the outer 29% of its radius. Inside this point, it rotates like a solid body, whereas throughout the convection zone it shows appreciable differential rotation. In addition, helioseismology has allowed us to make a very accurate model of the Sun, and this showed that the solar neutrino problem, the deficit of detected versus predicted neutrinos from the Sun, could not be explained by an incomplete understanding of the Sun's interior. As we now know, the resolution of this problem came from new particle physics in the form of neutrino oscillations (*6*). Helioseismology has also allowed us to measure the evolutionary state of the Sun, including constraints on its age and chemical composition.

If the Sun were as distant as other stars, we would no longer be able to see the millions of pulsation modes; we would be limited to around 100 or less. Having refined our techniques and experience based on the Sun, however, this number of pulsation modes is more than sufficient for providing accurate constraints on the parameters and structure of other stars (*7*). This is fortunate, because stellar models that are calibrated to the Sun are unlikely to be as accurate for stars of different mass and temperature. The work of Michel *et al.* represents a substantial first step in understanding stars somewhat more massive than the Sun (the masses of their targets were between 1.17 and 1.4 solar masses). Their data show that the oscillation amplitudes are 1.5 times as large as those of the Sun, which is still about 25% lower than theoretical estimates. This is noteworthy because these amplitudes depend on the nature of convection in the outer layers of these stars. In addition, they were able to measure properties of the granules seen at the surface of these stars, a further manifestation of convection within them. This provides us with the opportunity to test and refine our theories of convection, which is one of the largest sources of uncertainty in the modeling of stars. And, perhaps most important, their results showed that the expected oscillations were present with measurable amplitudes, thereby demonstrating the viability of space-based investigations.

Further observations by CoRoT and the upcoming NASA mission Kepler should yield a wealth of information on other solar-like stars. It will then be possible to place tighter constraints on quantities such as the total mass, luminosity, radius, age, and rotation rate of a very large number of stars. These space-based observations will test our understanding of physical processes, such as convection in the envelopes and cores of stars, but will also enhance our understanding of the ages and evolutionary states of objects throughout our Galaxy and the local universe.

References

1. E. Michel *et al.*, *Science* **322**, 558 (2008).
2. A. Sandage, G. A. Tammann, *Astrophys. J.* **151**, 531 (1968).
3. D. E. Winget *et al.*, *Astrophys. J.* **430**, 839 (1994).
4. D. E. Winget *et al.*, *Astrophys. J.* **378**, 326 (1991).
5. D. O. Gough *et al.*, *Science* **272**, 1281 (1996).
6. Q. R. Ahmad *et al.*, *Phys. Rev. Lett.* **89**, 011301 (2002).
7. T. S. Metcalfe, http://arXiv.org/abs/0808.3136 (2008).

10.1126/science.1164633

GENETICS

GenBank—Natural History in the 21st Century?

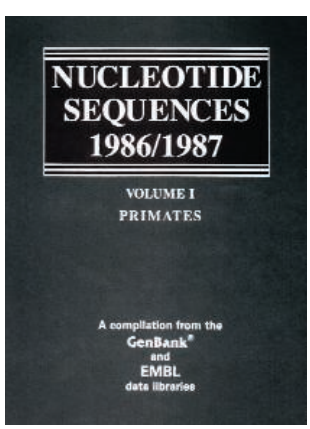
Since its foundation, the nucleic acid sequence database GenBank has merged the values of natural history with those of the experimental sciences.

Bruno J. Strasser

The American nucleic acid sequence database GenBank, as well as its European (European Molecular Biology Laboratory, EMBL) and Japanese (DNA Databank of Japan, DDBJ) mirror organizations, each contain far more nucleotides than "the number of stars in the Milky Way," as the U.S. National Institutes of Health (NIH) once put it in a press release (*1*). The early history of this database illustrates the transformation of biology into a new science that links the methods of natural history with those of experimentation.

GenBank represents the cutting edge of biology, but it also belongs to the centuries-old tradition of natural history—a tradition best characterized as the practice of collecting, describing, naming, comparing, and organizing natural objects. The method applies equally to plants, bones, or molecular sequences. This view challenges the received historical picture, in which the experimental sciences overtook natural history in the late 19th century and triumphed in the mid-20th century with the rise of molecular biology. As GenBank and other databases attest, the practices of natural history have been imported into the experimental sciences.

In March 1979, 30 molecular biologists and computer scientists meeting at the Rockefeller University in New York agreed on the necessity to create a national, computerized database (*2*). The impetus was the same as that behind most natural history collections, which were often created in reaction to a perceived overabundance of information—for example, when the expansion of European travel to the New World led to the accumulation of previously unknown specimens. This time, it was the explosive growth in the number of known DNA sequences and the promise of producing biological knowledge by analyzing and comparing them that made a database seem indispensable. Several scientists were maintaining individual sequences collections, but none was comprehensive.



