Prevalence and Patterns of Microarray Data Sharing

Heather A. Piwowar and Wendy W. Chapman
Department of Biomedical Informatics, University of Pittsburgh School of Medicine

Poster presented at PSB 2007 (Pacific Symposium on Biocomputing)
This abstract and the poster image have been archived at Nature Precedings, March 2008.
Paper in development.
Data from this study has been shared on our website.
For more information on our data sharing research please email or visit!

Sharing research data is a cornerstone of science. Although many tools and policies exist to encourage data sharing, the prevalence with which datasets are shared is not well understood. We report our preliminary results on patterns of sharing microarray data in public databases.

The most comprehensive method for measuring occurrences of public data sharing is manual curation of research reports, since data sharing plans are usually communicated in free text within the body of an article. Our early findings from manual curation of 100 papers suggest that 30% of investigators publicly share their full microarray datasets. Of these, 70% of the datasets are deposited at NCBI's Gene Expression Omnibus (GEO) database, 20% at EBI's ArrayExpress, and 10% in smaller databases or lab or publisher websites.

Next, we supplemented this manual process with a rough automated estimate of data sharing prevalence. Using PubMed, we identified research articles with MeSH terms for both "Gene Expression Profiling" and "Oligonucleotide Array Sequence Analysis" and published in 2006. We then searched GEO and ArrayExpress for links to these PubMed IDs to determine which of the articles had been credited as an originating data source.

Of the 2503 articles, 440 (18%) articles had links from either GEO or ArrayExpress. Of these 440 articles, 70% had links from GEO and 30% from ArrayExpress, with an overlapping 12% from both GEO and ArrayExpress.

Interestingly, studies with free full text at PubMed were twice (Odds Ratio=2.1; 95% confidence interval: [1.7 to 2.5]) as likely to be linked as a data source within GEO or ArrayExpress than those without free full text. Studies with human data were less likely to have a link (OR=0.8 [0.6 to 0.9]) than studies with only non-human data. The proportion of articles with a link within these two databases has increased over time: the odds of a data-source link for studies was 2.5 [2.0 to 3.1] times greater for studies published in 2006 than 2002.

As might be expected, studies with the fewest funding sources had the fewest data-sharing links: only 28 (6%) of the 433 studies with no funding source were listed within GEO or ArrayExpress. In contrast, studies funded by the NIH, the US government, or a non-US government source had data-sharing links in 282 of 1556 cases (18%), while studies funded by two or more of these mechanisms were listed in the databases in 130 out of 514 cases (25%).

In summary, our initial manual approach for identifying studies which shared their data was comprehensive but time-consuming; natural language processing techniques could be helpful. Our subsequent automated approach yielded conservative estimates for total data sharing prevalence, nonetheless revealing several promising hypotheses for data sharing behavior

We hope these preliminary results will inspire additional investigations into data sharing behavior, and in turn the development of effective policies and tools to facilitate this important aspect of scientific research.