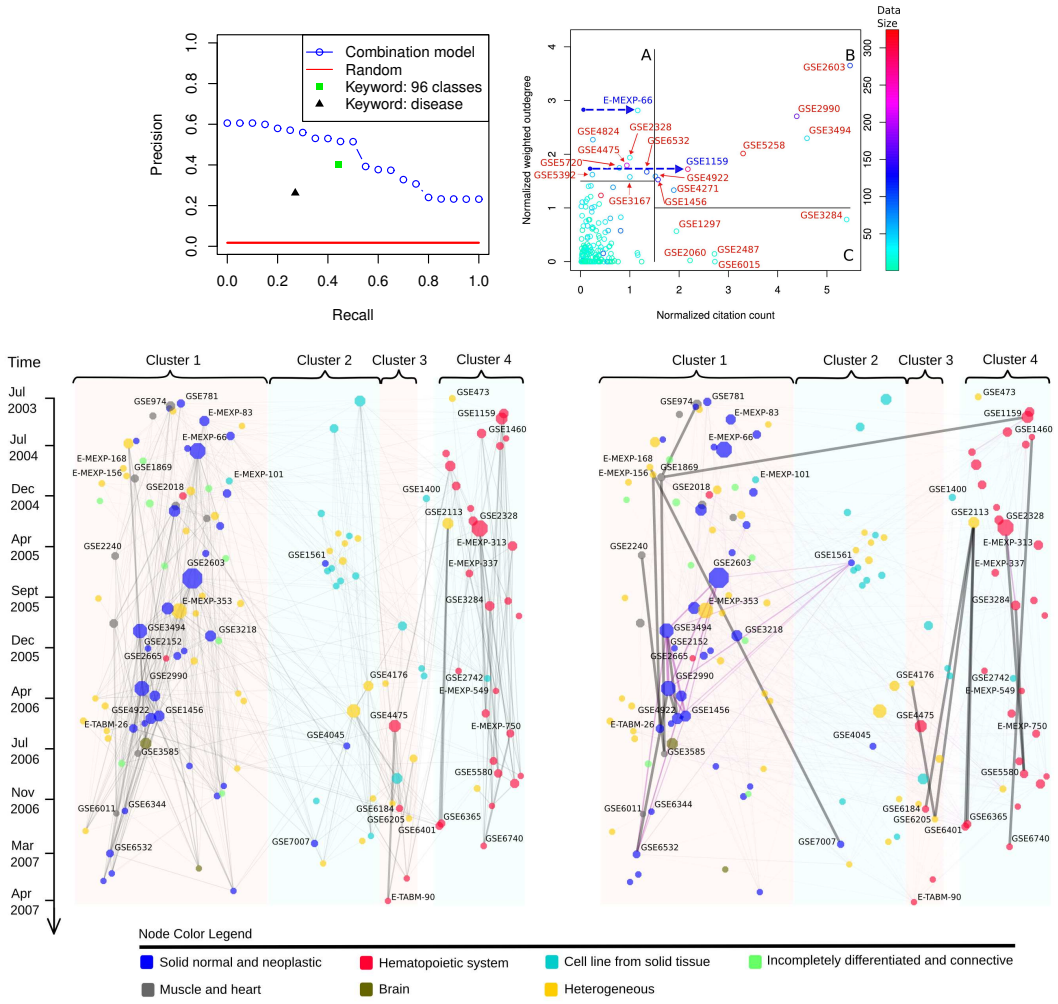

Figure 1: **Top left: Data-driven retrieval outperforms keyword search.** Blue: Precision-recall curve. Experiments sharing a biological category were considered relevant. In keyword retrieval, category names ("Keyword: 96 classes") or disease annotations ("Keyword: disease") were used as keywords. Datasets with ≥ 10 samples used as queries; curves are averages over queries. **Bottom: Relevance network of datasets.** Left: each dataset was used as a query to retrieve earlier datasets; a link from an earlier dataset to a later one means the earlier dataset is relevant as a partial model of activity in the later dataset. Link width is proportional to normalized relevance weight (θ_j^q; links with $\theta_j^q \geq 0.025$ shown, datasets without links discarded). Right: links are direct (gray) and indirect (purple) citations, node size proportional to estimated influence (total outgoing weight), colors: tissue types. Node layout details in [1]. **Top Right: Data-driven prediction of usefulness of datasets vs. their citation counts.** Manual checks comparing sets for which the two scores differed revealed inconsistent database records for two datasets; the blue arrows point to their corrected locations, which are more in line with the data-driven model. Regions A, B, and C: see [1].

