GuestEditorial

RUSS ALTMAN, MD, PhD

Share and Share Alike:

A Proposed Set of Guidelines for Both Data and Software

A pair of challenges increasingly threaten the success of bioinformatics research: convincing biologists to share their data and convincing computational colleagues to share their code. Many of us learned to share in preschool, but we also learned that there are sometimes reasons to keep a strong grip on your own stuff. The barriers to academic sharing include protection of graduate student and post-doc publication priority, protection of intellectual property for institutional patents, protection of the patient confidentiality and privacy, and protection of rights for future publication and funding applications.

But these concerns can be overcome with appropriate guidelines. And I believe the parallel issues of data sharing and code sharing can be addressed together. Proven successes in both fields—the sharing of genomics data, and the open source movement—suggest the effort will be well worthwhile.

The ethic of data sharing permeates the genomics community, a fact that derives from the earliest examples of open biological databases: Genbank for storing DNA sequences and the Protein Data Bank for storing macromolecular three-dimensional structures. Through the efforts of visionary scientists, funding agency staff, and journal editors, the submission of data to databases simultaneous with publication became a standard in these fields. The genome sequencing project was successful in part because the organizers created rules very early on—the "Bermuda rules" required nightly release of sequence data. This led directly to the creation of databases to support the sharing of microarray gene expression data, such as the Stanford Microarray Database (SMD), and the Gene Expression Omnibus (GEO) at NCBI. But biomedical computation researchers who enter new fields with visions of Genbank, PDB, and GEO as the relevant precedents may be surprised at the degree of resistance to data sharing in other subdisciplines.

At the same time that biomedical computation researchers are struggling to convince their biology colleagues to share data, they are engaged in an intriguing debate about the merits of open-source code sharing. The open source movement points to the emergence of Linux as a major precedent that shows the power of shared code. Closer to biomedicine, myriad examples of public domain software have energized certain fields, including EMBOSS for molecular biology, VTK for general visualization, and others. These tend to be larger projects with explicit dissemination goals. Of course, the NIH program that funds the seven National Centers for Biomedical Computation (NCBC) has software dissemination as a major goal, and has led to the creation of domain-specific portals such as Simtk.org. It is more difficult, however, to procure software created by an individual lab that is competing with other labs to create novel methods for biomedical computation.

These parallel problems merit a common solution. I would suggest that the community is converging on the included guidelines.

PROPOSED GUIDELINES

1 Biological data sets and software for storing, analyzing, and visualizing biological data should be released to the public when the first-pass analysis and publication is substantially complete, and no later than one year after the appearance of the first full scale analysis.

2 Users should have no expectation of "support" for working with the data or software, beyond basic documentation sufficient for a motivated graduate student to understand and use it.

3 Citation of the original source paper, consistent with scholarly standards, should be mandated, and failure to cite should be considered scientific misconduct.

4 Downloads should be instrumented, and information about frequency of downloads and other measures of impact should be included in hiring and promotion materials. They should be routinely addressed and evaluated in letters of recommendation written by peers.

5 Funding agency staff and biomedical journal editors must be firm in enforcing the sharing of data and code. Manuscripts should have an identifier that lists the eventual location of the data or code, and a date when the data or code will be available.

The current climate of tight funding for biomedical research, with the end of the NIH budget doubling, could threaten the trend towards more open sharing as investigators become nervous about competitive advantage. However, it is critical to preserve the gains in this area achieved over the last decade, and to institutionalize the processes that guarantee continued sharing. □