The Experimenter's Museum: GenBank, Natural History, and the Moral Economies of Biomedicine
Author(s): Bruno J. Strasser
Source: *Isis*, Vol. 102, No. 1 (March 2011), pp. 60-96
Published by: The University of Chicago Press on behalf of The History of Science Society
Stable URL: http://www.jstor.org/stable/10.1086/658657
Accessed: 12/04/2011 15:54

The Experimenter's Museum

GenBank, Natural History, and the Moral Economies of Biomedicine

By Bruno J. Strasser*

ABSTRACT

Today, the production of knowledge in the experimental life sciences relies crucially on the use of biological data collections, such as DNA sequence databases. These collections, in both their creation and their current use, are embedded in the experimentalist tradition. At the same time, however, they exemplify the natural historical tradition, based on collecting and comparing natural facts. This essay focuses on the issues attending the establishment in 1982 of GenBank, the largest and most frequently accessed collection of experimental knowledge in the world. The debates leading to its creation—about the collection and distribution of data, the attribution of credit and authorship, and the proprietary nature of knowledge—illuminate the different moral economies at work in the life sciences in the late twentieth century. They offer perspective on the recent rise of public access publishing and data sharing in science. More broadly, this essay challenges the big picture according to which the rise of experimentalism led to the decline of natural history in the twentieth century. It argues that both traditions have been articulated into a new way of producing knowledge that has become a key practice in science at the beginning of the twenty-first century.

EXPERIMENTATION IS OFTEN SINGLED OUT as the most distinctive feature of modern science. Our very idea of modern science, extending back to the Scientific Revolution, gives a central place to the emergence of this particular way of producing

* Program in the History of Science and Medicine, Section of the History of Medicine, Yale University, P.O. Box 208015, New Haven, Connecticut 06520-8015; bruno.strasser@yale.edu.

I would like to thank Carl W. Anderson, Winona C. Barker, Dennis Benson, Howard S. Bilofsky, Douglas L. Brutlag, Christian Burks, Graham N. Cameron, Christine K. Carrico, Judith E. Dayhoff, Ruth E. Dayhoff, Peter Friedland, Greg Hamm, Elke Jordan, Laurence H. Kedes, Ruth L. Kirschstein, Robert S. Ledley, David J. Lipman, Richard J. Roberts, Temple F. Smith, C. Frank Starmer, Michael S. Waterman, and Norton Zinder for sharing material and memories. I also thank the participants at the workshops and conferences at Princeton University, the University of Pennsylvania, Johns Hopkins University, the Max Planck Institute for the History of Science, and Yale University, as well as Helen Curry, Angela Creager, Michael Mahoney, Brendan Matz, Dan Kevles, Robert Kohler, Joe November, Bill Summers, Frank Snowden, the three anonymous referees for *Isis*, and Bernie Lightman, for enlightening comments.

Isis, 2011, 102:60–96

0021-1753/2011/10201-0003$10.00

knowledge.[1] The current epistemic, social, and cultural authority of science rests largely on the possibility of experimentation in the laboratory. Even though the historiography of science over the last four centuries reminds us that experimentation has been only one of many ways in which scientific knowledge has been produced, by most accounts experimentation has come to dominate all other methods in most fields, from high energy physics to psychology.[2] Historians have argued that, since the nineteenth century, the rise of the experimental approach in the life sciences has led to the progressive demise of the natural history tradition, culminating at the end of the twentieth century in the triumph of experimentalism as epitomized by the success of molecular biology.[3] By casting experimentalism against the natural history tradition and publicly displaying their contempt for natural history, experimental biologists—and molecular biologists in particular—have endorsed this narrative in support of their quest for disciplinary control over the life sciences.[4]

In this essay, I would like to question this received picture by drawing attention to the importance of collections and natural historical practices for the production of knowledge in late twentieth-century biomedical sciences.[5] I intend to support a larger historical argument: namely, that the twentieth century did not witness the replacement of natural history by experimentalism, nor even their juxtaposition, but, rather, a new articulation of these two traditions into a "hybrid culture" reminiscent of Baconian "experimental history."[6] This way of producing knowledge has become a key practice in science at the

[1] See, e.g., Steven Shapin, *The Scientific Revolution* (Chicago: Univ. Chicago Press, 1996).

[2] For an overview of "ways of knowing" in the history of science, technology, and medicine see John V. Pickstone, *Ways of Knowing: A New History of Science, Technology, and Medicine* (Manchester: Manchester Univ. Press, 2000); and Pickstone, "Working Knowledges Before and After circa 1800: Practices and Disciplines in the History of Science, Technology, and Medicine," *Isis*, 2007, *98*:489–516.

[3] This narrative is the main theme of William Coleman's and Garland Allen's classic work. See William R. Coleman, *Biology in the Nineteenth Century: Problems of Form, Function, and Transformation* (Cambridge: Cambridge Univ. Press, 1971); and Garland E. Allen, *Life Science in the Twentieth Century* (Cambridge/London: Cambridge Univ. Press, 1978). These volumes still significantly influence contemporary historiography; see, e.g., Peter J. Bowler and Iwan Rhys Morus, *Making Modern Science: A Historical Survey* (Chicago: Univ. Chicago Press, 2005). Some authors have taken a more nuanced view, however. Lynn K. Nyhart, for example, has claimed that natural history was declining relatively and growing absolutely around 1900, owing to the general expansion of biology's territory; see Lynn K. Nyhart, "Natural History and the 'New' Biology," in *Cultures of Natural History*, ed. Nicholas Jardine, James A. Secord, and Emma C. Spary (London: Cambridge Univ. Press, 1996), pp. 426–443, esp. p. 422. For Keith R. Benson, natural history remained "alive and well, primarily within museums"; see Keith R. Benson, "From Museum Research to Laboratory Research: The Transformation of Natural History into Academic Biology," in *The American Development of Biology*, ed. Ronald Rainger, Benson, and Jane Maienschein (Philadelphia: Univ. Pennsylvania Press, 1988), pp. 49–83, on p. 77. Paul Farber, while endorsing the narrative of the opposition between natural history and experimentation, noted pointedly that, from an intellectual point of view, the experimental approach (physiology) and natural history "did not have to be competitors"; see Paul Lawrence Farber, *Finding Order in Nature: The Naturalist Tradition from Linnaeus to E. O. Wilson* (Baltimore/London: Johns Hopkins Univ. Press, 2000), p. 80. The most direct challenge to this vision comes from Robert E. Kohler, *Landscapes and Labscapes: Exploring the Lab–Field Border in Biology* (Chicago: Univ. Chicago Press, 2002); Joel B. Hagen, "Experimental Taxonomy, 1920–1950: The Impact of Cytology, Ecology, and Genetics on the Ideas of Biological Classification" (Ph.D. diss., Oregon State Univ., 1984); and Hagen, "Naturalists, Molecular Biology, and the Challenge of Molecular Evolution," *Journal of the History of Biology*, 1999, *32*:321–341.

[4] See, e.g., the evolutionary biologist Edward O. Wilson's description of the "molecular wars" at Harvard: Edward O. Wilson, *Naturalist* (Washington, D.C.: Island, 1994), Ch. 12.

[5] On the role of protein collections in molecular biology see Soraya de Chadarevian, "Following Molecules: Haemoglobin between the Clinic and the Laboratory," in *Molecularizing Biology and Medicine: New Practices and Alliances, 1910s–1970s*, ed. de Chadarevian and Harmke Kamminga (Amsterdam: Harwood, 1998), pp. 171–201. On collecting, natural history, and contemporary bioprospecting see Bronwyn Parry, *Trading the Genome: Investigating the Commodification of Bio-Information* (New York: Columbia Univ. Press, 2004).

[6] On "experimental history" see Ursula Klein, "Experiments at the Intersection of Experimental History,

beginning of the twenty-first century. It has also led to a profound historical transformation in the life sciences—the development of public access and data-sharing policies—that has received little attention compared with that paid to the concomitant rise of intellectual property rights.[7]

By "natural history," I do not refer to the study of whole organisms—a recent meaning of the term—but to the different practices of collecting, describing, naming, comparing, and organizing natural objects, practices usually associated not with the laboratory but with the wonder cabinet, the botanical garden, or the natural history museum. Indeed, if there is any feature distinctive to the natural history approach, it is its reliance on collections, which have played a crucial role in natural history from the early modern period to the late nineteenth century, when Victorian sensibilities brought such collections to widespread popularity.[8] In addition to being tools for display, they were often means for producing knowledge about the taxonomies of living organisms, their anatomies, and their histories. Bringing specimens together in a single place and organizing them in a systematic way made comparisons among them possible and, by analogical reasoning, facilitated their identification and inscription into broader theoretical systems. Those who created the early modern cabinets of curiosity, the royal gardens of the seventeenth and eighteenth centuries, and the great zoological museums of the nineteenth and twentieth centuries all faced the challenges of bringing to a central location specimens that were often dispersed all over the world, securing the participation of individual naturalists, and negotiating the status of the specimens in the collection. The contemporary collectors of experimental knowledge whom I will examine here have been confronted with similar challenges, but in a very different context. They were operating in a community not of naturalists or *savants*, but of professional experimentalists who had widely different ideas about the epistemic value of collections, the ownership of knowledge, and, more generally, the moral economy of science.

This essay focuses on the creation of one of the most widely used types of collections in contemporary biomedical sciences: the molecular sequence database.[9] It traces the establishment in 1982 of GenBank, the largest and most frequently accessed collection of

Technological Inquiry, and Conceptually Driven Analysis: A Case Study from Early Nineteenth-Century France," *Perspectives on Science*, 2005, *13*:1–48. Pickstone's conceptual categories can help us solve the apparent contradiction resulting from the coexistence of experimental and natural historical approaches in the contemporary life sciences. Whereas we have been accustomed to think about the development of science in terms of a succession of episodes, such as Kuhnian paradigms replacing each other, Pickstone's "ways of knowing" are conceived as different epistemic layers coexisting simultaneously in the makeup of science.

[7] The literature on this topic is abundant; for an excellent overview see Daniel J. Kevles, *History of Patenting Life in the U.S. with Comparative Attention to Europe and Canada* (New York: Diane, 2000).

[8] On natural history as "organismic biology" see Erika Lorraine Milam, "The Equally Wonderful Field: Ernst Mayr and Organismic Biology," *Historical Studies in the Natural Sciences*, 2010, *40*:279–317. On natural history more generally see Farber, *Finding Order in Nature* (cit. n. 3); Michael T. Ghiselin and Alan E. Leviton, *Cultures and Institutions of Natural History: Essays in the History and Philosophy of Science* (Los Angeles: California Academy of Sciences, 2000); and Jardine *et al.*, eds., *Cultures of Natural History* (cit. n. 3) (for an excellent overview of themes in the history of natural history up to around 1900).

[9] A number of biological collections were founded in the twentieth century and have played a central role in the rise of the experimental life sciences. The American Type Culture Collection, for example, founded by Charles E. A. Winslow in 1911 at the American Museum of Natural History, is today the largest repository of microorganisms; the Cambridge Structural Database, founded by Olga Kennard, a student of James D. Bernal, at the University of Cambridge in 1965, contains structural information about small molecules, derived from X-ray diffraction experiments; and Victor McKusick's Mendelian Inheritance in Man, a database of hereditary diseases, was founded at the Johns Hopkins Medical School in 1966. For a participant's overview of the rise of genetic databases, including GenBank, see Temple E. Smith, "The History of the Genetic Sequence Databases," *Genomics*, 1990, *6*:701–707.

experimental knowledge in the world. GenBank, a public database of nucleic acid sequences, now contains many more nucleotides than "the number of stars in the Milky Way," as the National Institutes of Health (NIH) once put it in a press release (more than 200 billion nucleotides; just 100 billion stars).[10] The creation of GenBank, like that of the heavens, was no small achievement, and it similarly represented a significant historical turning point. The debates leading to its creation—about the collection and distribution of data, the attribution of credit and authorship, and the proprietary nature of knowledge—illuminate the challenges of making a natural historical practice compatible with the moral economy of the experimental sciences in the late twentieth century.

Databases, like earlier natural history collections, are not mere repositories; they are tools for producing knowledge. Researchers routinely compare the sequences they have determined in their laboratories with those in the database, using sophisticated software to infer by analogy the function of genes or the evolutionary relationships between species. In 2011 the sequences collected in GenBank, representing more than three hundred thousand different species, had been provided by tens of thousands of researchers.[11] Each day similar numbers of individuals from around the world access GenBank from their computers. Indeed, the material culture of this way of knowing depends on computers and computer networks. The results of the Human Genome Project, for example, were made available online on a daily basis through GenBank. Today, *in silico* biology complements *in vivo* and *in vitro* approaches, and it is vital to the success of the experimental enterprise. The science of bioinformatics has blossomed on exactly this premise.

Even though the emergence of the digital sequence database was closely interwoven with the computer revolution, the most significant challenge for its establishment was not technological but, rather, social and cultural. Indeed, as I will try to show, a number of tensions between the collecting and the experimental enterprises resulted from a clash of what E. P. Thompson has called "moral economies."[12] Those who contemplated the possibility of establishing large collections of experimental knowledge were confronted with changing understandings about issues of credit attribution, data access, and knowledge ownership. The resulting tensions reveal some of the essential features of the moral economies of contemporary experimental life sciences. They also bring into perspective the recent rise of public access publishing and data sharing in science.

[10] NIH press release, "Public Collections of DNA and RNA Sequence Reach 100 Gigabases," 22 Aug. 2005, available from http://www.nlm.nih.gov/news/press_releases/dna_rna_100_gig.html (acessed 1 Mar. 2009). The GenBank database contains the same data as its two partners, the EMBL Nucleotide Sequence Database and the DNA Data Bank of Japan (DDBJ).

[11] From an ENTREZ search on www.ncbi.nlm.nih.gov (accessed 1 Mar. 2009).

[12] The notion of "moral economy" has been popularized by the social historian E. P. Thompson as an alternative to economic and mob psychology explanations of peasant food riots in eighteenth-century England; see Edward Palmer Thompson, *The Making of the English Working Class* (New York: Vintage, 1963). He argued that the riots were driven not just by unfocused anger but by a sentiment of injustice and betrayal of a system of moral norms defining "just price" and exchange and the distribution of resources. The notion of moral economy has been imported into science studies and used in a variety of ways, most fruitfully in Robert E. Kohler, *Lords of the Fly:* Drosophila *Genetics and the Experimental Life* (Chicago: Univ. Chicago Press, 1994). Here, in a manner close to Thompson's and Kohler's usage, I will define it as the system of values that underlies the exchange of scientific knowledge, with particular regard to how knowledge is tied to issues of property, privacy, and priority. It is essential to remember that moral economies, unlike Mertonian norms, are locally and historically situated and thus can differ between scientific communities—in our case, between experimentalists and collectors of experimental data.

