*Science* 16 February 2001:
Vol. 291. no. 5507, pp. 1260 - 1261
DOI: 10.1126/science.291.5507.1260

VIEWPOINT

**COMPUTATIONAL BIOLOGY:**

# Bioinformatics--Trying to Swim in a Sea of Data

**David S. Roos***

Advances in many areas of genomics research are heavily rooted in engineering technology, from the capillary electrophoresis units used in large-scale DNA sequencing projects, to the photolithography and robotics technology used in chip manufacture, to the confocal imaging systems used to read those chips, to the beam and detector technology driving high-throughput mass spectroscopy. Further advances in (for example) materials science and nanotechnology promise to improve the sensitivity and cost of these technologies greatly in the near future. Genomic research makes it possible to look at biological phenomena on a scale not previously possible: all genes in a genome, all transcripts in a cell, all metabolic processes in a tissue.

One feature that all of these approaches share is the production of massive quantities of data. GenBank, for example, now accommodates >$10^{10}$ nucleotides of nucleic acid sequence data and continues to more than double in size every year. New technologies for assaying gene expression patterns, protein structure, protein-protein interactions, etc., will provide even more data. How to handle these data, make sense of them, and render them accessible to biologists working on a wide variety of problems is the challenge facing bioinformatics--an emerging field that seeks to integrate computer science with applications derived from molecular biology. We are swimming in a rapidly rising sea of data...how do we keep from drowning?

Bioinformatics faces its share of growing pains, many of which presage problems that all biologists will soon encounter as we focus on large-scale science projects. For starters, few scientists can claim a strong background on both sides of the divide separating computer science from biomedical research. This shortage means a lack of mentors who might train the next generation of "bioinformaticians." Lack of familiarity with the intellectual questions that motivate each side can also lead to misunderstandings. For example, writing a computer program that assembles overlapping expressed sequence tag (EST) sequences may be of great importance to the biologist without breaking any new ground in computer science. Similarly, proving that it is impossible to determine a globally optimal phylogenetic tree under certain conditions may constitute a significant finding in computer science, while being of little practical use to the biologist. Identifying problems of intellectual value to all concerned is an important goal for the maturation of computational biology as a distinct discipline. "Real" biology is increasingly carried out in front of a computer, while an increasing number of projects in computer science will be driven by biological problems.

Further difficulties stem from the fact that bioinformatics is an inherently integrative discipline, requiring access to data from a wide range of sources. Without the underlying data, and the ability to combine these data in new and interesting ways, the field of bioinformatics would be very much limited in scope. For example, the widespread utility of BLAST for the identification of gene similarity (1) is attributable not only to the algorithm itself (and its implementation), but also to the availability of databases such as GenBank, the European Molecular Biology Laboratory (EMBL), and the DNA Data Bank of Japan (DDBJ), which pool genomic data from a variety of sources. BLAST would be of limited utility without a broad-based database to query.

One core aspect of research in computational biology focuses on database development: how to integrate and optimally query data from (for example) genomic DNA sequence, spatial and temporal patterns of mRNA expression, protein structure, immunological reactivity, clinical outcomes, publication records, and other sources. A second focus involves pattern recognition algorithms for such areas as nucleic acid or protein sequence assembly, sequence alignment for similarity comparisons or phylogeny reconstruction, motif recognition in linear sequences or higher-order structure, and common patterns of gene expression. Both database integration and pattern recognition depend absolutely on accessing data from diverse sources, and being able to integrate, transform, and reproduce these data in new formats.

As noted above, computational biology is a fundamentally collaborative discipline, owing its very existence to the availability of rich and extensive data sets for analysis, integration, and manipulation. Data accessibility and usability are therefore critical, raising concerns about data release policies--what constitutes primary data, who owns this resource, when and how data should be released, and what restrictions may be placed on further use. Two challenges have emerged that could potentially restrict the advancement of bioinformatics research: (i) questions related to the appropriate use of data released before publication and (ii) restrictions on the reposting of published data.

