

# Retrieval of Experiments by Efficient Comparison of Marginal Likelihoods

Sohan Seth<sup>1</sup>, John Shawe-Taylor<sup>2</sup>, Samuel Kaski<sup>1,3</sup>

1 Helsinki Institute for Information Technology HIIT,  
Department of Information and Computer Science, Aalto University, Espoo, Finland

2 Centre for Computational Statistics and Machine Learning,

Department of Computer Science, University College London, UK

3 Helsinki Institute for Information Technology HIIT,  
Department of Computer Science, University of Helsinki, Helsinki, Finland



HELSINKI  
INSTITUTE FOR  
INFORMATION  
TECHNOLOGY



Aalto University



UNIVERSITY OF HELSINKI



# Background

- ▶ **Information retrieval**: obtain information relevant to a user's need  
e.g, web pages, documents, images etc.
- ▶ **Objective**: information retrieval for biological datasets or experiments
  - by 'experiment' we mean a collection of measurements from a set of 'covariates' and the associated 'outcomes'  
i.e., in general any experiment performed, e.g., to validate a hypothesis

# Background

- ▶ **Information retrieval**: obtain information relevant to a user's need  
e.g, web pages, documents, images etc.
- ▶ **Objective**: information retrieval for biological datasets or experiments
  - by 'experiment' we mean a collection of measurements from a set of 'covariates' and the associated 'outcomes'  
i.e., in general any experiment performed, e.g., to validate a hypothesis

in particular,

in functional genomics: microarray measurements from patients and healthy persons

in toxicogenomics: post-treatment microarray measurements from cell lines, and the associated toxicity values

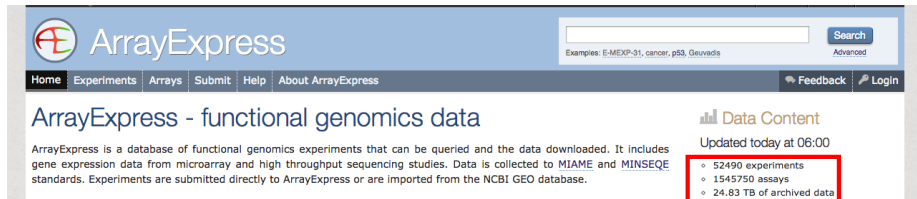
# Searching relevant datasets: current status

run experiment → publish findings → release **data** to databank

# Searching relevant datasets: current status

run experiment → publish findings → release **data** to databank

e.g, to ArrayExpress for functional genomics experiments



The screenshot shows the ArrayExpress website interface. At the top left is the ArrayExpress logo. To its right is a search bar with a 'Search' button and an 'Advanced' link. Below the search bar is a navigation menu with links for Home, Experiments, Arrays, Submit, Help, and About ArrayExpress. On the right side of the navigation menu are links for Feedback and Login. The main content area features the title 'ArrayExpress - functional genomics data' and a paragraph describing the database. To the right of this text is a 'Data Content' section with a bar chart icon, updated at 06:00. A red box highlights the following statistics:

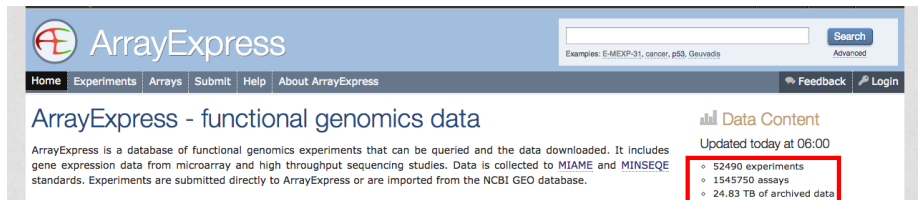
- 52490 experiments
- 1545750 assays
- 24.83 TB of archived data

<http://www.ebi.ac.uk/arrayexpress/>

# Searching relevant datasets: current status

run experiment → publish findings → release **data** to databank

e.g, to ArrayExpress for functional genomics experiments



The screenshot shows the ArrayExpress website interface. At the top left is the ArrayExpress logo. To its right is a search bar with a 'Search' button and an 'Advanced' link. Below the search bar is a navigation menu with links for Home, Experiments, Arrays, Submit, Help, and About ArrayExpress. On the right side of the navigation menu are links for Feedback and Login. The main content area features the title 'ArrayExpress - functional genomics data' and a paragraph describing the database. To the right of this text is a 'Data Content' section with a bar chart icon, updated timestamp, and a list of statistics: 52490 experiments, 1545750 assays, and 24.83 TB of archived data. The list of statistics is enclosed in a red rectangular box.