INFORMATION OVERLOAD ON THE HORIZON

In the natural history tradition, collections have been created for a number of different reasons, but one of them has been particularly enduring: the reaction to a perceived "information overload." In the sixteenth century, for example, the expansion of European travel led to the accumulation throughout the continent of previously unknown specimens and to the rise of natural history collections.[13] Collections were a practical means to bring order to a burgeoning diversity of natural forms. They made possible the immediate comparison of widely different organisms for the purpose of identifying individual specimens, producing general knowledge about organisms, or even ultimately making sense of the Creator's plan.[14] Finally, collections were often created by patrons or nation-states as displays of power and wealth; early modern wonder cabinets and nineteenth-century natural history museums attest to this clearly.[15]

The impetus for the creation of GenBank in 1982 was parallel to that for the founding of so many natural history collections. It was a reaction to a perceived "information overload," augmented by a new recognition of the scientific promise of the knowledge such a database would contain and the potential for individual and institutional prestige that would accompany its development. In the preceding decade, a number of key scientific and technological developments had radically transformed the intellectual landscape in the field of DNA sequences. But before DNA sequences became the main focus of attention, protein sequences held center stage. In 1953, after several years of painstaking effort, the British biochemist Frederick Sanger, working at the University of Cambridge, had succeeded in determining the first sequence of a protein, insulin, an achievement for which he was awarded his first Nobel Prize five years later.[16] In the following decade, the number of known protein sequences grew slowly—until the development of the automatic sequencer in 1967 enabled the number of known sequences to increase "explosively," as scientists frequently observed. By the end of the decade, that number reached into the hundreds.[17]

Sequencing long stretches of DNA, on the other hand, remained technically impossible

[13] On the early modern "information overload" see Daniel Rosenberg, "Early Modern Information Overload," *Journal of the History of Ideas*, 2003, *64*:1–9; Brian W. Ogilvie, "The Many Books of Nature: Renaissance Naturalists and Information Overload," *ibid.*, pp. 29–40; and Ogilvie, *The Science of Describing: Natural History in Renaissance Europe* (Chicago: Univ. Chicago Press, 2006).

[14] Collections were key tools not only for systematics but also, e.g., for studies in anatomy and evolution. The case of comparative anatomy is particularly illuminating. See, e.g., Richard W. Burkhardt, Jr., "The Leopard in the Garden: Life in Close Quarters at the Museum d'Histoire Naturelle," *Isis*, 2007, *98*:675–694; and Toby A. Appel, *The Cuvier–Geoffroy Debate: French Biology in the Decades before Darwin* (New York/Oxford: Oxford Univ. Press, 1987).

[15] On early modern collections see Paula Findlen, *Possessing Nature: Museums, Collecting, and Scientific Culture in Early Modern Italy* (Berkeley: Univ. California Press, 1994).

[16] On Sanger's sequencing work see Frederick Sanger, "Sequences, Sequences, and Sequences," *Annual Review of Biochemistry*, 1988, *57*:1–28; Soraya de Chadarevian, "Sequences, Conformation, Information: Biochemists and Molecular Biologists in the 1950s," *J. Hist. Biol.*, 1996, *29*:361–386; and Miguel Garcia-Sancho, "A New Insight into Sanger's Development of Sequencing: From Proteins to DNA, 1943–1977," *ibid.*, 2010, *43*:265–323.

[17] For one reference to the "explosive" growth of known sequences see M. O. Dayhoff to C. Berkley, 27 Feb. 1967, Archives of the National Biomedical Research Foundation, Georgetown, Washington, D.C. (hereafter cited as **NBRF Archives**). The archives, currently unsorted, are being processed at the National Library of Medicine; no further location information can be provided. For known protein sequences at the end of the decade see Margaret O. Dayhoff, *Atlas of Protein Sequence and Structure* (Silver Spring, Md.: National Biomedical Research Foundation, 1969).

until 1975.[18] That year, Sanger devised a method that made DNA sequencing relatively easy; two years later, the American molecular biologists Allan M. Maxam and Walter Gilbert at Harvard devised a second such method (Sanger and Gilbert received the Nobel Prize for their sequencing methods in 1980).[19] As a result, the number of known DNA sequences began to climb exponentially, leading to the feeling among molecular biologists that they would soon be overwhelmed by new DNA sequence data. In 1976 fewer than ten papers reporting nucleic acid sequences were published; in 1979 there were more than a hundred.[20] The bulk of known sequences began to shift from proteins to DNA, and it seemed clear that the number of DNA sequences would continue to grow at an increasing rate. One contemporary observer was particularly struck by the exponential rise in sequence data: the historian of science Derek J. de Solla Price. His 1963 *Little Science, Big Science* built on the observation that scientific knowledge, as measured by the number of published papers, was growing exponentially. So when he read in *Science* that DNA sequences were accumulating at a rate of 15 percent per month—more than any of his earlier estimates—he explored the matter further with one of the sequence data collectors, who acknowledged that this rise was indeed "extraordinary in the history of science."[21]

The significance of molecular sequences had also undergone a radical transformation in the 1970s. Originally, they were themselves considered objects of scientific interest, and their determination represented demonstrations of experimental virtuosity. However, as sequencing methods became increasingly automated, the sequences came to be considered highly prized pieces of data, used to draw new biological conclusions or new hypotheses that would then be explored experimentally.

The greatest excitement about DNA sequences focused on the structure and function of genes. Whereas the function of a protein was always known before its sequence was determined, the new methods for sequencing DNA produced vast amounts of data that at first seemed meaningless. However, if the sequence of a DNA fragment could be matched against another sequence—from another organism, for example—it could be inferred that the two sequences probably had similar functions, provided that they were of common evolutionary origin (homologous). The first result of such an approach, indicating that the DNA sequences of two virus proteins were similar, was published in 1978. Furthermore, comparisons between numerous sequences could show the presence of some common pattern, suggesting that it might have a functional role. The discovery in 1977 that genes were often composed of subunits ("introns" and "exons") and surrounded by several

[18] RNA sequences were first determined experimentally in 1965, although the process was slow. For the first RNA sequence see R. W. Holley, J. Apgar, G. A. Everett, J. T. Madison, M. Marquisee, S. H. Merrill, J. R. Penswick, and A. Zamir, "Structure of a Ribonucleic Acid," *Science*, 1965, *147*:1462–1465.

[19] Frederick Sanger and Alan R. Coulson, "A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase," *Journal of Molecular Biology*, 1975, *94*:441–448; and Sanger, Steve Nicklen, and Coulson, "DNA Sequencing with Chain-Terminating Inhibitors," *Proceedings of the National Academy of Sciences, USA*, 1977, *74*:5463–5467. On Sanger's sequencing methods see Garcia-Sancho, "New Insight into Sanger's Development of Sequencing" (cit. n. 16).

[20] On the increase in the number of papers reporting nucleic acid sequences between 1976 and 1979 see "Sequences Add Up," *Nature*, 1982, *297*:96. At the 1979 meeting convened to discuss a centralized sequence database, the "increasing rate at which nucleic acid sequence information is becoming available" was cited as the first reason for the need to create such a resource; see C. W. Anderson to H. Lewis, 14 Nov. 1980, Appendix II, NBRF Archives (cover letter for the minutes of the 1979 Rockefeller University meeting, which included a variety of documents).

[21] D. de Solla Price to Dayhoff, 11 Sept. 1980; and Dayhoff to Price, 17 Sept. 1980 (quotation): NBRF Archives. For the book see Derek J. de Solla Price, *Little Science, Big Science* (New York: Columbia Univ. Press, 1963).

functional elements, such as "TATA boxes," also raised much interest in the analysis and comparison of large numbers of DNA sequences.[22]

In short, a comprehensive database of DNA sequences seemed indispensable for making sense of the abundant new data that was being produced. As two molecular biologists would put it soon after, "the rate limiting step in the process of nucleic acid sequencing is now shifting from data acquisition towards the organization and analysis of that data."[23] These achievements and concerns converged in March 1979 in a crucial meeting at the Rockefeller University in New York City, which resulted in the first call from the scientific community for the creation of a centralized sequence database.

This meeting was convened by the molecular biologists Carl W. Anderson, Robert Pollack, and Norton Zinder to "discuss ways to collect, verify and make available to the world wide scientific community nucleic acid sequence information."[24] The organizers explained the necessity of such a gathering by pointing to the "rapidly increasing rate" of DNA sequences and the "wide range of biological questions that can be asked using a sequence data base."[25] Attendees included representatives from the European Molecular Biology Laboratory (EMBL), the National Institutes of Health Division of Research Resources (DRR), and the National Science Foundation (NSF), which sponsored the meeting. Among the participants were more than thirty scientists with special expertise in the field of computers applied to the life sciences, in the management of biomedical databases, or in molecular biology.[26]

A review of some of the meeting participants indicates not only the fields represented but also the tools and resources then available to further the organizers' aims. Among those with experience using computers in the life sciences, Joshua Lederberg, the Nobel Prize–winning molecular biologist who, as president of the Rockefeller University, opened the meeting, was best known for his discovery of bacterial sex. However, he had also vigorously promoted the use of computers and artificial intelligence in the biomedical sciences since the 1960s at Stanford, where he had founded the shared computer resource SUMEX-AIM. Another participant, the chemist and computer scientist Howard S. Bilofsky, from Bolt, Beranek, and Newman (BBN), the company that had developed the ARPANET for the Department of Defense in 1969, was working for the PROPHET

[22] For the first result see Theodore Friedmann, Russell F. Doolittle, and Gernot Walter, "Amino-Acid Sequence Homology between Polyoma and SV40 Tumor Antigens Deduced from Nucleotide-Sequences," *Nature*, 1978, *274*:291–293. Through the DNA-hybridization method, DNA sequences could be compared indirectly before sequencing techniques were available. On the discovery of gene subunits and their consequences see Michel Morange, *A History of Molecular Biology* (Cambridge, Mass.: Harvard Univ. Press, 2000), Ch. 17.

[23] Thomas R. Gingeras and Richard J. Roberts, "Steps toward Computer Analysis of Nucleotide Sequences," *Science*, 1980, *209*:1322–1328.

[24] Anderson to Dayhoff, 9 Jan. 1979, NBRF Archives. The meeting also had a more local agenda—namely, to assess the possibility of establishing a "centralized computer facility" to collect and analyze nucleic acid sequences at the Rockefeller University. See "Report to the National Science Foundation," attached to Anderson to Lewis, 14 Nov. 1980, NBRF Archives; and Bruno J. Strasser interview with Norton Zinder, Rockefeller University, 10 Feb. 2006.

[25] Anderson to Dayhoff, 9 Jan. 1979. The terms "data base," "data bank," and "data library" were often used interchangeably by the historical actors.

[26] C. W. Anderson, "Report to the National Science Foundation," 14 Nov. 1980, Appendix II, NBRF Archives. The participants were C. W. Anderson, H. Bilofsky, M. Billeter, F. Blattner, M. O. Dayhoff, G. Edelman, B. Erickson, R. J. Feldmann, W. Fitch, P. Freidland, T. Gingeras, J. S. Haemer, J. Hahn, C. Hutchinson, E. Kabat, L. Kedes, O. Kennard, L. Korn, J. Lederberg, C. Levinthal, H. Lewis, J. Maizel, A. M. Maxam, J. Milazzo, J. Pasta, G. Pieczenik, C. Queen, R. J. Roberts, T. Smith, R. Sommer, C. Squires, R. Staden, J. Vournakis, M. Waterman, and S. M. Weissman.

project, another shared computer resource for pharmacologists that BBN had established in 1973. Finally, the mathematician Michael S. Waterman and the physicist Temple F. Smith were developing algorithms to analyze sequence data at Los Alamos Scientific Laboratory.[27] In the field of database management, the physical chemist Margaret O. Dayhoff had been expanding a computerized collection of protein sequences (published as *Atlas of Protein Sequence and Structure*) at the National Biomedical Research Foundation (NBRF) in Washington, D.C., since 1965, and the biochemist Elvin A. Kabat had assembled his own specialized collection of immunoglobulin sequences at the NIH in Bethesda, Maryland, collaborating with Bilofsky and using his computer system. The crystallographer Olga Kennard was maintaining the Cambridge Crystallographic Data Center she had founded in 1965 to collect and distribute structural data on small organic molecules; and—though he was not directly involved—Carl W. Anderson, from the Brookhaven National Laboratory, was well aware of the progress of the Protein Data Bank hosted there, which had been collecting and distributing the atomic coordinates of protein structures since 1973. Molecular biologists in attendance included such luminaries as Walter Gilbert, Richard J. Roberts, and Sydney Brenner.

In addition to heated discussion about the opportunity of launching a DNA database, the participants engaged in practical demonstrations of how computers could be used for a future database. Dayhoff, for example, showed how her sequence database, located at Georgetown University, could be accessed remotely; another participant demonstrated access to the SUMEX-AIM computer facility at Stanford University; and a third showed how sequences could be compared using an "inexpensive 'personal' computer produced by Radio Shack." These technical possibilities were new to many of the experimental biologists present at the meeting, who were more familiar with wet laboratory instruments—electrophoresis apparatus and ultracentrifuges—than with computers and networks. At the end of the meeting, the participants concluded that a "centralized data bank" of nucleic acid sequences was "highly desirable and essential for the organized and efficient use of nucleic acid sequence information."[28]

Behind the apparent agreement among the participants, however, a number of concerns remained unresolved. First, some researchers worried that a single centralized facility would jeopardize the collecting efforts of individual laboratories. Whereas physicists had long been familiar with the centralized facilities intrinsic to postwar big science, biologists were often reluctant to imitate them, taking pride in the smaller scale of their laboratories. As two physicists pointed out shortly afterward: "Now, molecular biology is 'little science' *par excellence*, practiced with relatively little apparatus—by the standards of physics." It was no accident that the Protein Data Bank, for example, was hosted at Brookhaven National Laboratory, an institution devoted to physical research, rather than at some academic biomedical laboratory. Second, the participants wondered how the privacy of preliminary data could be maintained in a publicly accessible database. The

[27] On the PROPHET project see Paul A. Castleman *et al*., "The Implementation of the Prophet System," *AFIPS Conference Proceedings*, 1974, *43*:457–468. Waterman and Smith's most notable contribution would be an algorithm for local sequence alignment: Temple F. Smith and Michael Waterman, "Identification of Common Molecular Subsequences," *J. Molec. Biol*., 1981, *147*:195–197. For their collaborations prior to the Rockefeller meeting see Waterman, Smith, M. Singh, and W. A. Beyer, "Additive Evolutionary Trees," *Journal of Theoretical Biology*, 1977, *64*:199–213; and Waterman and Smith, "On the Similarity of Dendrograms," *ibid*., 1978, *73*:789–800. Regarding Waterman's work more generally see Waterman, *Skiing the Sun* (2007), p. 13 (this is a pdf pamphlet available on Waterman's homepage: www.cmb.usc.edu/people/msw/ [accessed 1 Mar. 2009]). Los Alamos Scientific Laboratory was renamed Los Alamos National Laboratory (LANL) in 1981.

[28] C. W. Anderson, "Report to the National Science Foundation," 14 Nov. 1980, NBRF Archives, pp. 2, 3.

issue was how to protect the priority claims of those who had determined sequences. Third, they wondered how to make the content of the database available on an equitable basis, without giving an unfair advantage to the host laboratory.[29] Indeed, if it was located in a research laboratory rather than a service company, the hosts might be tempted to exploit the content of the database before it was made publicly available.

This point was forcefully made by Olga Kennard, who was maintaining the Cambridge Crystallographic Data Center. Together with Dayhoff, she had the most experience in data collecting and was thus speaking authoritatively when she pointed out that in order to gain the "interest and confidence of the scientific community," which was essential for the success of data collection, the database organizers themselves must be well-recognized figures in that community. But at the same time, in order to allay any fears as to whether the organizers might appropriate the content of the database for themselves, it would be crucial that "every assurance" be given that that content would be "distributed world wide" and at "a minimum cost" for individuals.[30]

Kennard's perceptive analysis pointed to an essential contradiction in the requirements for a sequence database: the collector had to be a recognized figure in the field of DNA sequences yet not display any personal interest in the data it contained. Most great natural history collectors of the past, such as Joseph Hooker at Kew Gardens, Augustin Pyramus de Candolle at the Geneva Botanical Garden, or George Gaylord Simpson at the American Museum of Natural History, had been keenly interested in the items they had collected and did not practice the separation of collection and study that Kennard saw as necessary. Taken together, these concerns indicated that, as much as the participants favored collaboration, preserving individual interests remained a key issue. The moral tensions between different conceptions of credit attribution, data access, and knowledge ownership structured the debates on the establishment of a centralized database. More than the legal forms of intellectual property such as patents and copyrights, or the related commercialization of knowledge, it was the "informal" modes of appropriation that were the major preoccupation of the participants.