The first challenge to bioinformatics research relates to the analysis of data posted on the Web in advance of publication. Recognizing the value of early data release for a wide range of studies, the Human Genome Project adopted a policy of prepublication data release (2), and many genome projects (and the funding agencies that support them) now adhere to similar rules. Because bioinformatics depends absolutely on the ability to integrate data from a wide variety of sources, it is to be hoped that other projects that generate genomic-scale data (including expression analysis and proteomics research) will follow a similar policy (3), because immensely valuable results can emerge from large-scale comparative studies of genome structure, microarray data, protein interactions, and so on (4-6). The success of such altruistic data release policies, however, requires that those who generate primary sequence data (often on behalf of the community at large) receive appropriate recognition and are able to derive intellectual satisfaction from their work. Rowen *et al.* (7) have recently proposed treating unpublished data available on the Web as analogous to "personal communication," thereby establishing some degree of intellectual property protection.

The difficulty with this approach comes in determining what types of analysis should require permission from the submitters, and what types of analysis can reasonably be prohibited. Clearly, the identification of individual genes of interest for further experimental analysis must be acceptable--perhaps even without the need for formal permission--otherwise, early data release serves no purpose at all. Conversely, second-party publication of raw, unpublished, sequence data posted on the Web must be viewed as violating ethical standards--analogous to the verbatim plagiarism of unpublished results from a meeting presentation. Where to draw the line in intermediate cases will ultimately depend on the intellectual contributions provided by the manuscript in question, and whether such work might reasonably have been expected to emerge in due course from those who generated the original data (7). Such considerations of "value added" are not terribly different from those normally applied during manuscript review, but require special consideration by reviewers and editors of the anticipated contributions from the original submitter.

Experience with the *Plasmodium falciparum* genome project (8-15) suggests that disagreements over what kinds of data and analyses are permissible for publication are sometimes attributable to the failure of second parties to adequately consider the interests and involvement of those generating the primary data. More often, however, disputes are attributable to a lack of understanding: either on the part of biologists, who do not fully appreciate the long lag that may reasonably be expected between (for example) the first appearance of shotgun sequencing results and final sequence closure and annotation, or on the part of those generating the primary data, who may not fully appreciate the intellectual contributions of biologists/bioinformaticians. One hopes that as the gulf between those engaged in the application of genomic technologies, bioinformatics research, and laboratory analysis is bridged by understanding, these problems will diminish in importance. Increased acceptance of Web-based release as a form of publication (for hiring, promotion, tenure decisions, etc.), as well as increased understanding of the nature of "big science" projects in biology, will also reduce tensions.

The second challenge to bioinformatics research derives not from restrictions on data access but from restrictions on downstream use, such as incorporation into new or existing databases. This challenge is of a more fundamental nature, involving not just when bioinformatic analysis is permissible, but what kinds of analyses can be carried out. Today's publication of a draft analysis of the human genome by Celera Genomics (16) focuses a spotlight on this question, because the primary data themselves are being released only through a private company that places restrictions on the reposting and redistribution of their data. Other genome-scale projects, including a recent analysis of protein-protein interactions in *Helicobacter pylori* (17), have placed similar restrictions on the reposting of primary data.

As described in the accompanying editorial (18), *Science* has taken care to craft a policy which guarantees that the data on which Celera's analyses are based will be available for examination. But the purpose of insisting that primary scientific data be released is not merely to ensure that the published conclusions are correct, but also to permit building on these results, to allow further scientific advancement. Bioinformatics research is particularly dependent on unencumbered access to data, including the ability to reanalyze and repost results. Thus the statement that "... any scientist can examine and work with Celera's sequence in order to verify or confirm the conclusions of the paper, perform their own basic research, and publish the results" (19) is inaccurate with respect to research in bioinformatics. For example, a genome-wide analysis and reannotation of additional features identified in Celera's database could not be published or posted on the Web without compromising the proprietary nature of the underlying data. Nor could this information be combined with the resources available from other databases--such as the information from additional species necessary for cross-species comparisons, or data from microarray and proteomics resources that would permit queries based on a combination of genome sequence data, expression patterns, and structural information. It is certainly true that the present state of genomics research would never have been achieved without the freedom to use (properly attributed) information from GenBank/EMBL/DDBJ.

