POLICY FORUM: INTELLECTUAL PROPERTY

Publication Rights in the Era of Open Data Release Policies

Lee Rowen, Gane K. S. Wong, Robert P. Lane, Leroy Hood

In 1996, an international group of principal investigators from the large sequencing centers working on the Human Genome Project (HGP) adopted data release principles known as the Bermuda Conventions (*1*). The key principle states "human genomic sequence information generated by centers funded for large-scale sequencing should be freely available and in the public domain…." Adherence was formalized as a condition for receiving funds from the National Human Genome Research Institute, the Department of Energy, the Wellcome Trust, and other agencies.

Immediate data release generates a potential conflict between data producers and third parties in that producers often draw a distinction between preliminary and final data. Official notification that the data are final is often captured by publication of comprehensive analyses in peer-reviewed journals. Publication of such analyses by third parties before the data producers have officially signed off preempts what producers consider to be their prerogative (*2*). This conflict underlies the recent controversy (*3*) regarding who has the right to publish analyses of the genomes of the malaria and sleeping sickness parasites.

The underlying problem is that the data are "out there" with no formal restrictions on use. Thus, it was possible for two review articles (*4*) based on the gene content of ~1.5 mB of annotated sequence from the mouse major histocompatibility complex to be published before the project was finished and primary publications were written by the data producers. Nothing except a journal's editorial policy (*5*) or the peer-review process prevents a third party from publishing sequence analyses accompanied merely by a reference to the database accession number. An alternative practice is to include an acknowledgement to the appropriate genome center(s). Neither form of attribution does much to benefit careers of data producers, particularly as accession numbers are not considered prior publications (*6*).

In the past, etiquette has guided decisions about publishing analyses of other people's sequences or annotations. Often, informal contacts have resulted in permission to publish, coauthorship, back-to-back papers, or other agreeable options. However, the Web fosters a climate of anonymity, in which the data content is divorced from the context of its acquisition, with the result that credit is often not properly given to data producers.

What to Do?

When a community benefits from freely shared information, it must take responsibility for ensuring that the producers of this information are appropriately rewarded. In the space sciences, there is a mechanism to reward the people who oversee construction of the large observatories, like the Hubble Space Telescope, that everyone shares (*7*). Scientists involved in developing the instruments are given a guaranteed amount of observing time, with an opportunity to specify their research objectives and the objects they plan to observe. Their program is made publicly available at the time of a general Call for Proposals, and other proposals are not allowed to duplicate their stated plans. They then have a proprietary period of time after the observations are completed to analyze the data and publish results.

What is different about large-scale sequencing is that the data are already freely available. However, citing an accession number is analogous to referencing a personal communication. Journals typically require letters of permission for an author to cite a personal communication of unpublished data. We therefore propose that permission from data producers be required before third parties can publish certain types of analyses.

Disputes are inevitable, given the vagueness with which boundary conditions on legitimate ownership can be formulated. Nonetheless, we believe that the research community must develop a policy with specific guidelines on the kinds of analyses for which data producers can claim priority. We suggest two criteria:

1) The analysis must be based directly on sequence from a limited number of producers. This eliminates cross-species or global analyses, which typically require unpublished data from many unrelated producers, and where it is almost impossible not to use data produced outside of one's own laboratory.

2) The analysis must address a question that the sequence producers could reasonably have planned. This eliminates unexpected discoveries. However, a cataloguing of all the genes in a particular organism or a major locus should be restricted because it is an "obvious" paper.

We suggest that databases add a tag or qualifier to an accessioned sequence entry that indicates whether the submitters require permission for publication of analyses based directly on that sequence. If the submitters do not require a request for permission, then no restrictions would be placed, and the accession number would be considered an acceptable reference. For accessioned sequences tagged as requiring permission, an indication of the sort of publication planned must be included as part of the database entry. An abstract would be required, so that potential journal editors and reviewers could decide whether a publication claim right has been violated. Finally, there must be a mechanism, perhaps instituted by journals, for resolving disputes and enforcing the policy. There will be instances where third parties insist that sequence producers have claimed too much territory or where sequence producers have not published their analyses within a reasonable time.

Such a policy would diffuse the argument that "if we don't publish it someone else will" and would have implications beyond sequencing. Similar conflicts are destined to arise in functional genomics and proteomics. If this issue is not addressed, then the laudable precedent set by the HGP is less likely to be adopted by other data-gathering efforts. This would be a loss to both science and society.

L. Rowen and L. Hood are at the Institute for Systems Biology, Seattle, WA 98105, USA. G.K.S. Wong and R.P. Lane are in the Departments of Medicine and of Molecular Biotechnology, respectively, at the University of Washington, Seattle, WA 98195, USA.

References and Notes

1. www.wellcome.ac.uk/en/1/biopoldat.html and www.nhgri.nih.gov/Grant_info/Funding/Statements/RFA/data_release.html.
2. For example, see the discussion regarding publication at the malaria genome sequencing project's Web site http://e2kroos.cis.upenn.edu/release_oq.html
3. See *Nature* **405**, 719 (2000); C. Macilwain, *Nature* **405**, 601 (2000); M. Gottlieb *et al.*, *Nature* **406**, 121 (2000).
4. R. J. N. Allcock *et al.*, *Immunol. Today* **21**, 328 (2000); C. Y. Yu *et al.*, *Immunol. Today* **21**, 320 (2000).
5. For example, editorial policy for *Genome Research* states: "…any sequence information obtained from center sites MUST be acknowledged in the manner approved by that center, or, when the center has no standard guidelines, the center must be appropriately acknowledged as the source of the information,…in any instance where individuals at a sequencing center have contributed more than just standard release information, these individuals should be contacted, shown the manuscript, and their status on the paper...(whether they prefer authorship or not) determined prior to submission." www.genome.org/misc/ifora.stml.
6. *Nature* **405**, 719 (2000). "We believe that genomics databases, like preprint servers and conferences, represent a form of intra-community networking…. *Nature* does not count them as prior publications. This policy applies not only to raw sequence data but also to 'annotations'…."
7. www.stsci.edu/ftp/proposer/cycle10/cp/cp10-5.html #pgfld-1818439. We thank M. Hauser for explaining these policies to us.