Examining the uses of shared data

Heather A. Piwowar and Douglas B. Fridsma

Does your area of research use shared datasets?

- ***Re-using data has many benefits, including research synergy and efficient resource use***
- *Some research areas have tools, communities, and practices which facilitate re-use*
- ***Identifying these areas** will allow us to learn from them, and apply the lessons to areas which underutilize the **sharing and re-purposing of scientific data between investigators***

Nature Precedings : doi:10.1038/npre.2007.425.1 : Posted 11 Jul 2007

Which datasets?

This study examines the data re-use of microarray gene expression datasets.
Thousands of microarray gene expression datasets have been deposited in publicly available databases. Many studies reuse this data, but it is not well understood which datasets are reused and for what purpose. Here, we examined all publications in PubMed Central on April 1, 2007 containing the word "microarray."

How did we identify re-use?

We trained a machine-learning algorithm to automatically classify full-text gene expression microarray studies into two classes: those that generated original microarray data (n=900) and those which only re-used data (n=250).
SVlite, NLTK, feature selection

How did we identify patterns of re-use?

We then compared the Medical Subject Heading (MeSH) terms of two classes to identify MeSH topics which were over- or under-represented by publications with re-used data.

Results

Studies on humans, mice, chordata, and invertebrates were roughly equally likely to be conducted using original or shared microarray data, whereas shared data was used in a relatively high proportion of studies involving fungi (odds ratio (OR)=2.4), and a relatively low proportion involving rats, bacteria, viruses, plants, or genetically-altered or inbred animals (OR<0.5). Unsurprisingly, when we looked at Major MeSH terms to represent the primary purpose of the studies, statistical and computational methods clearly dominated. The only biomedical topics with a relatively high proportion of data reuse Major MeSH terms were Promoter Regions, Evolution, and Protein Interaction Mapping.

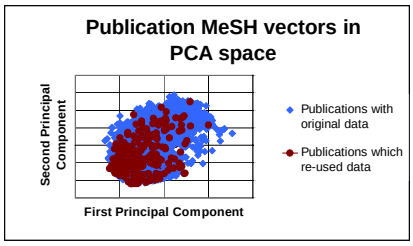

FIGURE 1: *Documents with re-used data have a different MeSH distribution than those with original data.*

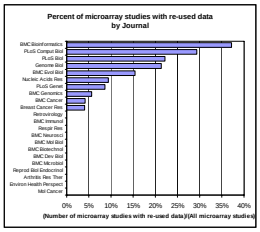

FIGURE 2: *As expected, journals with a bioinformatics focus published the highest proportion of studies with re-used microarray data.*

Hope

Identifying areas of particularly successful microarray data re-use -- such as Saccharomyces cerevisiae datasets and studies of promoter regions and evolution -- can highlight best practices to be used when developing research agendas, tools, standards, repositories, and communities in areas which have yet to receive major benefits from shared data.

Future Work

We plan to refine our prototype NLP tool for identifying studies which re-use data, and continue studying and measuring re-use and reusability.

Acknowledgements

We sincerely thank our funders:

- USA NLM for a training grant
- USA NSF for a travel grant

For Further Information

Please contact hpiwowar@alumni.pitt.edu

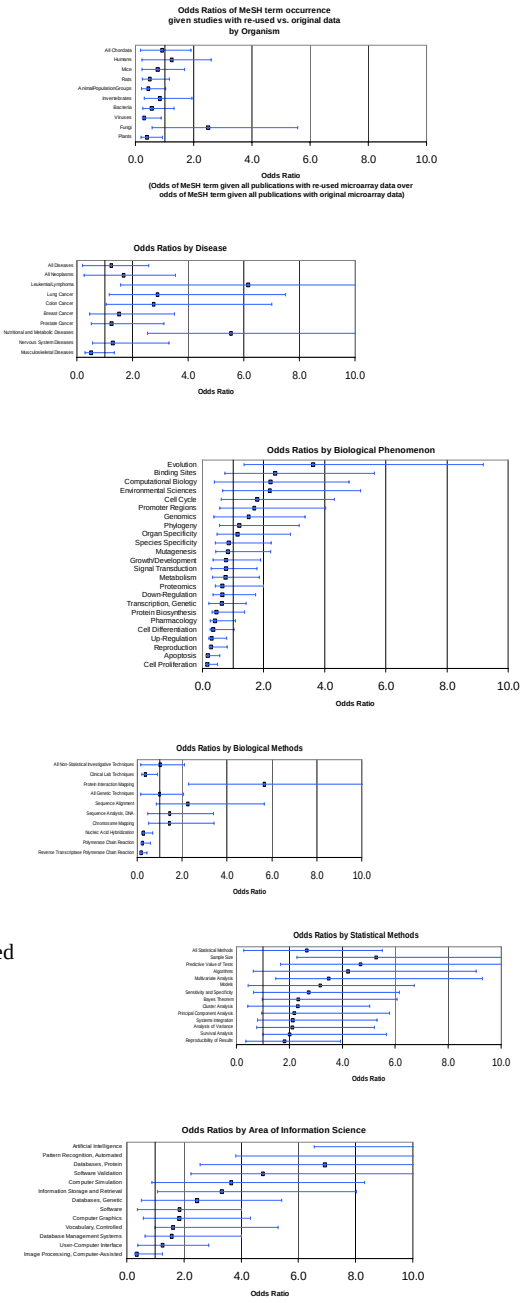

FIGURE 3: *The change in odds that a specific MeSH term will describe a publication with re-used data as compared to a publication with original microarray data is illustrated in the Odds Ratio graphs above.*

University of Pittsburgh
Department of Biomedical Informatics