Data-driven retrieval of experiments is more accurate than standard keyword search: We compared our model to state-of-the-art dataset retrieval by keyword search, in a scenario where a user queries with new datasets against a database of earlier released datasets. As base models, we used Latent Dirichlet Allocation and mixture of unigrams, for each data whichever had larger predictive likelihood. Retrieval by combination weight was consistently better than keyword search (Fig. 1, top left). We checked the result was not due to laboratory effects by discarding same-laboratory results: mean average precision decreased from 0.44 to 0.42 but supports the same conclusion. **Network of computationally recommended dataset connections reveals biological relationships:** When each dataset in turn is used as a query, the estimated combination weights form a "relevance net-

work" between datasets (Fig. 1, left side of the bottom subfigure), where each dataset is linked to the relevant earlier datasets. The network structure is dominated but not fully explained by tissue type. Normal and neoplastic solid tissues (cluster 1) are separate from cell lines (cluster 2) and hematopoietic tissue (cluster 4). The model has not seen the tissue types but found them from data. Finer structure is evident; muscle and heart datasets (gray) form a connected subnetwork with nodes at bottom of the image explained by upper nodes, those explained by nodes further up. Numerous links go across clusters and across tissue categories, e.g., the strongest link between two homogeneous datasets of different tissue types connects GSE3307 (comparing skeletal muscle samples from healthy individuals with patients having muscle diseases) to GSE5392 (measuring transcriptome profiles of normal brain and brain with bipolar disorder). Shortening of telomeres has been associated both with bipolar disorder muscular disorder, and treatment of bipolar disorder has been found to also slow down the onset of skeletal muscle disorder. We also investigated "outlier" datasets where the tissue type does not match the main tissue types of a cluster, and found reasonable explanations for them, see [1]. **Top dataset links overlap well with citation graph:** We compared the model-driven network to citation links (Fig. 1, right side of bottom subfigure) to find out to what extent citation practice matches the data-driven relationships. Of the top 200 data-driven edges, 50% overlapped with direct or indirect citation links. We compared densely connected sets of experiments between the two networks: e.g., for a citation clique of breast cancer datasets, and another clique of leukocyte datasets, for both cliques the corresponding edges in the relevance network are among the strongest for those datasets; see [1] for more analysis. Datasets with large weighted outdegree in the data-driven relevance network explain many other datasets; we checked whether their publications are highly cited. There is a statistically significant correlation between weighted outdegree and citation count (Fig. 1, top right; Spearman $\rho(169) = 0.2656$, $p < 0.001$). We examined whether influence of publication venue impact factor and h-index of the senior author could explain the low correlation; the answer was affirmative. Manual publication record check for datasets with low citation counts but high outdegrees (area A in Fig. 1, top right) revealed inconsistent publication records for two datasets (blue arrows in the figure point from original to corrected positions confirmed by Gene Expression Omnibus and ArrayExpress); the data-driven network revealed the inconsistency, and the new positions validate that the datasets are good explainers for others.

Conclusions

We tested feasibility of letting the data speak for themselves when relating new research to earlier studies, with positive conclusion: our scalable mixture modeling found both expected relationships such as tissue types, and relationships hard to find by keyword search such as treatments resembling conditions in other cell types. Such retrieval lessens the need for manual search. Data-driven relationships corresponded to citations when available but were richer and spotted errors in citations.

Acknowledgments. We thank M. Nelimarkka and T. Ruotsalo. Certain data herein are from Science Citation Index Expanded, Social Science Citation Index and Arts & Humanities Citation Index, by Thomson Reuters, ® Philadelphia, Pennsylvania, USA, © Copyright Thomson Reuters, ® 2011.

References

[1] Faisal A, Peltonen J, Georgii E, Rung J, Kaski S (2014) Toward computational cumulative biology by combining models of biological datasets. http://arxiv.org/pdf/1404.0329v1.pdf

[2] Schmid PR, Palmer NP, Kohane IS, Berger B (2012) Making sense out of massive data by going beyond differential expression. Proc Natl Acad Sci USA 109: 5594-5599.

[3] Tseng GC, Ghosh D, Feingold E (2012) Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res 40: 3785-3799.

[4] Rung J, Brazma A (2012) Reuse of public genome-wide gene expression data. Nat Rev Genet 14: 89-99.

[5] Punta M et al. (2012) The Pfam protein families database. Nucleic Acids Research 40: D290-D301.

[6] Lukk M, Kapushesky M, Nikkila J, Parkinson H, Goncalves A, Huber W, Ukkonen E, Brazma A. (2010) A global map of human gene expression. Nat Biotechnol 28: 322-324.

[7] Russ J, Futschik ME (2010) Comparison and consolidation of microarray data sets of human tissue expression. BMC Genomics 11: 305.

[8] Huttenhower C, Troyanskaya OG (2008) Assessing the functional structure of genomic data. Bioinformatics 24: i330-338.