The impact of the Rockefeller workshop was multifaceted, but above all it made it clear that there was a strong desire in the scientific community for a single computerized and nonproprietary database.[31] Two institutions were particularly well positioned to take the lead in developing such a facility in the United States: the National Biomedical Research

[29] Regarding the first concern (and for the quotation referring to molecular biology as "little science") see G. I. Bell and W. Goad to R. Ewald, 4 Dec. 1980, Water Goad Papers, American Philosophical Society, Philadelphia (hereafter cited as **APS Archives**), Ms. Coll. 114, Series III. For the second see "Report to the National Science Foundation," attached to Anderson to Lewis, 14 Nov. 1980, NBRF Archives, p. 2. For the third see *ibid*. and Appendix II, p. 5.

[30] Olga Kennard, "Notes on the Preliminary Report and Recommendations of the Workshop on Computer Facilities for the Analysis of Protein and Nucleic Acid Sequence Information," attached to Anderson to Lewis, 14 Nov. 1980, NBRF Archives, p. 1.

[31] The minutes of the Rockefeller meeting were not released until November 1980—i.e., almost two years after the meeting took place. Temple F. Smith has argued that this delay prevented the NIH from knowing about the conclusions of the Rockefeller meeting and perhaps delayed the development of the database project within the NIH. See Smith, "History of the Genetic Sequence Databases" (cit. n. 9). However, it seems unlikely that the NIH, and the National Institute of General Medical Sciences (NIGMS) in particular, was unaware of the conclusions of the meeting before the minutes were released. Indeed, a member of the NIH's DRR was present, and a number of participants, including Richard J. Roberts, were in close contact with the directorship of the NIGMS and most likely made the conclusions known: Bruno J. Strasser interview with Ruth L. Kirschstein, Bethesda, Md., 22 Feb. 2006.

Foundation and the Los Alamos Scientific Laboratory.[32] The diverse natures of these institutions, the very different personalities they hosted—Margaret O. Dayhoff at the NBRF and Walter B. Goad at Los Alamos—and their various research trajectories at the interface of computers and biology resulted in contrasting perspectives on the collection of biological data. Even though none of the key actors had any significant connection with the natural history enterprise, a comparison of their efforts at collecting sequences with those of other collecting enterprises in the natural history tradition is enlightening, as it reveals the reliance of all these undertakings on similar strategies.

MARGARET O. DAYHOFF AND HER CHALLENGER

Margaret O. Dayhoff (1925–1983) was by far the most experienced researcher in the field of sequence databases who attended the Rockefeller meeting. Dayhoff received a Ph.D. in quantum chemistry in 1948 under George E. Kimball at Columbia University, after having taken a B.A. in mathematics and an M.A. in chemistry (see Figure 1).[33] While a fellow at the IBM Watson Laboratory in 1947–1948, she used punched-card machines to calculate resonance energies in small molecules. After obtaining her Ph.D. she worked on problems of theoretical chemistry, first as a research assistant at the Rockefeller Institute and then at the University of Maryland. She joined the National Biomedical Research Foundation in 1960 and eventually became a professor of physiology and biophysics at Georgetown University and president of the Biophysical Society (1980–1981).[34]

The NBRF was a unique environment where computers and biology were brought into close proximity. Just outside of Washington, D.C., this private nonprofit institution had been founded by Robert S. Ledley in 1960 to explore the possible uses of electronic computers in biomedical research. Ledley, born in 1926, was trained as a dentist before obtaining an M.A. in theoretical physics from Columbia University and becoming interested in digital computers. In 1965 he published a nine-hundred-page monograph entitled *Uses of Computers in Biology and Medicine*.[35] It constituted the first introduction to the principles and methods of digital computing and their potential applications in biology and medicine. The publication of this book was only one example of Ledley's life-long commitment to promoting the use of digital computers in biomedicine, from the auto-

[32] Stanford University was considered an even more promising candidate than Los Alamos. However, for the sake of brevity, it will not be discussed here.

[33] M. O. Dayhoff, "Biographical Sketch: Margaret Oakley Dayhoff," 1965, NBRF Archives. On Dayhoff's background see Bruno J. Strasser, "Collecting and Experimenting: The Moral Economies of Biological Research, 1960s–1980s," *Preprints of the Max-Planck Institute for the History of Science*, 2006, *310*:105–123; and Strasser, "Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's *Atlas of Protein Sequence and Structure*, 1954–1965," *J. Hist. Biol.*, 2010, *43*:623–660. Published biographical sketches include Lois Hunt, "Margaret Oakley Dayhoff, 1925–1983," *Bulletin of Mathematical Biology*, 1984, *46*:467–472; Hunt, "Margaret O. Dayhoff, 1925–1983," *DNA*, 1983, *2*:97–98; and John R. Jungck, "Margaret Oakley Dayhoff: Harnessing the Computer Revolution," *American Biology Teacher*, 1985, *47*:9–10.

[34] For her work at the IBM Watson Laboratory see Margaret B. Oakley and George E. Kimball, "Punched Card Calculation of Resonance Energies," *Journal of Chemical Physics*, 1949, *17*:706–717. Isaac Asimov, who would also develop a keen interest in molecular biology, was a fellow the same year as Margaret O. Dayhoff (then Oakley). Regarding her move to the NBRF see R. S. Ledley, "Memorandum," 16 Nov. 1960, NBRF Archives.

[35] Robert S. Ledley, *Uses of Computers in Biology and Medicine* (New York: McGraw-Hill, 1965). On the founding of the NBRF see R. S. Ledley to Harvey E. Saveley, 29 June 1960, NBRF Archives. The NBRF eventually moved to Georgetown University Medical Center in Washington, D.C. On the early history of the NBRF, and on Ledley and Dayhoff's work with sequences, see Strasser, "Collecting, Comparing, and Computing Sequences" (cit. n. 33).

Figure 1. *Margaret O. Dayhoff (second from left), behind Robert S. Ledley, at the inauguration of their new computer, a DEC VAX-11/780, May 1979. Reproduced with permission from the National Biomedical Research Foundation Archives.*

mated recognition of chromosome images to computer-assisted medical diagnostics and the analysis of molecular sequences.[36]

On joining the NBRF in 1960, Dayhoff began to develop computer algorithms to assist biochemists in their sequencing efforts (see Figure 2).[37] Following Emile Zuckerkandl and Linus Pauling's initial insights into molecular evolution, published in 1962, she sought to use computers to compare sequences and construct evolutionary trees. As she later explained to a colleague: "There is a tremendous amount of information regarding evolutionary history and biochemical function implicit in each sequence and the number of known sequences is growing explosively. We feel it is important to collect this significant information, correlate it into a unified whole and interpret it." In order to carry out this project, Dayhoff and her—mainly female—collaborators began to assemble a database, compiling all known protein sequences.[38]

[36] On Ledley's transition to protein sequence studies see Strasser, "Collecting, Comparing, and Computing Sequences." On the introduction of computers in biology and medicine see Joseph A. November, "Digitalizing Life: The Introduction of Computers to Biology and Medicine" (Ph.D. diss., Princeton Univ., 2006); and Joel B. Hagen, "The Origins of Bioinformatics," *Nature Reviews*, 2000, *1*:231–236.

[37] Margaret O. Dayhoff and Robert S. Ledley, "Comprotein: A Computer Program to Aid Primary Protein Structure Determination," in *Proceedings of the Fall Joint Computer Conference* (Santa Monica, Calif.: American Federation of Information Processing Societies, 1962), pp. 262–274; and Dayhoff, "Computer Aids to Protein Sequence Determination," *J. Theoret. Biol.*, 1965, *8*:97–112. On Ledley's earlier attempts at the problem see Strasser, "Collecting, Comparing, and Computing Sequences."

[38] Dayhoff to Berkley, 27 Feb. 1967, NBRF Archives. On the uses of computers in molecular evolution and molecular systematics see Joel B. Hagen, "The Introduction of Computers into Systematic Research in the United States during the 1960s," *Studies in the History and Philosophy of Biological and Biomedical Sciences*, 2001, *32*:291–314. On Dayhoff's early use of computers and her work in molecular evolution see Strasser,

Figure 2. *Bead model of the amino acid sequence of the protein ribonuclease on computer listings, circa 1962. Reproduced with permission from the National Biomedical Research Foundation Archives.*

This simple task represented a considerable challenge because the published sequences were dispersed throughout a number of different journals and the word "sequence" was not yet indexed in bibliographic catalogues. The result of this collecting effort, which included about seventy protein sequences, was published in 1965 as a book entitled *Atlas of Protein Sequence and Structure*. Each page contained the sequence of a protein, its name and the organism it came from, its amino acid composition, and at least one reference to the literature where the sequence was first described. By the second edition, published only one year after the first, the *Atlas* had doubled in size. The first edition of the *Atlas* contained fewer than one hundred references to published sequences; the edition published seven years later included more than one thousand.[39] The number of papers containing descriptions of sequences was "exploding," as Dayhoff frequently put it, and collecting this overflow of information represented a growing workload for Dayhoff and her team.

Dayhoff had hoped to establish a system by which researchers who had determined sequences would share them voluntarily with her in a computer-readable format for inclusion in her database. In exchange, they would receive a free copy of the *Atlas*.[40] This gift economy, however, was largely unsuccessful. Even though her sequence collection

"Collecting, Comparing, and Computing Sequences." Dayhoff's collaborators included Richard V. Eck, the microbiologist Minnie R. Sochard (1931–), the applied mathematician Marie A. Chang (1937–), the biologist Lois D. Hunt, and four other women.

[39] Margaret O. Dayhoff *et al*., *Atlas of Protein Sequence and Structure* (Silver Spring, Md.: National Biomedical Research Foundation, 1965); Richard V. Eck and Dayhoff, *Atlas of Protein Sequence and Structure* (Silver Spring, Md.: National Biomedical Research Foundation, 1966); and Dayhoff, "LM 01206, Comprehensive Progress Report," 23 Aug. 1973, NBRF Archives.

[40] Eck and Dayhoff, *Atlas of Protein Sequence and Structure*, p. 1. Dayhoff wrote to a number of biochemists who were sequencing proteins to urge them to share their data: M. O. Dayhoff, "Staff Correspondence," 1965–1969, NBRF Archives.

was widely praised, the community of researchers did not collaborate with her as much as she had hoped, for reasons I will discuss later. She nevertheless succeeded in keeping up with the rapidly growing number of protein sequences by combing the scientific literature, and she published ever-thicker atlases. Beginning in 1972, she also distributed the data on magnetic tapes. However, those who acquired them had to agree not to redistribute the data, which was copyrighted. Reluctant NIH support and the rising costs of data collection led Dayhoff to seek revenue from sales of the collection in print and magnetic format. In 1977, for example, she sold the tape containing her database for $400. The same year, the Protein Data Bank sold the tape containing theirs for less than $35.[41] Even though the price charged by Dayhoff remained modest by all standards, it put her project on the side of commercial ventures rather than publicly available resources. The fact that she copyrighted the data included in the *Atlas*, and also resisted distributing it in a computer-readable format, reinforced that impression and irritated some users. One of them wrote to Dayhoff and asked rhetorically: "You are in somewhat the position of a folksong collector who copyrights his published material; do I have to pay him if I sing *John Henry*?"[42] Thus, long before the fierce debates on the patenting of genes and organisms of the 1980s, the appropriation of biological knowledge was already a subject of contention.

In addition to collecting sequences and maintaining a database, Dayhoff also analyzed data submitted to the *Atlas*. In 1966, for example, she discovered that the sequence of the ferredoxin protein contained an internal duplication, indicating how it had evolved.[43] She also developed different methods and computer programs to compare sequences and build phylogenetic trees.[44] The fact that Dayhoff published conclusions derived from the sequence data provided to the *Atlas* provoked "resentment within the scientific community," especially among those who had determined the sequences in their laboratories. As one researcher put it, Dayhoff seemed to consider that the sequences in her collection constituted her own intellectual "private hunting grounds."[45] But the molecular biologists and biochemists who had identified sequences, a painstaking effort of many months or even years, had a strong sense of ownership of their work and were not ready to give it up to a sequence collector to analyze.

In the natural history tradition, items such as specimens were generally owned by individual naturalists. Carl Linnaeus, for example, like so many botanists after him, had a private herbarium with specimens that had often been provided by distinguished colleagues or anonymous amateur naturalists. Collectors and, later, museums treated

[41] M. O. Dayhoff, "LM 01206, Comprehensive Progress Report," 23 Aug. 1973, NBRF Archives (nonredistribution agreement); Dayhoff to P. Edman, 14 Feb. 1977, NBRF Archives (sale of the database); and Frances C. Bernstein *et al.*, "Protein Data Bank: Computer-Based Archival File for Macromolecular Structures," *European Journal of Biochemistry*, 1977, *80*:319–324 (see p. 321 for the sale price).

[42] B. S. Guttman to Dayhoff, 10 June 1968, NBRF Archives.

[43] R. V. Eck and M. O. Dayhoff, "Evolution of the Structure of Ferredoxin Based on Living Relics of Primitive Amino Acid Sequences," *Science*, 1966, *152*:363–366. Two other groups made the same discovery simultaneously. See Russell F. Doolittle, S. J. Singer, and Henry Metzger, "Evolution of Immunoglobulin Polypeptide Chains: Carboxy-Terminal of an IgM Heavy Chain," *ibid.*, 1966, *154*:1561–1562; and Walter M. Fitch, "Evidence Suggesting a Partial, Internal Duplication in the Ancestral Gene for Heme-Containing Globins," *J. Molec. Biol.*, 1966, *16*:17–27.

[44] E.g., the Point Accepted Mutation matrix, or PAM matrix, became widely used by evolutionary biologists and was often referred to as the "Dayhoff matrix." See Joseph Felsenstein, *Inferring Phylogenies* (Sunderland, Mass.: Sinauer, 2004), Ch. 10.

[45] R. Holmquist to Dayhoff, 23 Dec. 1979, NBRF Archives; and W. Salser to Goad, 31 Dec. 1979, APS Archives. See also Russell F. Doolittle, "On the Trail of Protein Sequences," *Bioinformatics*, 2000, *16*:24–33; and Bruno J. Strasser interview with Temple S. Smith, Boston, 16 Feb. 2006.

specimens as their own and engaged in lending, trading, and even selling them to other collectors.[46] They were also free to produce systematic work—descriptions of higher taxa—based on specimens others had provided to their collections.[47] But what was perhaps an acceptable mode of interaction in the naturalist tradition was perceived by the experimentalist community of the twentieth century as an unacceptable transgression of its different moral economy.