ArrayExpress - functional genomics data

ArrayExpress is a database of functional genomics experiments that can be queried and the data downloaded. It includes gene expression data from microarray and high throughput sequencing studies. Data is collected to [MIAME](#) and [MINSEQE](#) standards. Experiments are submitted directly to ArrayExpress or are imported from the NCBI GEO database.

**Data Content**  
Updated today at 06:00

- 52490 experiments
- 1545750 assays
- 24.83 TB of archived data

<http://www.ebi.ac.uk/arrayexpress/>

- **data** ≡ measurements over covariates and outcomes + associated metadata  
e.g., in functional genomics: disease, disease state, cell type

# Searching relevant datasets: current status

- ▶ search relevant scientific articles, use **citations**

# Searching relevant datasets: current status

- ▶ search relevant scientific articles, use **citations**
- ▶ search relevant metadata (**keywords**) in databanks, e.g., use experimental factor ontology

EMBL-EBI

Services Research Training About us

Experimental Factor Ontology

Search EFO... Search

Examples: cancer, HeLa, Li-Fraumeni syndrome

Home Browse EFO Submit Term Semantic Web Project

## Representing experimental variables with EFO

The **Experimental Factor Ontology (EFO)** provides a systematic description of many experimental variables available in EBI databases, and for external projects such as the NHGRI GWAS catalogue. It combines parts of several biological ontologies, such as anatomy, disease and chemical compounds. The scope of EFO is to support the annotation, analysis and visualization of data handled by the EBI Functional Genomics Team. We also add terms for external users when requested. If you are new to ontologies, there is a short introduction on the subject available and a blog post by James Malone on what ontologies are for.

<http://www.ebi.ac.uk/efo/>



# Searching relevant datasets: issues

- ▶ metadata usually vary with user, e.g., cancer and carcinoma,
- ▶ metadata can often be incomplete,

# Searching relevant datasets: issues

- ▶ metadata usually vary with user, e.g., cancer and carcinoma,
- ▶ metadata can often be incomplete,

**BIOINFORMATICS**

Vol. 23 ISMB/ECCB 2007, pages i41–i48  
doi:10.1093/bioinformatics/btm229

---

## **Manual curation is not sufficient for annotation of genomic databases**

William A. Baumgartner, Jr.<sup>1,\*†</sup>, K. Bretonnel Cohen<sup>1,†</sup>, Lynne M. Fox<sup>2</sup>,  
George Acquah-Mensah<sup>3</sup> and Lawrence Hunter<sup>1,\*</sup>

<sup>1</sup>Center for Computational Pharmacology, University of Colorado School of Medicine, <sup>2</sup>Denison Library, University of Colorado Health Science Center and <sup>3</sup>Department of Pharmaceutical Sciences, Massachusetts College of Pharmacy and Health Sciences, USA

---

# Searching relevant datasets: next step

- ▶ search by comparing the measurements not metadata (annotations), e.g., for microarray datasets search with samples  $\times$  probes matrix

# Searching relevant datasets: next step

- ▶ search by comparing the measurements not metadata (annotations), e.g., for microarray datasets search with samples  $\times$  probes matrix

**BIOINFORMATICS**

Vol. 25 ISMB 2009, pages i145–i153  
doi:10.1093/bioinformatics/btp215

---

## **Probabilistic retrieval and visualization of biologically relevant microarray experiments**

José Caldas<sup>1,\*</sup>, Nils Gehlenborg<sup>2,3</sup>, Ali Faisal<sup>1</sup>, Alvis Brazma<sup>2</sup> and Samuel Kaski<sup>1</sup>

<sup>1</sup>Helsinki Institute for Information Technology, Department of Information and Computer Science, Helsinki University of Technology, Finland, <sup>2</sup>Microarray Team, European Bioinformatics Institute and <sup>3</sup>Graduate School of Life Sciences, University of Cambridge, Cambridge, UK

---

“find unexpected things in addition to the already known things available for metadata searches”

- ▶ the basic intuition is to compare characteristics of the measurements e.g., are the same genes being enriched?  
e.g., are the same genes being associated?

