Comment

Seek and ye shall maybe find

Gregory A Petsko

Address: Rosenstiel Basic Medical Sciences Research Center, Brandeis University, Waltham MA 02454-9110, USA.
E-mail: petsko@brandeis.edu

Published: 10 November 2000

Genome ***Biology*** 2000, **1(5)**:comment1005.1–1005.2

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2000/1/5/comment/1005

The other day I became what George Orwell might have called an unperson. I was searching the US National Library of Medicine's Medline database for an article that I wrote a number of years ago. I wanted to cite it in a paper I was writing (hardly anyone else cites it these days, so I like to; besides, it actually was relevant), and it was easier just to download the citation than to dig it out of my reference database since I was logged on to Medline anyway. But I couldn't find it. Searching with my name, the names of my coauthors, the journal and year, the title keywords - all failed. Anyone who saw the classic, creepy 1960s television series *The Twilight Zone* will understand how I felt. I actually went to my files and dug out the reprint so I could be certain I really had written the paper and wasn't in the grip of some grotesque delusion. (At my age, it never hurts to check one's memory.)

It took me a while to realize what had happened. The paper was published in *Science*, a journal that covers all scientific disciplines, and although the subject matter was a new technique that we applied to structural biology, the title might have led one to conclude that the paper was about computer science. Someone obviously did conclude that, and so the reference was never entered into Medline, which focuses on medicine and the basic life sciences. Anyone searching for past publications on that topic who used only Medline to conduct the search would, in all probability, never learn that our work had been done.

The experience caused me to consider another of the consequences of the 'Age of Genomics': we are all becoming conditioned to rely on databases, and to believe that they constitute a complete repository of the available relevant information. This belief is largely justified for sequence databases (with the significant exception of some sequence data from the private sector), a fact that has contributed to a much less justified faith in other databases.

Nowhere is this phenomenon more apparent than in the blind trust students - and, let's face it, the rest of us - place in computerized literature searches. There is no denying their value. Assembling a collection of references for a research article or review is orders of magnitude faster using such tools than it would be by hand. But this efficiency comes with a hidden cost: many of the literature databases, especially the free ones, limit their coverage in ways that are not obvious to the casual user. I don't mean to pick on Medline specifically, but it is the most widely-used literature database, at least in the US, so it is useful to consider it as an example. As my experience indicates, not every article that belongs in Medline finds its way there.

Nor does every journal. The *Journal of the American Chemical Society* is not covered, despite the fact that every issue contains more than a dozen articles on biological chemistry. Most other chemistry journals are also omitted. And a number of 'physics-related' journals such as *Acta Crystallographica* are also not comprehensively covered, despite the fact that they contain important articles on structural biology methods and results. These are all excellent journals, yet the best advice one could give a modern young life-scientist is to avoid publishing in them if one is trying to build a reputation, because papers published there risk invisibility.

Then there is the matter of history. Reference databases do not go back to the dawn of time. Medline goes only as far back as the 1960s for most journals, and the bulk of entries more than about twenty years old have no abstracts and so are less likely to be useful. Many search systems that access Medline do not search the entire database, and many other databases contain only the entries for the past ten or twenty years. It's bad enough that so many scientists today don't know the older literature - and so are more likely to reinvent the wheel - but it's far worse to realize that it won't be long before that literature won't ever be cited. There is a joke in science that papers more than ten years old aren't worth reading, but soon this may be no laughing matter.

I think there are two things that can help this situation. First, the scientific community needs to pressure database curators to be as inclusive as possible. This means not only a broad coverage in terms of journals, but also deep coverage into the older literature and more careful decisions about which articles belong in which database. Indeed, it is arguable that genomics, being interdisciplinary by its very nature, needs literature databases that cover nearly all of the physical and life sciences. Second, young scientists need to check periodically (pun intended) that their work has found its way into the proper databases. Doing a regular search for one's own *oeuvre* and notifying database curators of omissions and errors would help prevent out-of-body experiences like mine. In the end, we are all going to be subject to the vagaries of such information repositories. Databases are our new servants, but they are also our new masters.