At the time of the Rockefeller meeting, Dayhoff was managing the largest collection of protein sequences in the world, containing more than 100,000 amino acids. Her collection also included a small number of nucleic acid sequences, essentially transfer-RNA sequences, which had been included in the *Atlas* since 1966, and she was "deeply involved" in increasing the size of her DNA collection. In 1978 she had released her first computer tape exclusively devoted to nucleic acid sequences; it contained 24,000 nucleotide residues.[48]

Even though Dayhoff had pioneered some of the early methods for sequence comparison and for building phylogenetic trees, increasingly complex computational methods were being developed in various places, including the Los Alamos Scientific Laboratory in New Mexico.[49] Two researchers who were frequent visitors to Los Alamos, the mathematician Michael S. Waterman and the physicist Temple F. Smith, were present at the Rockefeller meeting and brought the news about a projected national database back to New Mexico. It struck a chord in the Theoretical Biology and Biophysics (T-10) group that George I. Bell, a physicist who converted to theoretical immunology, had created in 1974. Since the time of the Manhattan Project, Los Alamos had hosted a small research group devoted to medical aspects of radiation. Radiation genetics became the main focus of the biological research carried out at Los Alamos during the Cold War because of the controversy over the effects of fallout from atmospheric nuclear testing. A number of physicists and mathematicians, such as Stanislaw M. Ulam and Bell, who had been involved in the Manhattan Project and subsequent weapons projects, decided out of guilt, boredom, or curiosity to turn their minds to more peaceful ends. Radiation genetics, biophysics more generally, theoretical biology, and computational biology represented ideal venues for the exercise of their skills.[50] The efforts to build a DNA sequence

[46] On Linnaeus's herbarium see Staffan Müller-Wille, "Carl Von Linnés Herbarschrank: Zur epistemischen Funktion eines Sammlungsmöbels," in *Sammeln als Wissen: Das Sammeln und seine Wissenschaftsgeschichtliche Bedeutung*, ed. Anke te Heesen and Emma C. Spary (Göttingen: Wallstein, 2001), pp. 22–38; and Müller-Wille, "Linnaeus' Herbarium Cabinet: A Piece of Furniture and Its Function," *Endeavour*, 2006, *30*:60–64. On exchange practices among naturalists in the mid-twentieth century see Ernst Mayr, *Methods and Principles of Systematic Zoology* (New York: McGraw-Hill, 1953), Ch. 4. On specimens as commodities see, e.g., Mark Barrow, "The Specimen Dealer: Entrepreneurial Natural History in America's Gilded Age," *J. Hist. Biol.*, 2000, *33*:493–534; and Bettina Dietz, "Mobile Objects: The Space of Shells in Eighteenth-Century France," *British Journal for the History of Science*, 2006, *39*:363–382. On collection ownership see Samuel J. M. M. Alberti, "Objects and the Museum," *Isis*, 2005, *96*:559–571.

[47] See, e.g., Joseph Hooker's systematic work, which is treated in Jim Endersby, *Imperial Nature: Joseph Hooker and the Practices of Victorian Science* (Chicago: Univ. Chicago Press, 2008), Ch. 5; or the remark by Ernst Mayr that the systematicist should try to "collect at least part of his own material": Mayr, *Methods and Principles of Systematic Zoology*, p. 12.

[48] Dayhoff to Anderson, 25 Jan. 1979, NBRF Archives.

[49] Goad to D. Kerr, 19 Sept. 1979, APS Archives.

[50] On the T-10 group see "Group T-10: Theoretical Biology and Biophysics [report]," 1977, APS Archives. Regarding Los Alamos work on medical aspects of radiation see Peter J. Westwick, *The National Labs: Science in an American System, 1947–1974* (Cambridge, Mass.: Harvard Univ. Press, 2003), pp. 241–256. On the postwar "biophysics bubble" see Nicolas Rasmussen, "The Mid-Century Biophysics Bubble: Hiroshima and the Biological Revolution in America, Revisited," *History of Science*, 1997, *35*:245–293.

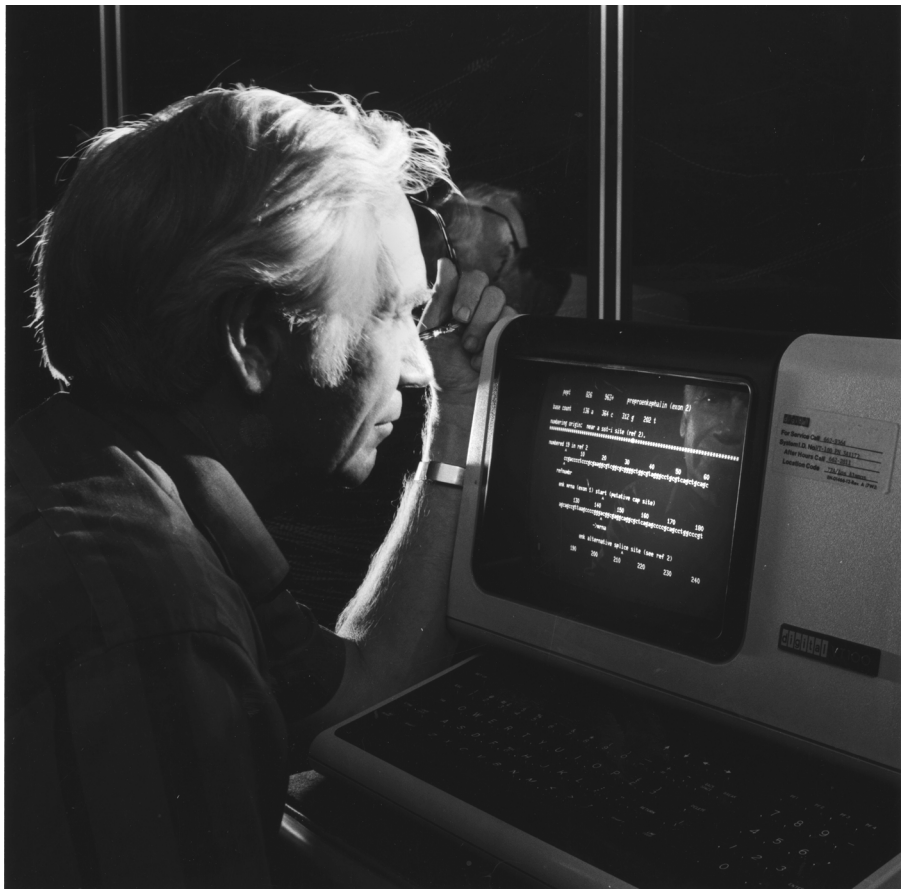

Figure 3. *Walter Goad in front of a terminal accessing the GenBank database, circa 1983 (Goad Papers). Reproduced with the permission of the American Philosophical Society Archives.*

database at Los Alamos were thus a direct result of the changing research agendas of the Cold War.

One of the members of the Los Alamos Theoretical Biology and Biophysics group, the theoretical physicist Walter B. Goad (1925–2000), became particularly interested in the prospects for the computerized sequence database outlined at the Rockefeller meeting (see Figure 3). He had received his Ph.D. in physics from Duke University in 1954; but he had been a member of the national laboratory, where he would spend his entire career, since 1950, and eventually he became associated with the team that developed the first thermonuclear bombs. In the 1960s he started to become interested in problems of theoretical biology, leading to year-long stays at the University of Colorado Medical Center and the Medical Research Council Laboratory of Molecular Biology in Cambridge, U.K. Since the creation of the Theoretical Biology and Biophysics group, Goad had devoted his entire time to biological problems. His biological research seemed to follow no clear direction, and he picked up new problems as they came along, sometimes applying the expertise in digital computers that he had gained while working on thermonuclear weapons. Unlike Dayhoff, Goad had no experience in collecting sequences; but when he heard about the

prospect of developing a national nucleic acid database he thought that Los Alamos was the right place to host it.[51]

EUROPE TAKES THE LEAD, AMERICA HASTENS TO RESPOND

The participants at the Rockefeller meeting recognized that the National Biomedical Research Foundation and Los Alamos Scientific Laboratory could each potentially host a nucleic acid database.[52] They also identified the European Molecular Biology Laboratory in Heidelberg, Germany, as a possible candidate, but they did not foresee that the EMBL would soon take the lead in creating a centralized nucleic acid database.

In Europe, as in the United States, several researchers had begun to collect nucleic acid sequences. For example, the molecular biologists Kurt Stüber at the University of Cologne and Richard Grantham at the University of Lyon had assembled small collections for their personal use. The EMBL, however, had greater plans. In January 1980 the European laboratory announced that it was hoping to collaborate with "whatever group in the USA" would become "responsible for computer storage and analysis of nucleic acid sequences."[53] Because the Rockefeller meeting ten months earlier hadn't been followed by any indication as to which institution would take the initiative in setting up a national facility, the EMBL decided to take the lead and convened its own meeting, on "Computing and DNA Sequences," near Heidelberg in April 1980. The goal of the meeting was to discuss the "use of the computer as an aid to sequence determination, . . . the utilization of data banks . . . and possible role for the EMBL in these matters." The agenda was thus very similar to that of the Rockefeller meeting and was likewise aimed at positioning its hosting institution with a view to the future development of a centralized facility. A large number of European researchers were in attendance, but the group also included several American scientists who had been present at the Rockefeller meeting. Like the Rockefeller gathering, the EMBL meeting ended with agreement that a sequence database should be centralized, that it should be computerized and available free of charge, and that it was urgently needed.[54]

Crucially, this time the results of the discussions of a small group of scientists were made public. The following week, *Nature* dedicated its main editorial to "Banking DNA Sequences." The author reflected on the recent increase in the number of sequences that were published and contemplated future "grandiose sequencing" projects, including the human genome. The editorial stressed that the need for a computerized DNA sequence databank that would make sequences "freely available" was "becoming urgent." "Although number, or rather letter, crunching is no substitute for thought," the author argued, computers would be an essential aid for a sequence database.[55] A consensus seemed to be

[51] "Group T-10: Theoretical Biology and Biophysics [report]," 1977, APS Archives. Goad's résumé is attached to BBN [Bolt, Beranek, and Newman], "Establishment of a Nucleic Acid Sequence Data Bank," Mar. 1982, NBRF Archives. For his own account of the creation of the database see Walter B. Goad, "Genbank," *Los Alamos Science*, 1983, *9*:52–63, esp. p. 55.

[52] C. W. Anderson, "Report to the National Science Foundation," 14 Nov. 1980, Appendix II, NBRF Archives.

[53] Goad to R. Roberts, 14 Jan. 1981 (on Stüber's and Grantham's collections); and Bell to Goad, 21 Jan. 1980 (quotation): APS Archives.

[54] "EMBL Workshop on Computing and DNA Sequences, 24th and 25th April 1980," Archives of the European Molecular Biology Institute, Hinxton, U.K. (hereafter cited as **EBI Archives**); F. Murray to Dayhoff, 19 Jan. 1980, NBRF Archives (quotation); and F. R. Blattner, "Report on EMBL Workshop on Computing & DNA Sequences," 24 June 1980, EBI Archives.

[55] "Banking DNA Sequences," *Nature*, 1980, *285*:59. The editorial overlooked the existence of Dayhoff's

emerging on both sides of the Atlantic as to the necessity of a computerized sequence database.

Only two months later, in June 1980, the EMBL announced that it had decided to make its nucleic acid database publicly available—a striking contrast to the slow pace at which the foundation for a national database was developing in the United States. The EMBL had been created in 1974 under the assumption that molecular biology, like high energy physics, would need expensive equipment that only an international laboratory, similar to CERN, could provide.[56] The prospect of hosting a nucleic acid sequence database on a centralized computer thus seemed perfectly in tune with this founding idea.

In the United States, Dayhoff, Goad, and their teams were preparing to compete for an eventual national contract for a DNA database. Only weeks after the Rockefeller meeting, Dayhoff had outlined a large-scale project to develop a nucleic acid sequence database and applied to the NIH for support. She put great emphasis on verifying the data for accuracy and on having the sequences "certified" by several experts, including the original authors. She argued that a carefully verified collection was "more economical in the long run than a 'quick and dirty' collection," a clear allusion to other sequence collectors who didn't put the same effort into verifying the data.[57]

Dayhoff simultaneously turned to NASA, a longtime sponsor of her activities, to seek funding for a "demonstration project" that would convince the NIH study committee to provide further support. This computerized database would be of crucial importance "to the NASA projects on the origins of life," Dayhoff argued. Indeed, her work on the evolution of proteins, such as ferredoxins or cytochromes, led her to infer ancestral sequences that could have been present in the first forms of life on earth. She also approached major biotech and pharmaceutical companies to support the development of her database. Claiming that access to the database would give these companies "a competitive advantage," she managed to elicit contributions of between $5,000 and $10,000 from Genex, Merck, Eli Lilly, DuPont, Hoffman–La Roche, and Upjohn, while Pfizer Medical Systems provided "computer time."[58]

On 15 September 1980, after publishing an announcement in *Science*, Dayhoff made her nucleic acid sequence database available for free over the telephone network. It comprised over 200,000 residues and was the largest sequence database worldwide, containing more than twice the amount of data in the second largest DNA sequence database, hosted at Los Alamos Scientific Laboratory.[59] Dayhoff's DNA sequence database was modeled after her protein sequence collection and included one sequence per entry, with annotations about its structure and function. The database was an immediate success, and in the first month of its operation more than a hundred scientists requested access. What Dayhoff did not mention in the published announcements was that in order to access the database the user had to request a password from the NBRF and sign an

database, and she wrote a letter to *Nature* to correct this omission: M. O. Dayhoff, R. M. Schwartz, H. R. Chen, L. T. Hunt, B. C. Orcutt, and W. C. Barker, "Banking DNA Sequences," *ibid.*, 1980, *286*:326.

[56] On the decision to open the EMBL database see K. Murray to Goad, 12 June 1980, APS Archives. On the assumptions behind the creation of the EMBL see John Krige, "The Birth of EMBO and the Difficult Road to EMBL," *Stud. Hist. Phil. Biol. Biomed. Sci.*, 2002, *33*:547–564.

[57] Dayhoff to E. Jordan, 13 Aug. 1980 (appeal to NIH); and M. O. Dayhoff *et al*., "Now Available over 150 Kilobases," 15 Aug. 1980 (quotation): NBRF Archives.

[58] Dayhoff to D. DeVincenzi, 20 Aug. 1980 (NASA request); and Dayhoff to various pharmaceutical and biotech companies, Aug.–Dec. 1980: NBRF Archives.

[59] Margaret O. Dayhoff *et al.*, "Nucleic Acid Sequence Bank," *Science*, 1980, *209*:1182; and Roberts to K. Stüber, 10 Dec. 1980, APS Archives (noting the comparative size of the Dayhoff and Los Alamos collections).

agreement not to redistribute the data. Whenever researchers accessed Dayhoff's database remotely, they would find the following notice on their screens: "Welcome to the NAS [Nucleic Acid Sequence] Reference Data System. You are licensed to use this data for your own research. As a licensee, you are legally obliged not to redistribute the data or otherwise make it available to any other party."[60]

In her letters to the NIH reporting on the progress of her work, Dayhoff stressed that the data was being made freely available; but that benefit came at a price—namely, substantial funding through research grants. Yet even NIH and NASA funding was not sufficient to make the database self-supporting. Thus only two days after making the database available at no charge, Dayhoff was negotiating with Laurence H. Kedes at IntelliGenetics, a small private company that had just been founded in Palo Alto by several Stanford University faculty members to sell computer software to the emerging biotechnology market. Dayhoff, confessing her "immediate cash problems," asked IntelliGenetics to distribute her database commercially, thus abandoning her pledge to have it distributed free of charge.[61]

In view of the NIH's uncertain support, and since the negotiations with IntelliGenetics proved fruitless (Dayhoff may have withdrawn out of concern that the company might become a competitor), Dayhoff decided in June 1981 to sell access to the database through a subscription. Commercial users were asked to contribute anywhere between $3,000 and $10,000 and noncommercial users between $750 and $1,000 per year. Even though these amounts were modest, the charges marked the crucial symbolic difference—as in the case of her protein database—between a free public good and a commercial product. Dayhoff put this unambiguously: "We have tried to get the database on a businesslike basis." Indeed, when she made her database available on the SUMEX-AIM computer she hoped that the increased visibility would help her find "new customers."[62]

The size of Dayhoff's collection continued to increase at a rapid rate. Three months after its opening, it had grown from over 227,000 to over 340,000 residues; eight months later it held more than 500,000 residues. Funding, on the other hand, remained extremely tight. In July 1981 the NIH informed Dayhoff that it would cease funding her nucleic acid database. "Databases do not inspire excitement," lamented Dayhoff in a letter to a colleague. Writing to the NIH, she pointed to a direct connection between the lack of public funding and her decision to market the database on a commercial basis.[63]

Even though Dayhoff's expertise and the quality of her *Atlas* were undisputed, the proprietary model on which she based her collecting enterprise was consistently chal-

[60] Dayhoff to Jordan, 23 Oct. 1980 (requests for access); T. Smith to NBRF, 30 Apr. 1982 (nonredistribution agreement); and M. O. Dayhoff, "Progress Report 08710: 1.2.1981–31.7.1982," 15 Sept. 1980 (on-screen notice): NBRF Archives.