# Searching relevant datasets: next step

- ▶ utilize researcher's expertise in retrieval in terms of modeling  
by model we mean generative model, or posterior distribution over parameters

$$\text{posterior} \propto \text{likelihood (measurements)} \times \text{prior (expertise)}$$

# Searching relevant datasets: next step

- ▶ utilize researcher's expertise in retrieval in terms of modeling  
by model we mean generative model, or posterior distribution over parameters

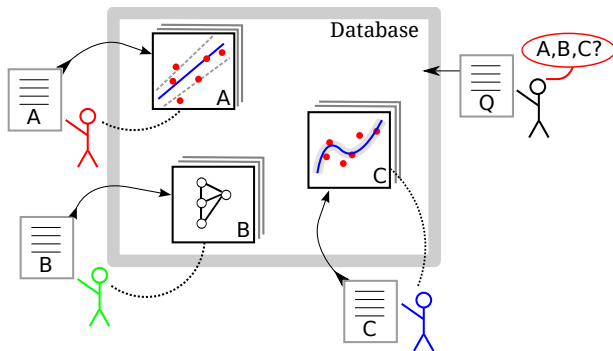
$$\text{posterior} \propto \text{likelihood (measurements)} \times \text{prior (expertise)}$$

- ▶ given model we can use marginal likelihood as a measure of similarity

$$\text{probability}(\text{query dataset} \mid \text{model of earlier dataset})$$

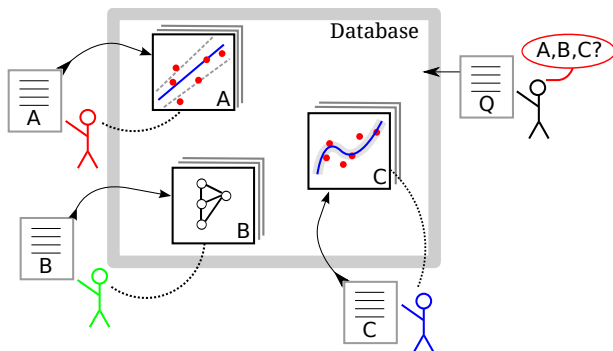
# Searching relevant datasets: summary

release **measurements + metadata + model** to databank



# Searching relevant datasets: summary

release **measurements + metadata + model** to databank



- ▶ however, hypothetical situation: we do not have models from researchers  
pilot studies are based on fitting our own model on datasets



# Current work: background

- experiment  $\equiv$  collection of measurements over covariates and outcomes, i.e.,  $\mathcal{E}_d = \{(\mathbf{c}_{di}, \mathbf{o}_{di})\}_{i=1}^{n_d}$ .
- each experiment  $\mathcal{E}_d$  has been modeled as  $\mathcal{M}_d$ ,
- model  $\equiv$  a collection of posterior MCMC samples, i.e.,  $\mathcal{M}_d = \{\theta_{dk}\}_{k=1}^{m_d}$

# Current work: background

- experiment  $\equiv$  collection of measurements over covariates and outcomes, i.e.,  $\mathcal{E}_d = \{(c_{di}, o_{di})\}_{i=1}^{n_d}$ .
- each experiment  $\mathcal{E}_d$  has been modeled as  $\mathcal{M}_d$ ,
- model  $\equiv$  a collection of posterior MCMC samples, i.e.,  $\mathcal{M}_d = \{\theta_{dk}\}_{k=1}^{m_d}$

model can be used for retrieval in different ways

- 1 explain query dataset  $\mathcal{E}_q$  as combination of previous datasets (Faisal et al.)
- 2 given query model  $\mathcal{M}_q$ , observe overlap with previous models (Dutta et al.)

# Current work: background

- experiment  $\equiv$  collection of measurements over covariates and outcomes, i.e.,  $\mathcal{E}_d = \{(c_{di}, o_{di})\}_{i=1}^{n_d}$ .
- each experiment  $\mathcal{E}_d$  has been modeled as  $\mathcal{M}_d$ ,
- model  $\equiv$  a collection of posterior MCMC samples, i.e.,  $\mathcal{M}_d = \{\theta_{dk}\}_{k=1}^{m_d}$

model can be used for retrieval in different ways

- 1 explain query dataset  $\mathcal{E}_q$  as combination of previous datasets (Faisal et al.)
- 2 given query model  $\mathcal{M}_q$ , observe overlap with previous models (Dutta et al.)
- 3 rank existing models  $\{\mathcal{M}_i : i = 1, \dots, d, \dots, D\}$  in terms of marginal likelihood

$$\text{ML}_{q|d} = \mathbb{E}_{p(\cdot|\mathcal{E}_d)} p(\mathcal{E}_q|\cdot)$$

# Current work: faster retrieval

given posterior samples

$$\text{(unweighted average) } \widehat{\text{ML}}_{q|d} = \frac{1}{m_d} \sum_{k=1}^{m_d} p(\mathcal{E}_q | \theta_{dk})$$