The potential for restricting downstream analysis offers the prospect of making a wealth of proprietary data generated by private companies accessible to the research community at large, but this potential comes at a very great cost. Imagine, for example, genomics research in a world where GenBank/EMBL/DDBJ did not exist and could not be assembled because of ownership restrictions. Five years ago, the Bermuda Conventions (2) established a standard for the release of genome sequence data that has served biologists very well; we should consider carefully what precedent to establish for the next 5 years, as considerations of data-release and data-use policy are likely to have far-reaching implications for all of biomedical research.

The "postgenomic era" holds phenomenal promise for identifying the mechanistic bases of organismal development, metabolic processes, and disease, and we can confidently predict that bioinformatics research will have a dramatic impact on improving our understanding of such diverse areas as the regulation of gene expression, protein structure determination, comparative evolution, and drug discovery. The availability of virtually complete data sets also makes negative data informative: by mapping entire pathways, for example, it becomes interesting to ask not only what is present, but also what is absent. As the potential of genomics-scale studies becomes more fully appreciated, it is likely that genomics research will increasingly come to be viewed as indistinguishable from biology itself. But such research is only possible if data remain available not only for examination, but also to build upon. It is hard to swim in a sea of data while bound and gagged!

**References and Notes**


1. S. F. Altschul *et al.*, *J. Mol. Biol.* **215**, 403 (1990) **[Medline]**.
2. **www.wellcome.ac.uk/en/1/biopoldat.html**; **www.nhgri.nih.gov/Grant_info/Funding/Statements/RFA/data_release.html**, see also **www.usinfo.state.gov/topical/global/biotech/00031401.htm**.
3. A. Brazma *et al.*, *Nature* **403**, 699 (2000) **[Medline]**.
4. R. L. Tatusov *et al.*, *Science* **278**, **631** (1997).
5. R. L. Tatusov *et al.*, *Nucleic Acid Res.* **29**, 22 (2001) **[Medline]**.
6. S. Chu *et al.*, *Science* **282**, **699** (1998).
7. L. Rowen, G. K. S. Wong, R. P. Lane, L. Hood, *Science* **289**, **1881** (2000); see also letter from E. Bell, response from L. Rowen and L. Hood, *Science* **290**, **1696** (2000), and letter from R. W. Hyman, *Science* **291**, **827** (2001).
8. M. J. Gardner, *Science* **282**, **1126** (1998).
9. S. Bowman *et al.*, *Nature* **400**, 532 (1999) **[Medline]**.
10. R. F. Waller *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 12352 (1998) **[Medline]**.
11. H. Jomaa *et al.*, *Science* **285**, **1573** (1999).
12. S. A. Kyes, J. A. Rowe, N. Kriek, C. I. Newbold, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 9333 (1999) **[Medline]**.
13. *Nature* **405**, 719 (2000) **[Medline]**.
14. C. Macilwain, *Nature* **405**, 601 (2000) **[Medline]**; see also letter from M. Gottlieb *et al.*, *Nature* **406**, 121 (2000) **[Medline]**.
15. The *Plasmodium* Genome Database Collaborative, *Nucleic Acids Res.* **29**, 66 (2001) **[Medline]**.
16. C. Venter *et al.*, *Science* **291**, **1304** (2001).

17. J.-C. Rain *et al*., *Nature* **409**, 211 (2001) **[Nature]**.
18. B. Jasny, D. Kennedy, *Science* **291**, **1153** (2001).
19. **http://www.sciencemag.org/feature/data/announcement/genomesequenceplan.shl**
20. I would like to thank my many colleagues in the computational biology research community for helpful discussions on the impact of data release policy decisions on bioinformatics research, and for comments on this manuscript.

The author is at the Department of Biology and Genomics Institute, University of Pennsylvania, Philadelphia, PA 19104, USA. E-mail: droos@sas.upenn.edu