Putting data to work

Optimal use of the extensive data sets being generated in the post-genomic era will require clear standards and a commitment to open communication.

Chemists and biologists have traditionally advanced science by defining a particular system of interest, formulating models for how the system works and testing these models at increasing levels of molecular resolution. The arrival of new technologies has permitted the more rapid interrogation of complex biological systems *in situ* and in real time, which has led to an explosion of potentially useful data, much of which is being housed in community-supported public databases. Chemical biology connects research from diverse parts of chemistry and biology, and therefore chemical biologists are key contributors to and users of a wide spectrum of these 'omic' data sets (*Nat. Chem. Biol.* **6**, 631, 2010). Developing more robust computational tools and implementing experimental and informatic standards will permit scientists to make the most of this exciting, data-rich era.

As discussed in a Commentary by Palsson and Zengler in the current issue (p. 787), the accumulation of omic data far outpaces the rate at which scientists can mine this information, generate new hypotheses and experimentally test new scientific models. In this piece, the authors examine the important question of how to balance acquisition of these large data sets with the significant investment of time and effort that is needed to translate raw omic data into new biological understanding.

As an example, a chemical biologist might want to visualize data sets from transcriptomic, phosphoproteomic and metabolomic profiling experiments—potentially across a wide range of conditions and cell types—to understand the mode of action and specificity of a kinase inhibitor. Navigating multiple databases, analyzing the data and ensuring that comparisons across data sets are scientifically sound currently requires substantial care and effort. Thus, as chemical biologists develop new tools for exploring increasingly complex systems, it is essential that the resulting data sets are able to 'talk' to each other and can be presented in a user-friendly manner.

These tools are required in part because of the diverse data types and formats across a growing number of public and private databases. Databases have emerged to serve the needs of a particular community. As such, data standards and formats agreed upon during the early stages of database development become adopted throughout a subfield as scientists deposit data and use the databases. Databases differ widely in areas such as data curation, the choice of identifiers for database entries and their user interfaces. Early databases such as GenBank and the Protein Data Bank (PDB) have been widely used for several decades, while others such as PubChem for chemical compounds and assay data or PRoteomics IDEntifications (PRIDE) for proteomics data are still in relatively early stages of development. All of these factors highlight the substantial challenge in integrating established as well as new data types into a unified and useful format for chemical biology researchers.

What is needed to enable better data integration and analysis? As highlighted by Palsson and Zengler, important cultural shifts will be necessary to make the best use of data from omic science. Given that multiple databases may exist for a particular type of data, agencies and researchers must adopt standards (for example, nomenclatures, identifiers and reporting formats) that enable users to seamlessly use relevant data from across these related repositories. This also requires diligence on the part of researchers to ensure that experiments are designed to produce output data sets that will be complete, self-contained and in accordance with community standards. Scientists working in these areas should organize 'jamborees'—community-centered data analysis efforts designed to transform databases into useful knowledge bases—that will support scientific advancement. Scientists working with model organisms (for example, *Caenorhabditis elegans*: http://www.wormbase.org) or on particular scientific problems (for example, mitosis: http://www.mitocheck.org) have already begun to develop integrated resources and data sites to advance science in their areas. For historical reasons, genomic and transcriptomic data has taken priority. However, to move all fields forward, a broader cross-section of data sets—proteomic data, metabolomic data and data on bioactive small molecules—needs to be plugged in to these broader initiatives.

Scientific publishing serves an important role in enabling access to data. *Nature Chemical Biology* and the other Nature journals require that authors make data sets freely available (http://www.nature.com/authors/editorial_policies/availability.html). For data types with established databases (gene sequences, protein structural data and so on) deposition of the data is mandatory. For data sets where database development is at an earlier stage, we encourage authors to deposit these data in appropriate publicly accessible databases. Currently, the publication processes for *Nature Chemical Biology* and *Nature Chemistry* include automated deposition of key chemical compounds into PubChem. In all cases, data deposition and publication of accession codes in the published paper enables readers to have access to essential data in perpetuity.

Journals can also facilitate communication about scientific data by adopting community standards for presenting information in published papers. For *Nature Chemical Biology* papers, this process occurs during copy editing, in which our copy editor works with the authors to optimize the scientific clarity and readability of the final paper. As part of this, we edit the paper in 'journal style', which ensures that scientific terms, such as the names or abbreviations of genes or chemical compounds, are presented consistently. Whenever possible, these conventions reflect community standards or are based on guidelines from organizations such as IUPAC or IUBMB. For example, we now request that authors of *Nature Chemical Biology* papers adopt the recommendations of the Strenda (Standards for the Reporting of Enzymological Data) Committee (see Correspondence, p. 785) as they conduct kinetic and binding experiments on enzymes and report the resulting data.

Science has reached a stage in which the acquisition of new data is not the limiting factor in making new discoveries. Instead the challenge is making the vast amounts of data available work for scientists. Chemical biologists need to be actively engaged in conversations about data standards and reporting and to commit themselves to broader integration of data in the support of scientific advancement. ■

© 2010 Nature America, Inc. All rights reserved.