The Scientist

Volume 23 | Issue 4 | Page 33

By Steven Wiley

Why Don't We Share Data?

There are so, so many reasons—and they make a lot of sense.

The most significant issue inhibiting data sharing is biologists' lack of motivation to do it.

We are constantly hearing suggestions to make all data gathered in biology experiments available online. This is an appealing idea because most data that we collect from experiments never sees the light of day. A smattering of our data appears in papers, of course, but we all recognize that this is usually a highly selected subset of all that is collected, intended to support the story that is being touted at the moment. If we could somehow make all of our data available to the community, the idea goes, biological progress would be greatly accelerated.

Despite the appeal of making all biological data accessible, there are enormous hurdles that currently make it impractical. For one, sharing all data requires that we agree on a set of standards. This is perhaps reasonable for large-scale automated technologies, such as microarrays, but the logistics of converting every western blot, ELISA, and protein assay into a structured and accessible data format would be a nightmare—and probably not worth the effort.

Related Articles

Give P2P a Chance

Systems Biology: Beyond the Buzz

Open Access 2.0

This does not mean that some instances of widespread data-sharing are not extraordinarily useful. However, these tend to be independent of a particular experimental context, the obvious example being DNA sequence or protein structure data. Some databases can also be very useful if the context is reasonably constrained. For example, tissue-specific expression

profiles have proven useful, as have datasets gathered during different stages of development.

Unfortunately, most experimental data is obtained ad hoc to answer specific questions and can rarely be used for other purposes. Good experimental design usually requires that we change only one variable at a time. There is some hope of controlling experimental conditions within our own labs so that the only significantly changing parameter will be our experimental perturbation. However, at another location, scientists might inadvertently do the same experiment under different conditions, making it difficult if not impossible to compare and integrate the results.

The most significant issue inhibiting data sharing, however, is biologists' lack of motivation to do it. In order to sufficiently control the experimental context to allow reliable data sharing, biologists would be forced to reduce the plethora of cell lines and experimental systems to a handful, and implement a common set of experimental conditions. Getting biologists to agree to such an approach is akin to asking people to agree on a single religion. If you're still not convinced, consider the experience of the Alliance for Cell Signaling (AfCS).

The AfCS, headed by Nobel Prize winner Al Gilman, was the original National Institutes of Health "Glue Grant," and had the goal of creating a comprehensive description of the cellular response to signaling molecules. Over a period of five years, members created a huge collection of data, documenting the response of the RAW 264.7 mouse macrophage cell line to a select panel of stimuli. This ambitious project required rigorous control of experimental conditions, reagents, data collection, and analysis. Although the AfCS stopped collecting data several years ago, the data are still available on a Web site that receives more than 100,000 weekly page views. Yet, over the last five years, these freely available data have been used in only a handful of papers.

Why is such an impressive set of primary experimental data so rarely used? I suspect that most of the investigators who use RAW 264.7 cells are not interested in systematic input-output data, and most investigators who are interested in modeling of signaling networks are not using RAW 264.7 cells. In my own case, I am interested in the EGF receptor and receptor tyrosine kinases. This aspect of cell signaling was not covered in their dataset, and thus it is of no interest to me.

And soon, discussions about the importance of sharing may become moot, since the rapid pace of technology development is likely to eliminate much of the perceived need for sharing primary experimental data. High throughput analytical technologies, such as proteomics and deep sequencing, can yield data of extremely high quality and can produce more data in a single run than was previously obtained from years of work. It will thus become more practical for research groups to generate their own integrated sets of data than try to stitch together disparate information from multiple sources.

Steven Wiley is a Pacific Northwest National Laboratory Fellow and director of PNNL's Biomolecular Systems Initiative.