CREDIT: ACADEMIC PRESS, INC.

It took almost 3 years for the NIH to come up with a funding scheme, and by that time, the EMBL had already made its own sequence database publicly available. This delay—somewhat embarrassing for the NIH—resulted not only from bureaucratic inertia, as some have argued, but also from uncertain scientific prospects for a natural history collection at a time when experimentation triumphed in revealing the secrets of nature. Frederick Sanger would later put it bluntly: " 'Doing' for a scientist implies doing experiments" (*3*). Pressed by a few vocal experimental scientists the NIH eventually issued a Request for Proposals. Two

Section of the History of Medicine, Yale University, New Haven, CT 06520, USA. E-mail: bruno.strasser@yale.edu

applications competed for the contract: one by a team headed by Margaret O. Dayhoff (1925 to 1983) at the National Biomedical Research Foundation (NBRF) and one by a group of researchers around Walter Goad (1925 to 2000) at Los Alamos National Laboratory, collaborating with the private company Bolt Beranek and Newman (BBN).

Dayhoff had pioneered sequence databases by gathering an extensive collection of protein sequences since the early 1960s. She believed that a "tremendous amount of information regarding evolutionary history and biochemical function [is] implicit in each sequence" and that it was essential to "collect this significant information, correlate it into a unified whole and interpret it" (*4*). In a series of volumes entitled *The Atlas of Protein Sequence and Structure*, published beginning in 1965, she presented the world's largest collection of protein and nucleic acid sequences, innovative methods to analyze them, and evolutionary inferences drawn from them.

The *Atlas* became immensely popular as a reference tool among molecular and evolutionary biologists. Dayhoff had hoped that researchers would share protein sequences directly with her before they were published. This model of data collection was typical of natural history, but proved unsuccessful with experimentalists, because inclusion in the *Atlas* established neither authorship nor priority. Dayhoff and her team were left with the painstaking task of surveying manually the published literature.

The competing project for the NIH contract came from Los Alamos, where modest biomedical research had been carried out since the Manhattan Project. When he heard about the conclusions of the Rockefeller meeting, Walter Goad became convinced that Los Alamos was "the natural place to locate a center for sequence analysis of DNA," mainly because of the national laboratory's "unique computer facility" (*5*). Goad, too, began to gather nucleic acid sequences, mostly from other collections such as those of Richard Grantham in France, Kurt Stüber in Germany, and Douglas L. Brutlag and Elvin A. Kabat in the United States.

The NBRF (Dayhoff) and Los Alamos-BBN (Goad) proposals to the NIH for a centralized database were very similar, yet they contained key differences concerning property, privacy, and priority in science. The NBRF proposed to collect sequences by exploring the published literature and inviting experimentalists to submit their data. It considered sequences much as naturalists considered specimens—as unencumbered natural objects free to be collected and appropriated. The Los Alamos–BBN proposal, by contrast, suggested that journal editors would be asked to make the submission of sequences to the database a condition for the publication of an article. This system was aligned with the reward system of the experimental sciences, in which research results were considered private knowledge until they were published and authorship was attributed. Publication was the main incentive and reward for making knowledge public.

The most essential difference between the proposals, however, was also the most subtle. In 1980, the U.S. Supreme Court had declared that "anything under the sun that is made by man," including genetically modified organisms, could be patented (*6*). The NIH grew deeply concerned about who would own the data in the future database. Goad stressed that he "did not intend to assert any proprietary interest whatsoever in any data," and pointed out that Dayhoff and her team had "sought revenues from sales of their database and prevented redistribution" (*4*), without mentioning that the revenues were only meant to cover expenses, not make a profit.

Los Alamos and BBN were further able to boost the openness of their database by offering to distribute it through the Department of Defense–operated computer network ARPANET, whereas the NBRF could offer only limited online access through telephone modems. On 30 June 1982, the NIH signed a contract with Los Alamos and BBN for the establishment of a public and free nucleic acid sequence databank, which would soon be called GenBank.

Yet, the success of GenBank in collecting all published sequences was only made possible by two crucial developments. First, it instituted a tight collaboration with the EMBL database, established a few months earlier at Heidelberg, and, after 1986, with the DDBJ. Each database assumed responsibility for a subset of journals—a division that mirrored the tendency of early natural history collections to focus on specimens of a certain region and relying on exchanges with other collections for the rest. The collaborations among GenBank, EMBL, and DDBJ required delicate negotiations to establish common entry standards; database managers could not rely on an equivalent of Linnaeus's binomial system, which allowed easy exchanges between natural history collectors.

Second, as the DNA databases increasingly lagged behind the exploding number of DNA sequences, they succeeded in convincing journal editors to make electronic submission a condition for publication, solving the problem of data collection. The substantial resistance to the Human Genome Project in the late 1980s, just like that to GenBank a few years earlier, was in part due to its association with the natural history enterprise. To some, a large-scale collection of gene sequences did not sound any more exciting than a large-scale collection of butterflies (*7*).