[61] Dayhoff, "Progress Report 08710: 1.2.1981–31.7.1982," 15 Sept. 1980; and L. H. Kedes to Dayhoff, 18 Sept. 1980, NBRF Archives (regarding Dayhoff's cash-flow problems and the request for help from IntelliGenetics). A similar arrangement was being made with a Japanese partner; see Dayhoff to K. Koike, 7 Nov. 1980, NBRF Archives. On IntelliGenetics see Timothy Lenoir, "Shaping Biomedicine as an Information Science," in *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems*, ed. Mary Ellen Bowden, Trudi Bellardo Hahn, and Robert V. Williams (Medford, N.J.: Information Today, 1999), pp. 27–45.

[62] Dayhoff to D. M. Moore, 25 Sept. 1981 (subscription fees); Dayhoff to D. A. Jackson (Genex), 20 Nov. 1981 ("businesslike basis"); and Dayhoff to Kedes, 12 June 1981 (hope for "new customers"): NBRF Archives.

[63] W. Barker to H. Aaslestad, 23 Feb. 1981 (more than 340,000 residues); Dayhoff to G. Milne, 18 May 1981 (more than 500,000 residues); Kirschstein to Dayhoff, 15 July 1981 (end of NIH funding); Dayhoff to Moore, 14 Sept. 1981 (lack of "excitement"); and Dayhoff to Kirschstein, 7 Aug. 1981 (connection between lack of funding and decision to market): NBRF Archives.

lenged by experimentalists. The fact that she had copyrighted her database, limited its redistribution, sought revenues from it, and used data submitted to the *Atlas* for her own research was considered by some of them as violating the moral economy of the experimental sciences. When James D. Watson, in his 1968 tell-all account of the discovery of the DNA double helix, revealed that he and Francis H. C. Crick had used some of Rosalind D. Franklin's unpublished crystallographic data to build their model, the reviewers almost unanimously condemned their behavior, as Franklin's data was considered to belong to her.[64] Dayhoff's standards of knowledge ownership were unacceptable to many experimentalists, who considered the data they produced to be their own and therefore to be published, distributed, and used only with their agreement. This tension would continue to plague Dayhoff's collecting enterprise in the years to come.

Meanwhile, over the summer following the Rockefeller meeting, Goad and Bell moved ahead on the possibility of contracting for the national database. They tried to convince other scientists that Los Alamos was "the natural place to locate a center for sequence analysis of DNA," primarily because of the national lab's unique "computer facility." The argument that computer power was essential for the success of a sequence database would be one of the cornerstones of the Los Alamos campaign to host the central facility.[65]

Starting from almost nothing at the time of the Rockefeller meeting, Goad and a small team comprising the computer scientists Minoru I. Kanehisa, the mathematician James W. Fickett, and the molecular biologist Christian Burks put much effort into creating a comprehensive database of DNA sequences. In May 1980, the Los Alamos group had invited its collaborators for "cake and coffee to celebrate 100,000 bases now in the DNA sequence library." The collection was then just about the size of Dayhoff's, but following her intense effort in the fall of 1980 the Los Alamos project fell far behind.[66] With little experience in data collecting, and no staff trained to scan the literature for published sequences, it was unclear how Goad and his team could possibly catch up with Dayhoff's rapidly growing database. Thus, two days after Dayhoff's collection became available, Temple F. Smith, soon to be a consultant for Los Alamos, asked her for a copy of her entire database. Smith did not hide the fact that he meant to become one of her "future competitors." Dayhoff turned him down. Again, in January 1981, Goad boldly decided to write to Dayhoff to request her most recent collection. After telling her how having access to her data had been useful in correcting errors in his own database, he asked, somewhat hesitantly, "I wonder if at some point you would consider the possibility of allowing your database to be resident in our files?" Goad did not have much to offer in exchange, since his collection was just a subset of Dayhoff's. On the other hand, it seemed to him that Dayhoff could hardly refuse to share data that she herself had not produced and over which she thus had little proprietary claim. Dayhoff, however, declined.[67] Goad was more successful in acquiring smaller collections from other researchers by proposing an ex-

[64] See the reviews included in the critical edition: James D. Watson, *The Double Helix: A Personal Account of the Discovery of the Structure of DNA: Text, Commentary, Reviews, Original Papers* (New York: Touchstone, 2001).

[65] T. T. Puck to C. A. Thomas, 16 July 1979, APS Archives. At the second general meeting on the development of a DNA database, in June 1980, Goad began to present "the big computers at Los Alamos" as a "solution": F. R. Blattner, "Report on EMBL Workshop on Computing & DNA," 24 June 1980, EBI Archives.

[66] Los Alamos Sequence Library, Mar. 1982, p. 3 (team composition; I thank Christian Burks for having shared this document with me); S. Simon to T10, 9 May 1980, APS Archives (quotation); and Dayhoff to DeVincenzi, 20 Aug. 1980, NBRF Archives (collection size).

[67] Smith to Dayhoff, 17 Sept. 1980; Goad to Dayhoff, 9 Jan. 1981; and Dayhoff to Goad, 7 Aug. 1981: NBRF Archives.

change for his own.[68] The EMBL also asked Dayhoff to contribute her collection for redistribution with the other data from the European laboratory's growing DNA sequence database. Here too Dayhoff refused to help, unless the EMBL was willing to protect her data by a nonredistribution clause. Gregory H. Hamm, who was in charge of the project at the EMBL, confessed that he was "somewhat puzzled" by this response and explained that he could not accept data into the EMBL database that was "subject to restriction, since this defeats the whole purpose of our effort."[69]

The contrasting attitudes of Dayhoff and Goad toward the ownership of data collections were already apparent in their early collecting efforts. Whereas Goad treated the Los Alamos sequence collections as free to be exchanged, Dayhoff considered her database as proprietary. This difference reflected alternative standards of knowledge ownership, but it also resulted from the uneven states of their collections. Goad, whose collection was much smaller than Dayhoff's, was more than willing to share it with her if she would reciprocate. In one way, however, Dayhoff and Goad were very similar: both resembled what Londa Schiebinger has called "armchair naturalists."[70] Unlike the more adventurous naturalists who actually traveled to remote places to collect specimens, armchair naturalists remained with their collections and focused on the organization and display of their specimens. These naturalists often built their collections by acquiring other collections, exchanging specimens with other collectors, or maintaining a network of correspondent-naturalists rather than by engaging personally in the search for specimens. Similarly, neither Dayhoff nor Goad had ever sequenced a protein or a piece of DNA; they relied on others to accomplish that. Goad tried to acquire sequences in bulk from other collectors, whereas Dayhoff obtained them by searching the literature and through daily interactions with those who had determined sequences in their laboratories. Because of the amount of work that went into acquiring each individual sequence and verifying it, Dayhoff perhaps felt more entitled than Goad to assert proprietary claims over the sequences she had assembled in her database, as did the earlier naturalists over the specimens in their collections.

MOBILIZING THE NATIONAL INSTITUTES OF HEALTH

Following the Rockefeller meeting of March 1979, and while Dayhoff, Goad, and the EMBL were each trying to lay the groundwork for a comprehensive collection of DNA sequences, the NIH began to discuss how to address the scientific community's call for such a centralized facility. The EMBL announced in June 1980 that it would make its database available in the near future, while the NIH was still preparing for its first workshop on "the need for a nucleic acid sequence data bank," to be held the following month. The EMBL's declaration played no small part in pushing the NIH to take a firm initiative to support the establishment of a database in the United States.[71]

[68] Goad obtained two important European collections: those of Kurt Stüber in Cologne and Richard Grantham in Lyon. See Goad to Stüber, 15 Jan. 1981; and Goad to Roberts, 14 Jan. 1981: APS Archives.

[69] G. Hamm to Dayhoff, 17 Mar. 1981, NBRF Archives.

[70] Londa L. Schiebinger, *Plants and Empire: Colonial Bioprospecting in the Atlantic World* (Cambridge, Mass.: Harvard Univ. Press, 2004), p. 57.

[71] "Minutes, Workshop on the Need for a Nucleic Acid Sequence Data Bank, July 14–15 1980, Bethesda," APS Archives. Possibly reflecting the growing competition between the American and the European projects, only American scientists were invited to the NIH meeting, whereas representatives from both continents had attended the previous EMBL meeting. The EMBL explicitly asked to be informed as to the NIH meeting's conclusions. See K. Murray to Jordan, 3 July 1980, NBRF Archives.

Even though there was considerable interest within the scientific community for establishing a database, there seemed to be "little agreement as to what kind of arrangement would best serve the field." Indeed, there were many possible ways to organize a database and just as many schemes for collecting the data, verifying its accuracy, and distributing it to the broader scientific community. The speakers at the NIH meeting presented their views on these different issues and debated, "often sharply," as to the best format for the database and its mode of distribution. By the end of the first day, however, a "consensus became evident."[72] On the second day, the participants drafted recommendations defining the needs of the scientific community. Those who were expected to compete for a possible contract were asked not to participate in these discussions. Dayhoff (NBRF), Goad (Los Alamos), and Kedes (IntelliGenetics) left the room.

"Clearly, we must act now," declared the authors of the recommendation report, after noting the exponential increase of sequence data and the decision of the EMBL to establish its own database. The authors insisted that the long-term goal should be not simply to constitute a collection of sequences but to construct a "sophisticated" and structured library. Its managers should "aggressively" collect and solicit sequence data—but only data that had been accepted for publication ("i.e. refereed data"). The "private communications" that had been included in previous databases, such as Dayhoff's, should thus be excluded from the future database. The data was to be made available through telephone and computer networks in order to provide "interactive access to the stored data."[73] Finally, the sequences should be in "the public domain." Participants thus reaffirmed the principles, first outlined at the Rockefeller meeting, of having a computerized and nonproprietary database. In an electronic message posted on the SUMEX-AIM system, Laurence H. Kedes summarized the result of the workshop: "A strong endorsement for the establishment of a national nucleic acid sequence data bank was hammered out today [and] the meeting adjourned with the optimistic expectation that there would never have to be another one."[74]

This expectation was overly optimistic, and numerous other meetings soon followed to work out the details. At the NIH, Elke Jordan, a phage geneticist and deputy director of the Genetics Program at the National Institute of General Medical Sciences (NIGMS), which funded most of basic sciences at the NIH, took the lead in organizing them and tried to convince other institutes to support a sequence database, since it would serve "scientists NIH-wide." Jordan, together with Ruth L. Kirschstein, the director of NIGMS, eventually succeeded in convincing different institutes within the NIH (NIGMS, NIAID, NCI, and DRR) to fund the project, together with the National Science Foundation, the Department of Energy, and the Department of Defense. The participation of the latter two departments in a biomedical project is less surprising given that they were trying to diversify the research priorities of the national laboratories toward topics more directly relevant to peacetime society.[75] In December 1981, the NIH finally issued a "Request for Proposals"

[72] Jordan to Dayhoff, 17 June 1980 ("little agreement"); Kedes, collective email, 15 July 1980 (debates); and M. Cassman and E. Jordan, "Minutes, Workshop on the Need for a Nucleic Acid Sequence Data Bank," 22 July 1980 (consensus): NBRF Archives.

[73] "Minutes, Workshop on the Need for a Nucleic Acid Sequence Data Bank, July 14–15 1980, Bethesda," APS Archives; and Kedes, collective email, 15 July 1980.

[74] M. Cassman and E. Jordan, "21 July 1980 Minutes, Workshop on the Need for a Nucleic Acid Sequence Data Bank," NBRF Archives; and Kedes, collective email, 15 July 1980.

[75] Jordan to W. Raub, 25 July 1980, National Center for Biotechnology Information Archives, Bethesda, Maryland (hereafter cited as **NCBI Archives**) (quotation); and Westwick, *National Labs* (cit. n. 50), Ch. 8. NIAID is the National Institute of Allergy and Infectious Diseases; NCI is the National Cancer Institute.

for the development and maintenance of a national nucleic acid sequence database containing all published sequences over fifty base pairs long. The most important stipulation was that the database was to be up to date within a year of the contract being awarded.[76]

It took almost three years after the Rockefeller meeting 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 somewhat embarrassing delay on the part of the NIH might have resulted from bureaucratic inertia, as some critics later charged. More to the point, the cautious attitude of the NIH reflected the fact it was unclear whether the NIGMS mission should include the funding of databases. Its stated mission was to support experimental research of importance for medicine, and the maintenance of a database clearly didn't fit that description.[77] But also, and more fundamentally, many doubted the scientific usefulness of a sequence collection, especially at a time when the experimental approach was triumphing. As an anonymous participant complained on the electronic billboard at SUMEX-AIM, there was resistance from "within the NIH among staff who feel that molecular geneticists really do not need such a facility."[78] The resistance within the NIH to the idea of funding a database—a kind of project that did not "inspire excitement," as Dayhoff had complained—reflected the priority given to experimental work. Frederick Sanger would later express this hierarchy clearly: "'Doing' for a scientist, implies doing experiments."[79] Collecting and comparing were common ways of producing knowledge in natural history but were often regarded as archaic by experimental biologists, even if these practices involved sophisticated computers.

At least three proposals were submitted to the NIH: one by the National Biomedical Research Foundation (Dayhoff); one by Los Alamos Scientific Laboratory (Goad), teaming up with Bolt, Beranek, and Newman (Bilofsky); and one by Los Alamos, together with IntelliGenetics (Kedes). The first two were selected by the NIH for further evaluation. These two proposals and the NIH referees' views on them offer a window into different solutions to the challenge of data collection and the problem of data ownership in the experimental sciences. The two proposals were similar in many ways, reflecting the convergence of views that had resulted from more than two years of meetings among those invested in the development of a database. However, they also reflected fundamental differences with respect to credit attribution, data access, and the ownership of knowledge.

COLLECTING DATA, NEGOTIATING CREDIT AND ACCESS

In the natural history tradition, a number of different strategies have been adopted to build collections. As Paula Findlen has argued, the networks of object exchange that helped fill the early modern cabinets of Ferrante Imperator and Ulisse Aldrovandi, for example, were based largely on patronage relationships. Eighteenth-century French gardens also depended on a gift exchange network between botanists, a "system of polite indebtedness,"

[76] S. W. Thornton, "RFP," 1 Dec. 1981, NBRF Archives.

[77] Smith, "History of the Genetic Sequence Databases" (cit. n. 9) (on the charge of bureaucratic inertia); and Strasser interview with Kirschstein, 22 Feb. 2006 (compatibility with NIGMS mission).

[78] Electronic message posted on SUMEX-AIM, 10 Sept. 1980, NCBI Archives. Elke Jordan circulated a copy of this message within the NIH, most likely to gather support: Jordan to Raub *et al.*, 11 Sept. 1980, NCBI Archives.

[79] Dayhoff to Moore, 14 Sept. 1981, NBRF Archives; and Sanger, "Sequences, Sequences, and Sequences" (cit. n. 16), p. 1.

as Emma Spary has put it.[80] The great natural history museums of the nineteenth century, such as the American Museum of Natural History and the French Museum, relied on commissioned expeditions, but also on the growing market for rare natural history specimens, to assemble their collections. The British and the French empires could also leverage the resources of their colonies to supply specimens. As Spary has summarized it so perfectly: "natural history is a science of networks."[81] These past networks depended crucially on the social configuration of the communities in which they operated and the values these communities embodied, especially with regard to the ownership of knowledge. In their twentieth-century proposals for the DNA sequence database, Dayhoff and Goad offered different strategies to address the problems of collecting. Dayhoff's approach, once again, reflected her idea that published knowledge belonged to the collector, whereas Goad's was more in tune with the idea that published knowledge belonged to the community as a whole.