# Current work: faster retrieval

given posterior samples

$$\text{(unweighted average) } \widehat{\text{ML}}_{q|d} = \frac{1}{m_d} \sum_{k=1}^{m_d} p(\mathcal{E}_q | \theta_{dk})$$

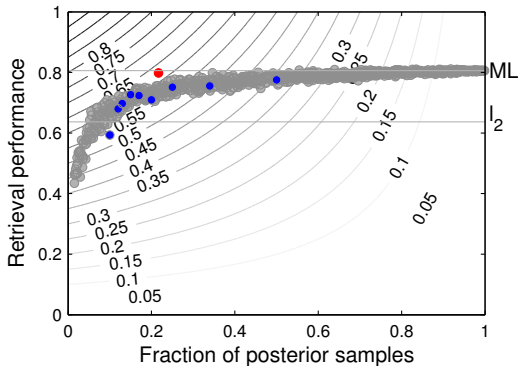
- ▶ however, evaluating  $m_d$  marginal likelihood can be expensive

# Current work: faster retrieval

given posterior samples

$$\text{(unweighted average)} \quad \widehat{\text{ML}}_{q|d} = \frac{1}{m_d} \sum_{k=1}^{m_d} p(\mathcal{E}_q | \theta_{dk})$$

- ▶ however, evaluating  $m_d$  marginal likelihood can be expensive



# Current work: approach

- ▶ find fewer important samples that can be used for retrieval

$$\text{(weighted average) } \widetilde{\text{ML}}_{q|d} \approx \sum_{k=1}^{m_d} w_{dk} p(\mathcal{E}_q | \theta_{dk})$$

where  $w_d = [w_{d1}, \dots, w_{dm_d}]$  is a vector of **sparse** weights (ideally) non-negative and sum to one.

- ▶ learn the weights by preserving ranking with respect to  $\widehat{\text{ML}}_{q|d}$

# Current work: optimization

- training set  $\{\mathcal{E}_d\}_{d=1}^D$
- consider a triplet  $(i_1, i_2, i_3) \in \{1, \dots, D\}^3$ ,  
use  $i_3 \equiv q$  as query and rank the  $(i_1, i_2)$
- without loss of generality, assume  $\widehat{\text{ML}}_{q|i_1} > \widehat{\text{ML}}_{q|i_2}$ , (unweighted average)  
then ensure  $\widetilde{\text{ML}}_{q|i_1} > \widetilde{\text{ML}}_{q|i_2}$ , (weighted average) i.e.,

$$\sum_k w_{i_1 k} p(\mathcal{E}_q | \theta_{i_1 k}) > \sum_k w_{i_2 k} p(\mathcal{E}_q | \theta_{i_2 k})$$



# Current work: optimization

- training set  $\{\mathcal{E}_d\}_{d=1}^D$
- consider a triplet  $(i_1, i_2, i_3) \in \{1, \dots, D\}^3$ ,  
use  $i_3 \equiv q$  as query and rank the  $(i_1, i_2)$
- without loss of generality, assume  $\widehat{\text{ML}}_{q|i_1} > \widehat{\text{ML}}_{q|i_2}$ , (unweighted average)  
then ensure  $\widetilde{\text{ML}}_{q|i_1} > \widetilde{\text{ML}}_{q|i_2}$ , (weighted average) i.e.,

$$\sum_k w_{i_1 k} p(\mathcal{E}_q | \theta_{i_1 k}) > \sum_k w_{i_2 k} p(\mathcal{E}_q | \theta_{i_2 k})$$

or,

$$\begin{aligned} & [+p(\mathcal{E}_q | \theta_{i_1 1}), \dots, +p(\mathcal{E}_q | \theta_{i_1 m_{i_1}}), -p(\mathcal{E}_q | \theta_{i_2 1}), \dots, -p(\mathcal{E}_q | \theta_{i_2 m_{i_2}})] \\ & [w_{i_1 1}, \dots, w_{i_1 m_{i_1}}, w_{i_2 1}, \dots, w_{i_2 m_{i_2}}]^\top > 0 \end{aligned}$$