Since then, the convergence between natural history and the experimental sciences has grown ever closer, for example, with the rise of DNA barcoding in taxonomy or the cross-linking of museum specimens and DNA sequence entries in GenBank, as is done with specimens at the Museum of Vertebrate Zoology at Berkeley. Taking into account GenBank's place in the natural history tradition can help solve some of its most vexing problems, such as obsolete or inaccurate annotations. For example, cumulative annotations have been used for centuries to accommodate the changing knowledge about specimens in natural history collections, and have now been proposed for GenBank (*8*).

Another challenge facing the DNA databases is the development of sequence and other biological databases based on the principles of open-content resources such as Wikipedia (*9*, *10*). These databases could make, critics argue, GenBank/EMBL/DDBJ obsolete, much as the *Encyclopedia Britannica* seems doomed to become. If it is any indication of the future, natural history has thrived by relying on the input of a broad community, including amateurs—not by relying solely on the expertise of its curators.

References

1. NIH, "Public collections of DNA and RNA sequence reach 100 gigabases" (22 August 2005).
2. T. E. Smith, *Genomics* **6**, 701 (1990).
3. Margaret O. Dayhoff Papers, Archives of the National Biomedical Research Foundation, Washington, DC.
4. F. Sanger, *Annu. Rev. Biochem.* **57**, 1 (1988).
5. Walter B. Goad Papers, Archives of the American Philosophical Society, Philadelphia.
6. *Diamond v. Chakrabarty*, 447 U.S. 303 (1980).
7. R. Cook-Deegan, *The Gene Wars: Science, Politics, and the Human Genome* (Norton, New York, 1994), chap. 8.
8. M. I. Bidartondo *et al.*, *Science* **319**, 1616 (2008).
9. E. Pennisi, *Science* **319**, 1598 (2008).
10. J. Giles, *Nature* **445**, 691 (2007).

10.1126/science.1163399

[1]Division of Paleontology, Steinmann Institute, University of Bonn, D-53115 Bonn, Germany. [2]Institut für Biologie und Sachunterricht und ihre Didaktik, Universität Flensburg, Auf dem Campus 1, D-24943 Flensburg, Germany.

*To whom correspondence should be addressed. E-mail: martin.sander@uni-bonn.de

References and Notes

1. K. A. Stevens, J. M. Parrish, *Science* **284**, 798 (1999).
2. K. A. Stevens, J. M. Parrish, in *The Sauropods: Evolution and Paleobiology,* K. A. Curry Rogers, J. A. Wilson, Eds. (Univ. of California Press, Berkeley, CA, 2005), pp. 178–200.
3. G. S. Paul, *Mod. Geol.* **23**, 179 (1998).
4. A. Christian, *Mitteilungen aus dem Museum für Naturkunde in Berlin, Geowissenschaftliche Reihe* **5**, 271 (2002).
5. G. Dzemski, A. Christian, *J. Morphol.* **268**, 701 (2007).
6. A. Christian, G. Dzemski, *Fossil Rec.* **10**, 38 (2007).
7. J. Hummel *et al.*, *Proc. R. Soc. London Ser. B* **275**, 1015 (2008).
8. J. S. McIntosh, W. E. Miller, K. L. Stadtman, D. D. Gillette, *Brigham Young Univ. Geol. Stud.* **41**, 73 (1996).
9. C. Wiman, *Palaeontol. Sinica Ser. C* **6**, 1 (1929).
10. A. H. Clarke, *J. Vestibular Res.* **15**, 65 (2005).
11. The authors thank all members of DFG Research Unit 533 for discussion. This is contribution number 52 of this research unit.

Specimens Versus Sequences

IN HIS PERSPECTIVE "GENBANK—NATURAL history in the 21st century?" (24 October 2008, p. 537), B. J. Strasser claims that GenBank follows the tradition of natural history studies. I argue that GenBank is inconsistent with some important aspects of the tradition of natural history and it alone does not and will not constitute natural history.

First, GenBank has been designed to store molecular information about model organisms or humans, whereas natural history serves to explore and document biodiversity. GenBank's format is incapable of handling unique aspects of biodiversity studies such as diverse and large collections of specimens, and taxonomic uncertainties and revisions. Second, GenBank does not require vouchering of specimens, DNA extracts, or other molecular products, whereas the study of natural history always anchors information with specimens. Without such anchoring, revisionary work (which is part of the tradition of natural history) cannot be conducted. As a result, GenBank contains a considerable amount of unidentified or misidentified sequences (*1*–*3*). Third, GenBank is unlikely to reach amateurs in the way that natural history has. Study of natural history generally does not require expensive equipment and is exciting and rewarding. In comparison, DNA sequencing is costly, offers little aesthetic reward or recreational value, and requires specialist knowledge. It is hard to imagine amateurs comparing or exchanging their collections of DNA sequence instead of actual specimens.