Even though Dayhoff already had the largest existing sequence database, she believed that her collection was as yet "a mere shadow of its ultimate grandeur." In order to realize her vision, she planned to collect data as she had done in the past, by surveying the literature; she estimated that twenty-nine journals contained more than 98 percent of all the published sequences. Dayhoff insisted on the importance of comprehensiveness, a key value in the natural history collecting tradition, just as precision was a key value for the experimental tradition.[82] "Comprehensiveness" was an ambiguous term, however. From Dayhoff's proposal it was clear that she would give priority to long sequences (over five hundred base pairs) over short sequences (fifty base pairs, which represented the bottom limit required by the NIH). This decision reflected the fact that she privileged the use of the database for research in evolutionary biology, rather than for assisting researchers in molecular genetics.

Goad and his partner at BBN, Howard S. Bilofsky, envisioned collecting data in a similar way, but with one crucial difference. Apparently more sensitive than Dayhoff to the fact that experimentalists had a strong sense of ownership over their sequences, they proposed to rely on cooperation with journal editors, rather than only on voluntary contributions from authors or the scanning of the published literature. They stressed that coordination with journal editors "on topics ranging from electronic uploading of published sequences to standards for annotation" was essential to the success of the database. They suggested a mechanism that had first been proposed at the EMBL to bring authors to collaborate in the collecting effort: "We believe—and a number of journal editors have already agreed in principle—that once a national centre is established, most journals will be willing to furnish or *require* authors to furnish, a copy of the original figures, or, preferably, a computer-readable copy of each sequence, to the national data bank."[83]

Goad and Bilofsky insisted that the cooperation of journals, rather than of individuals,

[80] Findlen, *Possessing Nature* (cit. n. 15), Ch. 8; and Emma C. Spary, *Utopia's Garden: French Natural History from Old Regime to Revolution* (Chicago: Univ. Chicago Press, 2000), p. 77.

[81] Spary, *Utopia's Garden*, p. 97. On the accumulation strategies of the museums see Barrow, "Specimen Dealer" (cit. n. 46). On the benefits of colonial resources for the supply of specimens see, e.g., Richard W. Burkhardt, Jr., "Naturalists' Practices and Nature's Empire: Paris and the Platypus, 1815–1833," *Pacific Science*, 2001, *55*:327–341.

[82] M. O. Dayhoff, "Technical Proposal: Establishment of a Nucleic Acid Sequence Data Bank," 1 Mar. 1982, NBRF Archives, pp. 12, 18. Dayhoff planned to explore the remaining 2 percent by manually searching through bibliographic indexes. On the experimental tradition see M. Norton Wise, *The Values of Precision* (Princeton, N.J.: Princeton Univ. Press, 1995).

[83] BBN, "Establishment of a Nucleic Acid Sequence Data Bank," Mar. 1982, NBRF Archives, pp. 16, 26.

was the key to the collecting enterprise. They expected the electronic transmission of data between journals and databases to become increasingly common "as computer to computer transmission grows more facile." The NIH reviewers judged that such reliance on journals would be an excellent mechanism for collecting data and were very confident that the Los Alamos team "should have no difficulty bringing the database up to date within the first year." Conversely, they criticized Dayhoff's traditional approach to data collecting, which rested essentially on (wo)manpower to scan published papers and on individual relationships with the authors of sequences. They judged that Dayhoff had given "little thought . . . to increasing the efficiency of data collection and dissemination," which raised concerns about her capacity to meet the deadline in view of the exploding number of sequences becoming available.[84]

Apart from the matter of efficiency, in relying on the authority of scientific journals and their role in the scientific reward system Goad and Bilofsky were appealing to a different system of values than Dayhoff. In the early 1980s, sequences were considered highly proprietary knowledge, and their publication in a journal constituted a key reward for the individual author or laboratory. But submission to a database established neither priority nor authorship. Worse, the disclosure of sequence data could give important hints to competing groups working on the same sequence. In the experimental sciences, publication—and thus the attribution of priority and authorship—was a main motive that brought scientists to make data public. Authorship, in turn, brought recognition and scientific credit, the key social rewards for producing knowledge in science. As the molecular biologist Lewis Wolpert reflected a couple of years later: "J. B. S. Haldane is reported to have said that his great pleasure was to see his ideas widely used even though he was not credited with their discovery. That may have been fine for someone as famous and perhaps noble as Haldane, but for most scientists recognition is the reward in science."[85]

Dayhoff's system of data collecting ran against one of the essential values of the experimental sciences' moral economy: namely, that the production of knowledge deserves individual, not collective, credit. As a molecular biologist promoting databases lamented in 1989, "scientists are fierce individualists who consider themselves lone seekers of new knowledge. . . . The idea that they are part of an unorganized community of minds involved in a collective effort to seek knowledge may be foreign to most of them."[86]

Neither Dayhoff, nor Goad, nor any of the participants at the initial meetings on the national database envisioned challenging the line of demarcation between public and private knowledge set by publication in a printed journal. Even Dayhoff had explicitly shied away from following this trajectory by stating in the preface of her early *Atlas of Protein Sequence and Structure* that the editors did not want to "become involved in

[84] *Ibid.*, p. 26; and NIGMS, "Second Staff Evaluation of Proposal from BBN," 21 June 1982, NBRF Archives, p. 1. When the NIH reviewers asked Dayhoff if she planned to collaborate with journal editors, she replied that she supported the mandatory submission scheme but that she would not make it a priority. In a phone conversation with Ken Murray, on 6 June 1982, she explained that she was quite skeptical about the collaboration of journal editors. See K. Murray, "Summary of Telephone Conversation with Margaret Dayhoff," 6 June 1982, EBI Archives, Folder 6.

[85] Lewis Wolpert, *The Unnatural Nature of Science* (London: Faber & Faber, 1993), p. 89. See also Warren Hagstrom, "Gift Giving as an Organizing Principle in Science," in *Science in Context: Readings in the Sociology of Science*, ed. Barry Barnes and David Edge (Cambridge, Mass.: MIT Press, 1982), pp. 21–34.

[86] Alain E. Bussard, "Data Proliferation: A Challenge for Science and for Codata," in *Biomolecular Data: A Resource in Transition*, ed. Rita Colwell (Oxford: Oxford Univ. Press, 1989), pp. 11–15, on p. 13.

questions of history or priority," notwithstanding the fact that they accepted unpublished data.[87]

Other databases, such as the Protein Data Bank at Brookhaven, had taken an even more conservative route with unpublished data in order to protect individual authors and so as not to challenge the authority of journals. Data could be deposited in the database without being made available to outside users for one to four years after the publication in a journal of the general conclusions derived from that data, in order to protect the authors' ability to exploit it further. Similar concerns prevailed among researchers in the field of DNA sequence databases, and in the absence of a mechanism to protect the privacy of their data these concerns hindered the collection efforts. As a reporter put it in *Science*, explaining some of the resistance to a centralized database, "many people were uncomfortable with the prospect that sequences might become freely available before principal investigators had had time to work with them and therefore benefit from their sequencing efforts."[88]

In the natural history tradition, the success of the data-collecting enterprise had rested on the authority of the collectors within the communities in which they were operating. In this respect Dayhoff and Goad were in a weak position, because they were personally and institutionally very peripheral to the community from which data would be collected. The NBRF was a small, nonprofit research organization that, aside from Dayhoff's theoretical work in molecular evolution, was best known for the development of computer applications for medicine, not for contributions to basic scientific research. Los Alamos had its own negatives: its association with military projects and its specific culture as a national laboratory. The national laboratory was best known as the home of the Manhattan Project during World War II and, during the Cold War, of the thermonuclear weapons project, in which Goad was personally involved. The fact that Los Alamos was considered an institution with major ties to the military isolated it from the biomedical community. In a résumé he sent to his superiors at Los Alamos, Goad stated that from 1950 to 1969 he had been "active in all phases of the theoretical work involved in nuclear weapons design and development, including weapon effects," and had served at the same time as "consultant for the US Air Force Foreign Weapon Evaluation Group" and on several other weapons research committees.[89] Significantly, Goad omitted all these activities from the résumé he submitted with his application to the NIH. In the wake of the public criticisms against the involvement of science, and especially physics, in the military-industrial complex voiced since the mid-1960s, biomedical researchers had become particularly wary of having any connection with the military. The participation of the Department of Defense (DOD) in the database project, for example, had caused "some practitioners a degree of nervousness," as a reporter in *Science* pointed out. Elliott Levinthal, of the DOD's Advanced Research Projects Agency, explained that the department was not interested in "chemical or

[87] Dayhoff *et al.*, *Atlas of Protein Sequence and Structure* (cit. n. 39), p. xiv.

[88] Roger Lewin, "Long-Awaited Decision on DNA Database," *Science*, 1982, *217*:817–818, on p. 817. Regarding the arrangements in force for the Protein Data Bank see Commission on Journals, "Deposition of Macromolecular Atomic Coordinates and Structure Factors with the Protein Data Bank—Modified Policy," *Acta Crystallographica*, 1982, *B38*:1050. Even so, concerns about the possibility that others might exploit their work led researchers to withhold crystallographic data from the Protein Data Bank; see Marcia Barinaga, "The Missing Crystallography Data," *Science*, 1989, *245*:1179–1181.

[89] "Report, Group T-10: Theoretic Biology and Biophysics," 1977, APS Archives. On the role of the national laboratory in the development of the H-bomb see Richard Rhodes, *Dark Sun: The Making of the Hydrogen Bomb* (Sloan Technology Series) (New York: Simon & Schuster, 1995).

biological warfare."[90] It seems unlikely that such disclaimers would have reassured molecular biologists regarding any DOD involvement in the database project. Having just confronted the turmoil of the recombinant DNA controversy in the aftermath of the Asilomar conference of 1975, molecular biologists seemed particularly unwilling to expose themselves to another potential source of public criticism. To those contemplating the establishment of a national sequence database, it was unclear whether molecular biologists would collaborate fully in a project hosted at Los Alamos, a critical uncertainty since, as Goad himself had explicitly recognized in his application, the success of the sequence data bank would depend crucially on "the level of cooperation and communication the contractor establishes with the scientific community."[91]

A second, related aspect of the Los Alamos identity that could threaten its "cooperation and communication" with the biomedical community was simply the fact that it was a national laboratory. Secrecy and security were perceived to be key elements of the national laboratory culture because of its close relationships with the military and the fact that its research often related to national security interests. The NIH referees, worried that this might constitute an obstacle to the necessary relationship of trust between the biomedical community and those operating the database, investigated the question in various oblique ways. They asked the Los Alamos–BBN team, for example, "Exactly what access will users have to the Cray computers at Los Alamos?" Goad and Bilofsky had to confess that the Crays that had figured so prominently in their application would be out of bounds for the general user: they were "not accessible from outside Los Alamos because of security restrictions."[92]

In submitting an application to the NIH, Goad was very much aware that Los Alamos was both an asset and a liability for his project. The national laboratory was undergoing a new security partitioning, and Goad expressed his concerns about the impression that biomedical researchers might have on visiting the site: "I have some misgivings about being within the secured area during the first six months of 1982. We expect to be evaluated during that time for the computer-based DNA sequence resource. . . . It is important that we be perceived by the molecular biology community, and particularly by our reviewers, as offering completely free and open access to the information and programs we will be collecting." Having the database on a site where access was restricted would be all the more damaging in that, as Goad pointed out, there were "people who already feel, however unfairly, that our openness is compromised by national security programs that demand security protection." Goad wanted to "avoid anything that unnecessarily tends to reinforce that view," and he made every effort to make Los Alamos appear more civilian and less military—more open and less secret—in order to accommodate the civilian ethos of the biomedical community. As Richard J. Roberts would later explain, "Biologists didn't want to be associated with a weapon lab; biologists thought they were pure, and physicists were not."[93]

[90] Lewin, "Long-Awaited Decision on DNA Database" (cit. n. 88), p. 818. See also Stuart W. Leslie, *The Cold War and American Science: The Military-Industrial-Academic Complex at MIT and Stanford* (New York: Columbia Univ. Press, 1993).

[91] BBN, "Establishment of a Nucleic Acid Sequence Data Bank," Mar. 1982, NBRF Archives, p. 16. Regarding molecular biologists' unwillingness to confront further controversy see Bruno J. Strasser interview with Richard J. Roberts, 2 July 2008.

[92] BBN to NIH, 7 May 1982, APS Archives. However, calculations could be made on the Crays and then transferred to an accessible computer by Los Alamos personnel, as Bilofsky and Goad tried to explain.

[93] Goad to P. Carruthers and M. Slaughter, 3 Nov. 1981, APS Archives; and Strasser interview with Roberts, 2 July 2008.

Goad and Bilofsky sought to emphasize the unique resources offered by Los Alamos to compensate for its negative cultural resonances. They advertised the powerful computers, including four Cray-1 supercomputers, that made the national laboratory "one of the most powerful computing centres in the world." This tremendous number-crunching capacity was indispensable for the database, the authors argued, because "the Los Alamos approach to sequence data collection" relied heavily on sophisticated computer software to verify and annotate the submitted sequences. The Cray computers, for example, would be very useful for the curators, who could make searches through the entire database to determine whether a new sequence was homologous to an existing one and annotate the new sequence entry in the database. These powerful computers would also be used to search each sequence for specific patterns indicating functional elements. Since some of the programs needed to accomplish these tasks were "computationally quite intensive," it was claimed that they could only "be operated cost-effectively on the Los Alamos Cray computers."[94] Dayhoff, on the other hand, had emphasized the human expertise of her team in verifying sequences. In any case, in terms of computational power Dayhoff's "modern, high speed computer" certainly could not compete with Goad's four Cray-1 machines, the fastest computers in the world. The NIH reviewers found the Los Alamos computing power "impressive and unique" but didn't question whether such extraordinary speed was in fact necessary for managing a sequence database. The use of a computer program to align two sequences and verify their statistical significance could typically take several minutes to several hours on a minicomputer, such as the popular PDP-11.[95] On a large computer, such as Dayhoff's DEC VAX-11/780, the same operation would take just seconds. On a supercomputer such as the Los Alamos Cray, it would be orders of magnitude faster.

DISTRIBUTING DATA, NEGOTIATING OWNERSHIP

Just as the NIH reviewers were examining the proposals for a sequence database, public debates were raging over the effects of patenting the living products and the techniques of molecular biology. In 1980 the United States legislature passed the Bayh-Dole Act, expanding universities' intellectual property rights over federally funded research. The same year, in Diamond v. Chakrabarty, the U.S. Supreme Court ruled that living organisms could be patented, after noting that a 1951 congressional report had concluded that "anything under the sun that is made by man" was patentable. Independently, in December the U.S. Patent Office issued a patent to Stanley Cohen, Herbert Boyer, Stanford University, and the University of California, San Francisco, for their basic genetic engineering technique. Fears ran high in the scientific community that the rise of intellectual property would lead to increasingly secretive practices and hinder the production of scientific knowledge.[96] In such a context, it is unsurprising that the greatest concern for the NIH,

[94] BBN, "Establishment of a Nucleic Acid Sequence Data Bank," Mar. 1982, NBRF Archives, p. 24.

[95] See Table 1, using the Needelman-Wunsh algorithms, in Rodney A. Jue, Neal W. Woodbury, and Russell F. Doolittle, "Sequence Homologies among *E. coli* Ribosomal Proteins: Evidence for Evolutionarily Related Groupings and Internal Duplications," *Journal of Molecular Evolution*, 1980, *15*:129–148, on p. 143.