# Current work: optimization

- training set  $\{\mathcal{E}_d\}_{d=1}^D$
- consider a triplet  $(i_1, i_2, i_3) \in \{1, \dots, D\}^3$ ,  
use  $i_3 \equiv q$  as query and rank the  $(i_1, i_2)$
- without loss of generality, assume  $\widehat{\text{ML}}_{q|i_1} > \widehat{\text{ML}}_{q|i_2}$ , (unweighted average)  
then ensure  $\widetilde{\text{ML}}_{q|i_1} > \widetilde{\text{ML}}_{q|i_2}$ , (weighted average) i.e.,

$$\sum_k w_{i_1 k} p(\mathcal{E}_q | \theta_{i_1 k}) > \sum_k w_{i_2 k} p(\mathcal{E}_q | \theta_{i_2 k})$$

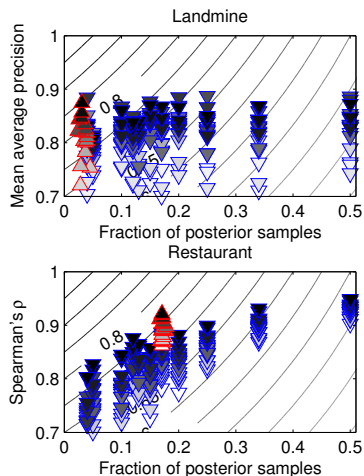
or,

$$\begin{aligned} & [ +p(\mathcal{E}_q | \theta_{i_1 1}), \dots, +p(\mathcal{E}_q | \theta_{i_1 m_{i_1}}), -p(\mathcal{E}_q | \theta_{i_2 1}), \dots, -p(\mathcal{E}_q | \theta_{i_2 m_{i_2}}) ] \\ & [ w_{i_1 1}, \dots, w_{i_1 m_{i_1}}, w_{i_2 1}, \dots, w_{i_2 m_{i_2}} ]^\top > 0 \end{aligned}$$

- each binary label corresponds to a triplet
- linear classification problem with sparse design matrix
- learn  $w = [w_1, \dots, w_d]$ , weight vector for each experiment

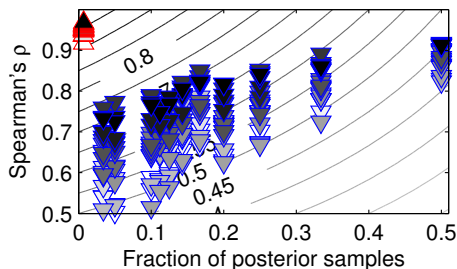
# Preliminary results

- ▶ Landmine
  - 29 experiments: two classes 16-13
  - each experiment is a classification problem
  - 9 features,  $\sim 500$  samples
- ▶ Restaurant
  - 119 experiments
  - each experiment is a regression problem
  - 22 binary features, 3-18 samples



# Preliminary results: toxicogenomic data

- covariates: post treatment gene expression, outcome: toxicity
- 65 drugs (experiments), 26-44 cell lines (samples)
- 1000 genes (LINCS, Library of Integrated Network-based Cellular Signatures),
- associated toxicity (CTD<sup>2</sup>, Cancer Target Discovery and Development)



Ack: Suleiman Ali Khan, HIIT; Aravind Subramanian, Broad Institute

# Summary

- ▶ general
  - retrieval of biological datasets or experiments
  - metadata driven search → content driven search
  - suggest releasing models, model captures expertise
- ▶ specific
  - reducing likelihood evaluation to speed up retrieval
  - preliminary results are promising
- ▶ ongoing
  - larger validation set: toxicogenomic datasets, ArrayExpress affymetrix dataset

# Acknowledgements

- ▶ HIIT personnel

Ali Faisal, Elisabeth Georgii, Jaakko Peltonen, Ritabrata Dutta

- ▶ EBI collaborators

Alvis Brazma, Ugis Sarkans

# Acknowledgements

- ▶ HIIT personnel

Ali Faisal, Elisabeth Georgii, Jaakko Peltonen, Ritabrata Dutta

- ▶ EBI collaborators

Alvis Brazma, Ugis Sarkans

Thank You! Questions?