Journal of the History of Biology (2018) 51:657–691
https://doi.org/10.1007/s10739-017-9490-y

Globalizing Genomics: The Origins of the International Nucleotide Sequence Database Collaboration

HALLAM STEVENS
School of Humanities and Social Sciences
Nanyang Technological University
14 Nanyang Drive #05-07
Singapore 637332
Singapore
E-mail: hstevens@ntu.edu.sg

Abstract. Genomics is increasingly considered a global enterprise – the fact that biological information can flow rapidly around the planet is taken to be important to what genomics is and what it can achieve. However, the large-scale international circulation of nucleotide sequence information did not begin with the Human Genome Project. Efforts to formalize and institutionalize the circulation of sequence information emerged concurrently with the development of centralized facilities for collecting that information. That is, the very first databases build for collecting and sharing DNA sequence information were, from their outset, international collaborative enterprises. This paper describes the origins of the International Nucleotide Sequence Database Collaboration between GenBank in the United States, the European Molecular Biology Laboratory Databank, and the DNA Database of Japan. The technical and social groundwork for the international exchange of nucleotide sequences created the conditions of possibility for imagining nucleotide sequences (and subsequently genomes) as a "global" objects. The "transnationalism" of nucleotide sequence was critical to their ontology – what DNA sequences came to be during the Human Genome Project was deeply influenced by international exchange.

Keywords: Genomics, Databases, Transnational history, GenBank, EMBL-Bank, DNA Database of Japan

Introduction

In 2001, at the celebration marking the conclusion of the Human Genome Project (HGP), Prime Minister Tony Blair spoke about the

The original version of this article was revised due to a retrospective Open Access order.

international dimensions of the project. “Scientists from Japan and Germany, France, China, and around the world have been involved, as well as the United Kingdom and the United States. And this undertaking, therefore, has brought together the public, private and non-profit sectors in an unprecedented international partnership” (NHGRI, 2000). This internationalism was considered symbolically important as well as critical to the project’s ultimate success. The “finished” human genome was not just a product of Anglo-American technoscience, but, at least rhetorically, a global project.

Genomics is increasingly considered a global enterprise – the fact that biological information can flow rapidly around the planet is taken to be important to what genomics is and what it can achieve (Thacker, 2006). However, the large-scale international circulation of nucleotide sequence information did not begin with the HGP. Efforts to formalize and institutionalize the circulation of sequence information emerged concurrently with the development of centralized facilities for collecting that information. That is, the very first databases build for collecting and sharing DNA sequence information were, from their outset, international collaborative enterprises.

Why was such international coordination considered important? Why did biologists perceive the need for it? How did it develop and what challenges did it face? And what implications did it have for the HGP and genomics more generally? This paper attempts to answer these questions by examining the origins of the International Nucleotide Sequence Database Collaboration (INSDC) between GenBank in the United States, the European Molecular Biology Laboratory Databank (EMBL-Bank), and the DNA Database of Japan (DDBJ). Although the INSDC was formalized under this name only in 2005, it emerged from an international collaboration that stretched back over 20 years. Here, I will examine the early parts of this collaboration, stretching from roughly 1979 to the mid-1990s.

The last decade of scholarship in the history of science and technology has increasingly adopted approaches that have been labelled “transnational.” In the introduction to their special issue of the *British Journal for the History of Science*, Turchetti, Herran, and Boudia define this as “producing historical analyses encompassing the integrated study of different forms of global circulation of scientific knowledge and products, including the construction and functioning of international institutional and professional spaces devoted to science” (Turchetti et al., 2012, p. 330). This literature has been interested – *inter alia* – in understanding the impact of transnational relationships and institutions in shaping scientific knowledge – what difference does it make that particular research programs or products are constructed "transnationally"? Zouyue Wang, for example, has argued that the exchanges of Chinese scientists between China and the US in the 1940s and 1950s had significant implications for the "globalization" of American science in cold war (Wang, 2010). "American" science during the cold war was importantly shaped by transnational movements of people, material objects, and ideas.

Adopting this transnational approach, I argue here that the technical and social groundwork for the international exchange of nucleotide sequences created the of possibility for imagining nucleotide sequences (and subsequently genomes) as a "global" objects. DNA and RNA sequences were not automatically the same everywhere and anywhere – they had to be made so via a complex set of negotiations that took place between roughly 1979 (the first meetings around establishing a nucleotide database) and rapid expansion of the HGP in the 1990s. As such, this paper makes the claim that the "transnationalism" of nucleotide sequence was critical to their ontology – *what DNA sequences came to be* during the HGP was deeply influenced by international exchange.

This paper is also a set of first, tentative steps towards producing a more international history of the HGP and a more international history of databases. Most popular and scholarly accounts (Cook-Deegan, 1996; Sulston and Ferry, 2002; McElheny, 2012) of the genome projects give little attention to French, German, Japanese, and Chinese aspects of the project. The vast majority of human genome sequence was produced in the United States and the United Kingdom; China contributed only 1% of the DNA, and Japan merely 6%. But such figures are not representative of the various collaborative and technological contributions made by international scientists, nor of the long term and symbolic significance of international participation.