[96] See, e.g., D. Dickson, "Stanford Ready to Fight for Patent," *Nature*, 1981, *292*:573; and Sally Smith Hughes, "Making Dollars out of DNA: The First Major Patent in Biotechnology and the Commercialization of Molecular Biology, 1974–1980," *Isis*, 2001, *92*:541–575. On these developments see also Daniel J. Kevles, "*Diamond v. Chakrabarty* and Beyond: The Political Economy of Patenting Life," in *Private Science: Biotechnology and the Rise of the Molecular Sciences*, ed. Arnold Thackray (Philadelphia: Univ. Pennsylvania Press, 1998), pp. 65–79; Kara Swanson, "Biotech in Court: A Legal Lesson on the Unity of Science," *Social Studies*

even beyond the question of the mechanism of data collection, was the issue of copyright on the sequence data and, more generally, the issue of ownership of the information included in the database.

The NIH prompted both applicants to explain how they planned to obtain copyright agreements with the journals from which the sequence data would be copied into the future database. Neither of the applicants declared an intention to obtain copyright permission from the journal publishers. This question was perhaps most embarrassing for Dayhoff, because she had been copyrighting her *Atlas of Protein Sequence and Structure* from the outset, as well as her demonstration DNA database, including its electronic edition. Reviewers implied that this practice might bring potential legal difficulties, but Dayhoff dismissed the argument by replying that in seventeen years her copyright had never been challenged by a journal. Robert S. Ledley, director of the NBRF, had sought legal advice on the subject and was informed that the inclusion of sequences from copyrighted articles would constitute a "fair use."[97]

However, the NIH reviewers pressed the matter further, questioning both applicants specifically as to whether the NIH would "own all data in the database . . . regardless of whether it was collected prior to inception of the bank?" Goad and Bilofsky replied the most clearly, explaining that they did not intend "to assert any proprietary interest whatsoever in any data." Furthermore, the Los Alamos–BBN team noted that Los Alamos had already made its database "freely available" to anyone, "without restriction on further distribution." The NBRF made similar claims concerning the future database, at least for as long as it would be supported by the NIH, emphasizing that the sequence data would be in "the public domain and available to all interested people" and that users would "be free to make whatever use they wish of the information, including redistribution." Dayhoff left some ambiguity, however, as to whether the NBRF would reclaim proprietary rights to the data that had been collected before the beginning of the contract once the contract had terminated.[98] Given that the NBRF had been running its database on a "businesslike basis," it seemed likely that it would want to revert to that model after the termination of an NIH contract. The Los Alamos–BBN team made sure that the NIH referees would remember this point: the NBRF, they noted in their own answers to the questions, had "sought revenues from sales of their database" and "prevented redistribution," including "to NIH users of the PROPHET system." Goad was clearly aware that Dayhoff's "businesslike" database was handicapping her application to the NIH when he wrote to a colleague: "we seem to be developing an edge . . . as our principal competitor becomes increasingly enmeshed in proprietary arrangements." Indeed, the NIH reviewers were clear about their distrust of Dayhoff's standing "proprietary arrangement," which they found "not reassuring" for the future of the public database.[99]

The issue of data ownership was a legal one, involving copyright, but also a practical one: namely, how would data physically be distributed? Dayhoff planned to distribute the DNA database as she had distributed the *Atlas*, by sending out magnetic tapes and printing

of Science, 2007, *37*:357–384; and Doogab Yi, "The Recombinant University: Genetic Engineering and the Emergence of Biotechnology at Stanford, 1959–1980" (Ph.D. diss., Princeton Univ., 2009).

[97] BBN to NIH, 7 May 1982, APS Archives (obtaining copyright agreements); Ledley and Dayhoff to S. Thornton, 7 May 1982, NBRF Archives (obtaining copyright agreements); and J. Seeber to Ledley, 5 Mar. 1982, NBRF Archives ("fair use").

[98] BBN to NIH, 7 May 1982; Ledley and Dayhoff to Thornton, 7 May 1982; and M. O. Dayhoff, "Replies to Information Requested by May 10, 1982" [undated, but 1982], NBRF Archives, p. 1.

[99] BBN to NIH, 7 May 1982; Goad to Carruthers and Slaughter, 3 Nov. 1981, APS Archives; and NIGMS, "Second Staff Evaluation of Proposal from NBRF," 21 June 1982, NBRF Archives, p. 1.

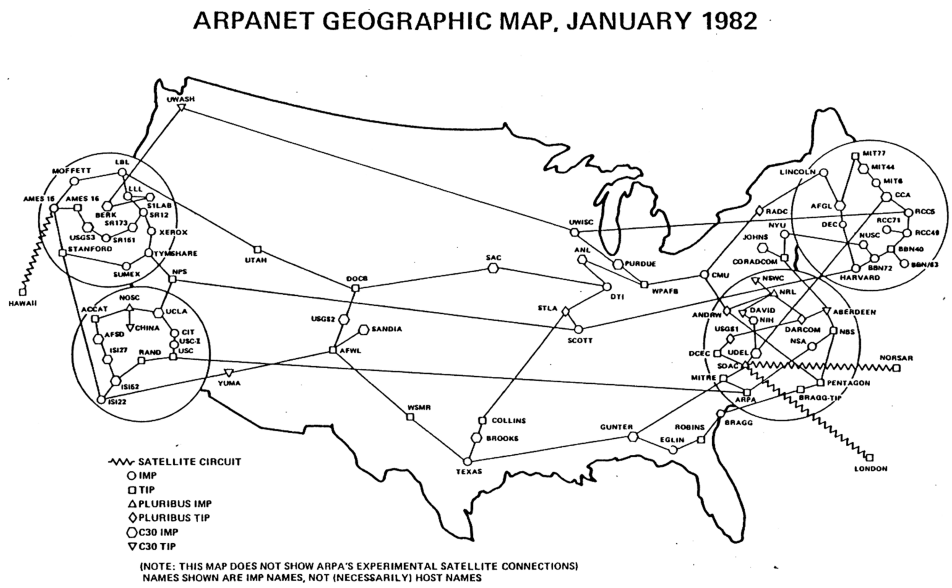

Figure 4. *The ARPANET network as presented in the Los Alamos–BBN application for the database contract, January 1982. Reproduced with permission from the National Biomedical Research Foundation Archives.*

sequences in book format. In addition, she proposed to offer three dial-up telephone lines to the NBRF computer, a DEC VAX-11/780, where the database was stored and which allowed remote computing.[100] Unsurprisingly, the Los Alamos–BBN proposal was more sophisticated technologically. Goad and Bilofsky emphasized the fact that the Los Alamos database had been "available to all scientists all over the world" through on-line connections to the Los Alamos computers and also through the BBN-based PROPHET system and the Stanford-based SUMEX-AIM system, which were both connected to national computer networks such as the ARPANET network (see Figure 4) and the commercial Telenet network. Goad and Bilofsky anticipated providing "on-line access to the Los Alamos facilities over national networks" to those contributing data and to those in charge of managing the database, and they envisioned that the future would involve "extensive on-line user access" and "electronic communication and collaboration among users." They also seemed well aware that microcomputers were becoming increasingly common in biomedical research, since they proposed to develop software to read the new 5¼-inch disks used on "the small word processing computer systems that have been installed in hundreds of laboratories around the country."[101] The NIH reviewers praised the extensive use of networks and microcomputers envisioned in the proposal.

When prompted by the NIH reviewers to say whether she too could offer network access to the database, in addition to the telephone connections, Dayhoff replied that it was a costly option and that she did not see it as indispensable for the distribution of data.

[100] M. O. Dayhoff, "Technical Proposal: Establishment of a Nucleic Acid Sequence Data Bank," 1 Mar. 1982, NBRF Archives. This computer was perhaps still "modern," but in fact it had been introduced four years earlier.

[101] BBN, "Establishment of a Nucleic Acid Sequence Data Bank," Mar. 1982, NBRF Archives, pp. 117, 24, 6, 27. The extension of ARPANET ca. 1982 is visualized in Allen Newell and Robert F. Sproull, "Computer-Networks—Prospects for Scientists," *Science*, 1982, *215*:843–852.

Furthermore, unlike the Los Alamos team, she could not hope to use the ARPANET network, which was run by BBN for the Department of Defense, because it was restricted to institutions—such as Stanford University and Los Alamos—that carried out DOD-related work. Los Alamos had obtained special permission from the DOD to use the ARPANET for the sequence database project.[102] The PROPHET computing system, even though it was funded by the NIH, was also controlled by BBN, and it was open only to very specific users. Without access to computer networks, Dayhoff relied on dial-up connections to the NBRF computer at Georgetown, using only the telephone network. She noted, as well, that connections could be made through small computers costing "less than $1,000." Thus, if Dayhoff somewhat underrated network access, she seemed just as aware as Goad of the growing potential of microcomputing at a time when "computer anxiety was still strong in the molecular biology community."[103]

The problem of network access should not be seen as a technical matter alone; it was also an epistemic and a cultural one. The availability of the database on the growing computer networks and the possibility of accessing the database from microcomputers would guarantee that the data could reach an increasingly broad audience and become accessible for wider review by the scientific community. The objectivity of science was perceived as resting precisely on the ideal of a public review process and the open disclosure of scientific facts. As Joshua Lederberg put it in a 1978 article on "digital communication," "the claim of science to universal validity is supportable only by virtue of a strenuous commitment to global communication." For Lederberg and others, computer networks were becoming key to the communication of experimental results in science, especially in view of the recent "information-explosion."[104] By the early 1980s, computer resources and networks such as SUMEX-AIM, Telenet, and ARPANET also increasingly embodied the cultural values of shared resources and free access to data—precisely the values that the NIH reviewers hoped the database would represent. Providing access to the database through computer networks became a matter of reaffirming the value of broader participation in the production of scientific knowledge.

Three months after the proposals had been submitted, at 5:00 p.m. on 30 June 1982, Dayhoff received a phone call from the NIH: the contract had just been signed with Los Alamos and BBN, providing $3.2 million over five years to set up and maintain a nucleic acid sequence database.[105] *Science* welcomed this "long-awaited" decision and, in rhetoric typical of concerns in the 1980s about American industrial decline, deplored the fact that rival facilities in Europe and Japan had beaten the NIH to the mark. Goad announced that the database would contain all published sequences within a year. Ledley, the president of the NBRF, was staggered and thought it "inconceivable" that his institution had lost the contract to Los Alamos and BBN, while Dayhoff expressed great "surprise" at the decision and privately showed "huge disappointment."[106] For almost twenty years, she had

[102] Ledley and Dayhoff to Thornton, 7 May 1982, NBRF Archives (network access not indispensable); and "Meeting at DOE [unsigned note]," 9 Jan. 1982, NCBI Archives (permission to use the ARPANET).

[103] Ledley and Dayhoff to Thornton, 7 May 1982; and Lewin, "Long-Awaited Decision on DNA Database" (cit. n. 88), p. 817.

[104] Joshua Lederberg, "Digital Communications and the Conduct of Science: New Literacy," *Proceedings of the IEEE*, 1978, *66*:1314–1319, on p. 1314. See also Elliott C. Levinthal, Raymond E. Carhart, Suzanne M. Johnson, and Lederberg, "When Computers 'Talk' to Computers," *Industrial Research*, 1975, *17*:35–42; and Newell and Sproull, "Computer-Networks—Prospects for Scientists" (cit. n. 101).

[105] R. S. Ledley, "Chronology of NBRF/BBN Protest," 14 Jan. 1983; and Department of Health and Human Services, "Negotiated Contract," 30 June 1982: NBRF Archives.

[106] Lewin, "Long-Awaited Decision on DNA Database" (cit. n. 88); Thornton to C. Fretts, 16 Aug. 1982,

been the world's leading sequence collector, at a time when data collecting was hardly considered a worthy scientific enterprise. When science funding agencies and the scientific community finally recognized the potential of sequence databases for the production of knowledge, she lost the contract to build such a database to a physicist with no prior experience in sequence collecting. Dayhoff decided to focus again on her protein sequence collection, leaving DNA collection to the Los Alamos National Laboratory. She did not see GenBank develop, however, as she died of heart failure eight months after the contract was awarded.

The development of GenBank at Los Alamos turned out to be far more difficult than Goad and his team had envisioned. Instead of being up to date with all published sequences within a year, as they had promised and as the NIH contract required, GenBank's collection lagged increasingly behind the rising number of published sequences, despite teaming up with its European and Japanese partners to collect sequences from their respective geographical areas.[107] For years, journal editors resisted making the submission of sequence data mandatory. At the end of the 1980s, they finally adopted and enforced such policies, thus durably solving the problem of data collection. By that time DNA sequences were being determined so rapidly—and automatically—that their epistemic value had decreased drastically, and it was becoming increasingly cumbersome to print them in the pages of journals anyway. The existence of GenBank became essential for the Human Genome Project, which began in 1990. GenBank served as a repository for newly determined sequences, as a tool for sequencers assembling genomes, and as a powerful database for the growing number of researchers in bioinformatics. In 1992 GenBank left Los Alamos and was integrated in the NIH's National Center for Biotechnology Information, part of the National Library of Medicine. This shift, from a military to a biomedical institution, reflected the changing fortunes of the physical and the life sciences in a post–Cold War world, as well as the loosening of some of the links between physics and biology that had been so productive for the emergence of molecular biology.[108] But it also indicated that collections were no longer viewed as relics of an archaic past associated with natural history but, rather, as essential tools for the production of knowledge in the experimental life sciences.

CONCLUSIONS

In the history of the creation of GenBank, we can revisit two major historical transformations in the experimental life sciences of the late twentieth century: changing moral economies (the rise of open access) and changing research practices made possible by electronic databases (the rise of comparative practices). The making of GenBank, arguably the most important collection of data in the life sciences today, reflected and at the same time contributed significantly to both of these deep historical changes.

There are several ways to read the competition between Dayhoff and Goad and its final outcome. While gender and personality issues may have played their part, I have argued

NBRF Archives (NBRF reaction); and Bruno J. Strasser interview with Winona Barker, Georgetown, 1 Sept. 2005 (Dayhoff's "huge disappointment").

[107] On the lagging collection see GenBank advisors meeting, Minutes, 6 Nov. 1987, EBI Archives.

[108] On GenBank's move to the National Center for Biotechnology Information see D. Benson, D. J. Lipman, and J. Ostell, "Genbank," *Nucleic Acids Research*, 1993, *21*:2963–2965. More generally, see Daniel J. Kevles, "Big Science and Big Politics in the United States: Reflections on the Death of the SSC and the Life of the Human Genome Project," *Historical Studies in the Physical and Biological Sciences*, 1997, *27*:269–297; and Rasmussen, "Mid-Century Biophysics Bubble" (cit. n. 50).

here that the key differences between the contenders related to issues of credit, access, and ownership in science. These were the major components of the first historical transformation, that of the shift in the moral economies of the life sciences. Dayhoff and Goad faced the complex challenge of adapting a natural history endeavor, based on the collection of natural objects, to the moral economy of the experimental life sciences in the late twentieth century. For Dayhoff, as for many naturalists in the past, collections and the items they contained were private property, and the collector was free to use them as commodities, gifts, or public goods. No item carried much value until it became part of a collection—that is, an element in a system designed for the preservation and production of knowledge. The relations among elements, revealed through their systematic comparison, were more valued than the elements themselves, and thus the collector could take credit and claim authorship for bringing these relationships to light. Naturalists studying collections such as those in museums of natural history were entitled to appropriate the work of the numerous individuals whose contributions had filled these collections.[109] Dayhoff did precisely that in her *Atlas* and in numerous scientific publications that drew conclusions from the sequence data provided to her for inclusion in the database.

