Future of biology rests in harnessing data avalanche

September 4 2008

(PhysOrg.com) -- Like most sciences, biology is inundated with data. However, a group of researchers warns in a *Nature* feature that the avalanche of biological information is at the point where the discipline may be unable to reach its full potential without improvements for curating data into on-line databases. The commentary appears in the September 4, issue of the journal and outlines specific remedies to harness the information overload.

By July 2008, data-extractors or curators had indexed over 18 million articles in PubMed and sequences of over 260,000 organisms into GenBank. Both are examples of databases where biological information is stored for public access. Data curation is very labor intensive.

“There is a lack of standardization or consistency in the way scientists

report their findings in different journals," remarked corresponding author Sue Rhee of the Carnegie Institution's Department of Plant Biology and principal investigator of The Arabidopsis Information Resource (TAIR). "In some cases the researchers don't even specify the species of a gene under study. That leaves biocurators, who have advance degrees in biology, and expertise with databases and scripting languages, to read the full text and transfer the essence of the information into specific fields in the database. They spend a lot of time just figuring out the basics. And that leaves a lot of room for error."

Curation is not just a data organization tool. Such input has become essential to biological research. The authors note that eleven different databases had ¾ of a million visitors who viewed 20 million pages in just one month. And with inference programs that feed on the curated data, researchers can now tap into other work that relates to theirs and use that data in their own experiments—a huge advancement that is accelerating the pace of biology. "With this vast universe of information, the whole nature of experimentation is changing," continued Rhee. "But the field is being held back with the curation backlog."

The group of authors outlined a series of solutions to the problem. The first is to have authors input their data directly into databases upon acceptance in refereed journals. This step has already begun with Plant Physiology and TAIR. When a manuscript in accepted, researchers now fill in a web form about Arabidopsis genes. Second, the commentators urge the biological community to adopt standard reporting formats that are universally agreed upon. And third, curation needs to be elevated by academic institutions and funding agencies. There should also be incentives for researchers to curate their own data, such as increases in academic recognition, career advancement, and funding. They additionally suggest that "community annotation" could be modeled after large-scale astronomy projects like the Sloan Digital Sky Survey, or the Galaxy Zoo, where 80,000 astronomers and interested amateurs

classified one million galaxies in less than three weeks.

“The effort and cost required to curate the data is small compared with the cost of carrying out the research in the first place, yet this additional step adds tremendously to the value of the research results to society,” commented Eva Huala, director of TAIR.

Wolf Frommer, acting director of Carnegie’s Department of Plant Biology noted that “advances in our understanding of biology will affect our food supply, our health-care system, the development of remedies for climate change, and many other aspects of daily life. Basic and applied research have to go hand in hand with curation of databases so that humanity can adapt to the quickly changing world as fast as possible.”

Provided by Carnegie Institution