Natural history documentation has many important facets, such as specimen locality data, images of morphology, and ecological notes. The Global Biodiversity Information Facility (www.gbif.org) and Morphbank (www.morphbank.net) are good examples. Specimen collections will continue serving as the primary record of natural history and biodiversity, while molecular data are supplementary.

GUANYANG ZHANG

Department of Entomology, University of California, Riverside, Riverside, CA 92521, USA. E-mail: Guanyang.Zhang@email.ucr.edu

References

1. P. D. Bridge, P. J. Roberts, B. M. Spooner, G. Panchal, *New Phytol*. **160**, 43 (2003).
2. R. H. Nilsson *et al.*, *PLoS ONE* **1**, e59 (2006).
3. G. Valkiūnas, C. T. Atkinson, S. Bensch, R. N. M. Sehgal, R. E. Ricklefs, *Trends Parasitol*. **24**, 247 (2008).

Response

ZHANG MISREPRESENTS THE PAST OF NATURAL history and the present of molecular sequence collection. His Letter reflects the exact tension I described in my paper between the naturalist and experimentalist perspectives.

First, although GenBank's objective may be to document model organisms, in February 2009 there were no fewer than 187,136 species represented in GenBank (*1*)—i.e., 5 to 10% of all known species. GenBank is thus already an extraordinary sampling of biological diversity, unmatched by any museums of natural history, and on par with the morphological databanks mentioned by Zhang.

Second, GenBank certainly contains a "considerable amount of unidentified or misidentified sequences," but all significant natural history museums likewise contain "unidentified and misidentified" specimens. Indeed, because systematic vouchering of specimens strengthens the reliability of the taxonomic assignment of sequence data (if the specimens are actually available to researchers), for more than a decade GenBank has included a data field for the specimen voucher. Presently, more than 623,000 sequences contain some kind of voucher information (*1*).

Third, sequence data in GenBank have indeed been provided mainly by professionals, although that might change very soon with the falling cost of genome sequencing. However, natural history collections are no different, given that most specimens have been identified by professionals, not by amateurs, as the history of taxonomy amply demonstrates (*2*). Zhang romanticizes natural history when he suggests that it "does not require expensive equipment"; most natural history endeavors have been anything but inexpensive. Surveys such as those of the fauna and flora of the West Indies in the 1780s, or of the American West in 1900 (*3*, *4*), were both characteristic Big Science projects, and the Kew Gardens or the American Museum of Natural History are not exactly cottage industries.

Zhang seems to fear that sequence data will replace specimen collections. He ignores the fact that natural history is not fundamentally about specimens, but about natural facts. Specimens are among these, but so are rocks, bones, and now molecular sequences. Thus, in his critique, Zhang repeats a concern that numerous naturalists have voiced since the beginning of the 20th century: that the experimental sciences would take over natural history. In the 1960s, molecular evolutionists led many of their naturalist colleagues to believe exactly that. They mocked the naturalists' methods based on morphological comparisons by asking "how many vertebrae does a sponge have?" (*5*).

In my Perspective, I argued that this controversy is becoming increasingly irrelevant in the 21st century, as the boundaries between specimen collections and molecular data collections are becoming increasingly blurred. With many GenBank sequences now linked to specimens in natural history museums, the convergence of specimen and sequence collections has proceeded smoothly precisely because both stand in the same tradition of natural history.

BRUNO J. STRASSER

Section of the History of Medicine, Yale University, New Haven, CT 06511, USA. E-mail: bruno.strasser@yale.edu

References

1. National Center for Biotechnology Information (www.ncbi.nlm.nih.gov).
2. J. Endersby, *Imperial Nature: Joseph Hooker and the Practices of Victorian Science* (Univ. of Chicago Press, Chicago, 2008).
3. L. L. Schiebinger, C. Swan, Eds., *Colonial Botany: Science, Commerce, and Politics in the Early Modern World* (Univ. of Pennsylvania Press, Philadelphia, 2005).
4. R. E. Kohler, *All Creatures: Naturalists, Collectors, and Biodiversity, 1850–1950* (Princeton Univ. Press, Princeton, NJ, 2006).
5. W. M. Fitch, E. Margoliash, *Science* **155**, 279 (1967).

Letters to the Editor

Letters (~300 words) discuss material published in *Science* in the previous 3 months or issues of general interest. They can be submitted through the Web (www.submit2science.org) or by regular mail (1200 New York Ave., NW, Washington, DC 20005, USA). Letters are not acknowledged upon receipt, nor are authors generally consulted before publication. Whether published in full or in part, letters are subject to editing for clarity and space.