A very different set of norms prevailed in the experimental life sciences in the late twentieth century. The production of knowledge there rested on revealing singular facts of nature in the laboratory. Elucidating the structure and function of molecules, for example, was considered by the experimentalist community a key intellectual achievement deserving credit and recognition of authorship, and the experimental scientists who succeeded in such work felt a sense of ownership over the knowledge they had produced. The fact that so many Nobel Prizes have rewarded the determination of molecular structures and functions indicates that these were considered major individual scientific accomplishments. Such were the premises on which Goad built his vision of a sequence database. He laid no claim to ownership over the data it would contain and made it as widely accessible as possible, eventually taking advantage of increasingly globalized computer networks. He also declined to exploit the database's scientific content, leaving that to the experimentalists who had determined the sequences and to the emerging community of computational biologists who would soon rally under the banner of "bioinformatics." In doing so, he successfully adapted natural history's key task—to collect objects of nature—to the moral economy of the experimental sciences. Goad took into account the growing resistance of some academic scientists and science administrators to the appropriation of biological knowledge and their corresponding efforts to make it publicly available. But at the same time, he was keenly aware that in the reward system of the experimental sciences the production of knowledge deserved individual, not collective, recognition. He thus astutely suggested that the database rely on the authority of journal editors, whose power to attribute authorship would compel researchers to share the knowledge they had produced.

In addition to these issues of credit and access, the problem of ownership also defined the debates over the creation of GenBank. Precisely during that time, powerful forces were at work to make scientific knowledge more relevant to the U.S. economy. The 1980 Bayh-Dole Act, which expanded universities' intellectual property rights over federally funded research, the *Diamond v. Chakrabarty* ruling in that same year, and the inauguration in 1981 of President Ronald Reagan's business-friendly administration all contrib-

[109] For an excellent example of the importance of collections for systematic work see Kristin Johnson, "Ernst Mayr, Karl Jordan, and the History of Systematics," *Hist. Sci.*, 2005, *43*:1–35.

uted to a climate in which the commercialization of knowledge, especially in biotechnology, was strongly encouraged.[110] It might seem paradoxical that in such a context the NIH and leading molecular biologists so strongly resisted any proprietary models for a sequence database. The personal and professional commitments of some of the most influential figures involved in building the sequence database shed light on this paradox. When prompted to say whose advice she had taken when GenBank was being set up, Ruth L. Kirschstein, the director of NIGMS, replied without hesitation: "Rich Roberts." Roberts had been an unusually strong advocate of sharing data and research materials. He had set up his own collection of restriction enzymes, which he distributed freely to the scientific community, an uncommonly generous practice. In the following years, he would become one of the most vocal advocates of open-access publishing. Also influential was the computer science background of some of the reviewers chosen by the NIH to examine the database proposals. Having emerged from the counterculture movement, many computer scientists resisted commercial appropriations of knowledge and valued the sharing of computer codes.[111] Most likely, they saw genetic code as a parallel to computer code and thus as a resource to be made freely available to others for the greatest benefit of the community.

The fact that GenBank grew out of the molecular biology community and was primarily destined to serve that community has played an equally important role in the development of open access. Biologists had become accustomed in the twentieth century to sharing experimental materials, especially organisms, for free. The stock collections of classical geneticists, such as Thomas H. Morgan's *Drosophila* mutants and Rollins A. Emerson's corn seeds, were explicitly made available at no cost. Charging for them would have been considered a transgression of the experimental geneticists' moral economy. The same situation later prevailed among molecular geneticists, such as those who were trained in Max Delbrück's "phage group" and other model organism communities and who had become leading molecular biologists by the time of GenBank's creation.[112] Chemists, by comparison, with their long history of close ties to industry, did not tend to value the free distribution of research materials and data as much. The Cambridge Structural Database, the closest equivalent of GenBank for chemists, has charged a fee for access since its inception in 1965, apparently without provoking much resentment among users. The American Chemical Society has also been at the forefront of the opposition to open-access publishing, resulting in the resignation of the molecular biologist Richard J. Roberts, one of its Nobel Prize–winning members.[113] It thus seems to be no historical accident that life

[110] Robert Teitelman, *Gene Dreams: Wall Street, Academia, and the Rise of Biotechnology* (New York: Basic, 1989); Eric James Vettel, *Biotech: The Countercultural Origins of an Industry* (Philadelphia: Univ. Pennsylvania Press, 2006); and, specifically on the *Chakrabarty* case, Kevles, "*Diamond v. Chakrabarty* and Beyond" (cit. n. 96).

[111] Strasser interview with Kirschstein, 22 Feb. 2006; Richard J. Roberts, "The Early Days of Bioinformatics Publishing," *Bioinformatics*, 2000, *16*:2–4; and Bevin P. Engelward and Roberts, "Open Access to Research Is in the Public Interest," *PLoS Biology*, 2007, *5*:e48. See also Fred Turner, *From Counterculture to Cyberculture: Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism* (Chicago: Univ. Chicago Press, 2006).

[112] On the fly group see Kohler, *Lords of the Fly* (cit. n. 12); on the phage group and, more generally, on the cooperative individualism fostered by the Rockefeller Foundation at Caltech see Lily E. Kay, *The Molecular Vision of Life: Caltech, the Rockefeller Foundation, and the Rise of the New Biology* (New York: Oxford Univ. Press, 1993).

[113] Roberts to Dr. [Tamara] Namaroff, 1 June 2005, available at https://mx2.arl.org/Lists/SPARC-OAForum/Message/1977.html (accessed 10 Apr. 2010).

scientists and computer scientists spearheaded the movement toward open access in science.

The creation of GenBank did more than just reflect the current moral economy of the experimental sciences and the culture of computer scientists: it served as a model and as a resource to promote open access to scientific knowledge. In the controversy over the scientific merits of the Human Genome Project, whose proponents were criticized as being a "small coterie of power-seeking enthusiasts," the argument that the results would be made publicly available through GenBank figured prominently in its favor. This argument became even more important after 1998, when the private company Celera entered into competition with the international public consortium to complete the sequencing of the human genome. In that debate, GenBank's open-access policy was heralded by the public consortium as differentiating between the academic and the commercial projects, thus diverting general attention from the methodological criticisms then being made of the public project.[114] Most important, GenBank served to expand open access not just to data such as sequences but to scientific knowledge generally. PubMed Central was created in 2000 as an online archive of freely accessible scientific publications in the biomedical sciences. Its goal has been to make published scientific knowledge as broadly available as possible. The case of GenBank was often used as a model to promote PubMed Central—as, for example, when Roberts and nine other Nobel Prize winners signed a public statement enjoining scientific journal editors to contribute to PubMed Central as "the GenBank of the published literature." In 2008 the NIH agreed on its Public Access Policy, making deposition in PubMed Central one year after publication mandatory for all federally funded research. The policy came into force on 7 April 2008.[115] That same day, in Bethesda, the NIH celebrated the twenty-fifth anniversary of GenBank, a more than timely coincidence.

The second historical transition in the experimental life sciences in the late twentieth century pertains to the changing research practices in the life sciences (and beyond). The argument that GenBank and the numerous other databases that have become indispensable for research in the experimental life sciences represent the outcome of a hybridization between the natural historical and the experimental traditions rests above all on the way these databases have been *used* to produce knowledge.[116] These databases undoubtedly represent an outcome of the experimental tradition, but at the same time they belong to a way of knowing that is perhaps best described as "natural historical," in that it rests on the collection and comparison of natural facts, often across many species.[117] This is not to say that modern databases are identical to the natural historical collections of plants and animals, past and present. A number of significant differences stand out, especially the fact that modern databases store information electronically, allowing data to be massively

[114] Salvatore E. Luria, "Human Genome Program," *Science*, 1989, *246*:873–874 (quotation); and John Sulston and Georgina Ferry, *The Common Thread: A Story of Science, Politics, Ethics, and the Human Genome* (London: Bantam, 2002), Ch. 7.

[115] Richard J. Roberts *et al.*, "Information Access: Building A 'Genbank' of the Published Literature," *Science*, 2001, *291*:2318–2319; and Harold Varmus, "Progress toward Public Access to Science," *PLoS Biol.*, 2008, *6*:e101.

[116] Each year, *Nucleic Acids Research* publishes a special issue devoted to databases in experimental biology. In 2010, over 150 databases were described. See M. Y. Galperin and G. R. Cochrane, "The 2010 Nucleic Acids Research Database Issue and Online Database Collection: A Community of Data Resources," *Nucleic Acids Res.*, 2010, *28*:1.

[117] For a more extensive discussion of this theme see Strasser, "Collecting, Comparing, and Computing Sequences" (cit. n. 33).

compared and widely distributed. But this is a difference more of degree than of kind. The specimens contained in museums of natural history, for example, have certainly been less mobile than electronic data, but they have nevertheless circulated according to the traditional policy of museums to lend specimens to qualified individuals and institutions for study. The key point is that databases and naturalist collections have made possible a similar way of knowing, one that is distinct from the experimental way of knowing.

Whether these databases represent "homologues" (the result of a historical continuity) or "analogues" (the result of a functional convergence) of the natural history museums and other naturalist collections is an important historical question.[118] The lack of connection between the main actors who promoted databases of experimental knowledge and the natural history tradition seems to point to the fact that databases and naturalist collections had different origins but were set up to serve similar purposes. Indeed, by bringing elements to a single place and organizing them in a standardized format, those who established collections made a specific epistemic practice possible: the systematic comparison of elements. This comparative perspective is as central for contemporary molecular biologists who use sophisticated algorithms (such as BLAST) to find similarities among gene sequences in GenBank as it was for anatomists such as Georges Cuvier who looked for morphological similarities among skeletons at the Muséum National d'Histoire Naturelle. The comparative perspective can reveal similarities, differences, and patterns that individual experiments or observations obviously cannot.

It is hard to overemphasize the epistemic difference between the comparative perspective, so essential to natural history, and the "exemplary" perspective, so prominent in the experimental sciences. These two perspectives relied on fundamentally different approaches to make universal claims. In the comparative perspective variations of the natural world could be overcome by systematic comparison across a broad range of species to reveal underlying regularities. The exemplary perspective, inspired by the physical sciences, based universal claims on observations made in a carefully controlled system, such as a single model organism. Such work assumed that "what is true of *E. coli* is true of the elephant," as the molecular biologist Jacques Monod famously put it.[119] Thus it comes as no surprise that when molecular biologists became increasingly interested in employing a comparative perspective to reveal the structure, function, and history of molecular sequences, for example, they established collections like so many naturalists before them. Furthermore, collections of experimental data represent not just a few model organisms but outstrip in size the collections of most natural history museums (in 2011 GenBank included sequences from a third of a million species).

Differences aside, the persistence of this comparative perspective in the long history of the life sciences could actually provide a link between contemporary databases and earlier natural history collections. While collectors of protein and DNA sequences, such as Dayhoff and Goad, might not have recognized any direct connection to natural history, the researchers who produced the data often did. A number of protein sequences were determined by researchers such as Frederick Sanger who had been inspired by the

[118] I thank Robert E. Kohler for this useful analogy.

[119] Jacques Monod and François Jacob, "General Conclusions: Teleonomic Mechanisms in Cellular Metabolism, Growth, and Differentiation," *Cold Spring Harbor Symposia on Quantitative Biology*, 1961, *21*:389–401, on p. 393. On the origins of this expression see Herbert C. Friedmann, "From 'Butyribacterium' to 'E. coli'—An Essay on Unity in Biochemistry," *Perspectives in Biology and Medicine*, 2004, *47*:47–66. In the earlier part of the twentieth century, knowledge derived from animal model organisms was believed to hold true only for animals, not plants and microbes, as would become the case in molecular biology starting in the late 1950s.

particular tradition of "comparative biochemistry." This subdiscipline of biochemistry, which has received scant historical attention, grew in the 1930s as an important alternative to "mainstream" biochemistry, which advocated the experimental study of carefully selected model systems rather than comparisons among numerous species.[120] The main proponents of comparative biochemistry, such as the Dutch, British, and Belgian biochemists Albert Jan Kluyver, Ernest Baldwin, and Marcel Florkin, took advantage of the evolutionary history and systematic relationships of organisms to understand the origins and functions of biochemical mechanisms. Following this approach, biochemists determined protein sequences from related species in order to compare them. They hoped that common sequences that had been preserved in the evolutionary process would indicate the presence of a functionally essential part of the molecule. From the early 1960s, a number of sequences from different species were also produced by biochemists who used molecular tools to revisit the classical problems of natural history in terms of the tenets of "molecular evolution." These researches grew out of the tradition of "experimental taxonomy," another key subdiscipline that, like comparative biochemistry, blurred the boundaries between experimental and natural historical endeavors. As I have argued elsewhere, in the mid-twentieth century the traditional problems of natural history, such as taxonomy and phylogeny, came increasingly to be studied experimentally, as the rich studies by Robert E. Kohler, Joel B. Hagen, and Erika L. Milam, for example, have also made clear.[121]

Both of the historical narratives I have outlined here contrast sharply with the two main stories that were told about the experimental life sciences in the twentieth century: the first centered on the triumph of the experimental sciences and the decline of natural history, often expressed as a story of the life sciences "going molecular"; while the second centered on the radical novelty of *in silico* biology or the transformation of biology into an "information science."[122] Understanding the nature of the relationships between current databases in the experimental sciences and earlier collections in natural history will require more historical research, but one point is already clear enough: the very distinction between experimentalism and natural history does not do justice to the complexity of the historical actors' research practices. Throughout the twentieth century, a number of researchers cultivated links between the two approaches, adopting simultaneously the comparative and the exemplary perspectives; meanwhile, generations of experimentalists on crusade were boasting about the autonomy of the experimental method and its superiority over natural historical practices (also characterized as the superiority of the

[120] See Strasser, "Collecting, Comparing, and Computing Sequences" (cit. n. 33). Comparative biochemistry has not been covered by Robert Kohler or Joseph Fruton, for example, in their histories of biochemistry. See Robert E. Kohler, *From Medical Chemistry to Biochemistry: The Making of a Biomedical Discipline* (Cambridge: Cambridge Univ. Press, 1982); and Joseph S. Fruton, *Protein, Enzymes, Genes* (New Haven, Conn.: Yale Univ. Press, 1999).

[121] Strasser, "Collecting, Comparing, and Computing Sequences"; Bruno J. Strasser, "Laboratories, Museums, and the Comparative Perspective: Alan A. Boyden's Serological Taxonomy, 1925–1962," *Hist. Stud. Nat. Sci.*, 2010, *40*:149–182; Kohler, *Landscapes and Labscapes* (cit. n. 3); Hagen, "Experimental Taxonomy, 1920–1950" (cit. n. 3); Hagen, "Naturalists, Molecular Biology, and the Challenge of Molecular Evolution" (cit. n. 3); and Erika Lorraine Milam, "'The Experimental Animal from the Naturalist's Point of View': Behavior and Evolution at the American Museum of Natural History, 1928–1954," in *Descended from Darwin: Insights into the History of Evolutionary Studies, 1900–1970*, ed. Joe Cain and Michael Ruse (Philadelphia: American Philosophical Society, 2009), pp. 157–178.

[122] Lenoir, "Shaping Biomedicine as an Information Science" (cit. n. 61); and David Baltimore, "How Biology Became an Information Science," in *The Invisible Future: The Seamless Integration of Technology into Everyday Life*, ed. Peter J. Denning (New York: McGraw-Hill, 2001), pp. 43–55.

laboratory over the museum). The hybridization of the two approaches in contemporary research, especially as seen through the use of databases of experimental data, is now too widespread for historians to ignore. Today, natural historical practices centered on the collection and comparison of data are increasingly recognized as legitimate ways to produce knowledge in the experimental life sciences. In this light, the current "hybrid culture" of experimental and natural historical practices might indicate that the triumph of a predominantly experimental tradition, which has defined research in the life sciences from the late nineteenth to the late twentieth century, is progressively drawing to a close.