Published in final edited form as:

Nat Biotechnol. 2006 December ; 24(12): 1471–1472.

NCBI GEO standards and services for microarray data

Ron Edgar and Tanya Barrett

National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), 45 Center Drive, Bethesda, MD 20892-6511, Email: geo@ncbi.nlm.nih.gov

The Minimum Information About a Microarray Experiment (MIAME) guidelines are a data content document developed by the Microarray Gene Expression Data (MGED) Society that outlines the information that should be provided when describing a microarray experiment¹. Many journals and funding agencies have adopted the guidelines, with the aim of facilitating access to the elements of a study that would enable independent evaluation of results. However, the MIAME requirements have been criticized recently^{2, 3}. The criticism stems, in part, from different interpretations of the level of detail required to adequately report a microarray experiment, and debates as to whether there is a genuine benefit to making microarray data public.

The Gene Expression Omnibus (GEO) database at the National Center for Biotechnology Information (NCBI)⁴ and ArrayExpress at the European Bioinformatics Institute (EBI)⁵ are the two major public databases of microarray data. Although they have different designs, both databases support capture of all data elements defined by MIAME. Figure 1 presents a timeline of major landmarks in the evolution of the GEO database, together with concomitant growth in submissions. GEO was launched in 2000, more than a year before the MIAME guidelines were proposed. Because there was not yet a consensus on reporting standards for microarray data, or even an obligation to make microarray data public, GEO initially allowed a minimal level of experimental detail to be supplied. Over the ensuing years we continually monitored the needs and requests of end-users, and gauged the level of effort submitters were realistically willing to invest in making their data public. We responded with incremental improvements to database design and curation standards, and we developed easy-to-generate batch deposit formats that significantly reduced the burden of submission and allowed contributors to focus on the content submitted rather than the mechanism of submission.

In June 2005, we released major database revisions that included specific provisions for all MIAME data elements. In 2006, mechanisms for provision of raw data were further streamlined, and several MIAME elements that were previously optional became mandatory. Yet, even with these advances, it is still possible for a submitter to supply data that do not strictly adhere to the MIAME requirements. The difficulty lies in the fact that MIAME is a subjective set of guidelines where the level of detail to report is open to interpretation and, thus, cannot be unequivocally validated or enforced by computational means.

All data submitted to GEO are syntactically validated for correct document structure, organization, and provision of basic elements. Next, each submission is inspected by curators for content integrity. GEO curators employ a pragmatic approach; we aim to ensure that sufficient information has been supplied to allow general interpretation of the experiment. Although encouraged, we have been less dogmatic with regards to provision of all-inclusive experimental protocols that would possibly permit practical replication of the entire experiment. Our reasoning is that provision of granulated experimental details adds a significant burden on the submitter, for (arguably) minimal real benefit for most end-users who are usually less concerned with this level of detail. When content or format problems are identified, curators work with the submitter until the issue is resolved. Submissions lacking

critical descriptive elements necessary for overall experiment interpretation are not approved for public release. However, given the large diversity of biological themes, technologies, and statistical transformations applied to microarray data, it is impractical for curators to decisively determine the accuracy and validity of the data, or to assess if all relevant information has been supplied. This is where the role of reviewers and editors becomes important.

The GEO database has had mechanisms for anonymous reviewer access to prepublication data since 2003. Over the last several years, authors have occasionally requested curator comment regarding the level of MIAME-compliance of their submissions, and we have been happy to offer feedback on areas that could be improved. GEO staff are similarly available to support reviewers and editors by providing tailored inspections of MIAME compliance of specific submissions upon request of the journal, as ArrayExpress is proposing to do⁶. If a reviewer determines that insufficient information has been supplied, the GEO database is designed such that authors can quickly respond by updating their records accordingly.

It has been challenging to find the optimal balance between submitter effort and the appropriate level of metadata detail to request, all within a rapidly evolving technological and social environment⁷. However, the relative simplicity of the GEO database structure, together with common-sense curation policies that focus on gathering germane MIAME elements, have made it possible for us to develop an extensive suite of utilities that make the volumes of complex data archived at GEO accessible and easy to use by the research community at large⁸. Ultimately, the value of a database is reflected by how it is used by the community it serves. In the past month, GEO received approximately one million query hits, and over 200,000 file transfer downloads amounting to over 2.5 terabytes of compressed data. Furthermore, it is clear that researchers are applying these data to their own studies, as evidenced by over 100 recent publications citing data found in GEO to support or otherwise complement their own studies⁹. We view this as testament that the effort involved in making expression data public via GEO is fully justified.

References

1. Brazma A, et al. *Nat Genet* 2001;29:365–371. [PubMed: 11726920]
2. Galbraith DW. *Nature* 2006;444:31. [PubMed: 17080064]
3. Shields R. *Trends Genet* 2006;22:65–66. [PubMed: 16380192]
4. <http://www.ncbi.nlm.nih.gov/geo/>.
5. <http://www.ebi.ac.uk/arrayexpress/>.
6. Brazma A. *Nat Biotech*. 2006in press
7. Edgar R. *Nature* 2006;443:394. [PubMed: 17006493]
8. Barrett T, et al. *Nucleic Acids Res* 2005;33:D562–566. [PubMed: 15608262]
9. <http://www.ncbi.nlm.nih.gov/geo/info/ucitations.html>.

Figure 1. Timeline of GEO growth and major landmarks in evolution of GEO database, and a screenshot of GEO tools which allow users to query, analyze, and visualize the data in GEO.