We also now have a wealth of literature on the history of databases in the sciences (for example, Bowker and Star, 1999; Bowker, 2005; McCray, 2014; Mackenzie et al., 2015). In biology, databases have been described as "communication regimes" (Hilgartner, 1995), as scientific instruments (Hine, 2006), as tools for the "reuse" of data (Leonelli, 2010), as "spaces of convergence" for biology and computing (Chow-White and García-Sancho, 2011), and as form of "theory" for biology (Stevens, 2013). We have learned from these studies that databases are not merely passive information stores but rather that they play a range of critical roles in knowledge formation in a variety of contexts.

One critical feature of databases is their ability to cross space and to make the data within them "travel" (Howlett and Morgan eds., 2010).

There has been some attention given to the multi-sited nature of databases and the importance of long-distance data sharing, especially in genomics. Hilgarter, in particular, has highlighted the vital role of coordination amongst and between genome centers that drove the HGP (Hilgartner, 2004, 2013). Other work has shown how the sharing of data and the making of data and databases has depended on, and been shaped by, differences in research methods and practices, ethos and ethics, and culture at different sites (Leonelli, 2009; Davies et al., 2013; Farquhar and Rajan, 2014).

What the present narrative adds to this literature is not only a detailed account of how this occurred in one specific and important case (the primary DNA sequence libraries) but also an example of how such negotiation, translation, and standardization occurred successfully across national boundaries. One of the major success stories of the HGP has been the development of regimes for "open data" and "open science" (see Ankeny et al., this volume). The origins of INSDC demonstrate how DNA sequencing and databasing was, almost from its inception, considered a global project and that the realization of this transnational sharing was a product of technical and social labour by a specific group of database workers. Moreover, the aim is to show how this transnational process of geographic exchanges and negotiations had *ontological* effects, generating new kinds of *geographically mobile objects* that became central to biology in the twenty-first century.

Establishing Connections

Why did international collaboration emerge around nucleotide sequence databases? The first meetings towards establishing centralized facilities for collecting nucleotide sequences in databases took place at about the same time in both Europe and the US. In March 1979, a National Science Foundation meeting was held at Rockefeller University with the purpose of discussing the feasibility of establishing a database.[1] EMBL (based in Heidelberg, Germany) held a similar meeting at Schönau in April 1980. The Europeans acted faster, hiring the computer scientist Greg Hamm in October 1980 to establish what became EMBL-Bank.[2] But from the beginning of his operation Hamm

[1] Archival sources will be cited in the footnotes throughout. See "Archival Sources" section for full descriptions of archival collections used. "Report from the collaborative meeting: EMBL/DDBJ/Genbank" EMBL, Heidelberg, 24–28 June 1991, p. 5 [ASH/01419].

[2] For a detailed account of the founding of EMBL-Bank see García-Sancho (2012).

was looking for overseas collaborators. Ken Murray, one of the senior scientists involved with EMBL-Bank, wrote to Elke Jordan at the National Institutes of Health (NIH) in July 1980: "we would certainly like to cooperate with those involved in whatever way appears to be the most generally useful."[3] In fact, Hamm's group was already collaborating with Walter Goad's group at Los Alamos National Laboratory (who had established a pilot database) and Richard Grantham at the *Centre d'Evolution Moleculaire* in Lyon.[4] Writing in *Nature* in 1982, Hamm argued that cooperation, not competition, was the way to proceed in DNA sequence databasing (Walgate, 1982, p. 596).

The most obvious reason for this emphasis on collaboration was that the task ahead of Hamm and his colleagues appeared overwhelming. It was taking EMBL-Bank a great deal of effort to collect existing nucleotide sequence information scattered around Europe, and sequence information was rapidly proliferating, making collection and centralization a massive task. "There is certainly enough work for everyone" reported the first issue of EMBL Nucleotide Sequence Data Library News.[5] The reason for this difficulty was because of the way in which most sequences had to enter the database: they had to be keyed in by hand from published scientific journal pages. This was a labour intensive, time consuming, and error-prone. Later, other methods were developed for getting data into the databases, but this step remained a problematic bottleneck for many years.[6]

Another reason for fostering collaboration was a shortage of computing and communication resources. Nucleotide sequence databases aimed to make data available for user-biologists across the globe, yet initially neither GenBank nor EMBL-Bank had the technical capability to achieve this. On-line access to GenBank, for example, was largely limited to those within the US; only by working together was it possible to conceive of making a truly comprehensive and centralized resource available to the widest possible group of biologists.

[3] Correspondence, Ken Murray to Elke Jordan, 3 July 1980 [BEN/01152].

[4] "Nucleotide Sequence Data Library News" No. 1. March/April 1982, p. 3 [CAM/03516] For more on the evolution of the Los Alamos databank into GenBank see Strasser (2011) and Stevens (2013).

[5] "Nucleotide Sequence Data Library News" No. 1. March/April 1982, p. 3 [CAM/03516]. The reasons for the eventual joining of the effort by the Japanese was also the increasing growth of data. Takeo Maruyama, "About the Cooperative Framework of DNA Data Entry" DDBJ Newsletter, No. 6, February 1987 [NIG (trans.)].

[6] Memo Graham Cameron to Lennart Philipson, 28 January 1984 [CAM/03842].

In the year after GenBank's formal establishment in 1982, the European and US databases quickly came to an "informal agreement."[7] To deal with the data-entry problem, each group would be responsible for scanning and entering data from particular journals; the groups would then exchange data by magnetic tape sent through the mail, allowing each group to have the complete set of data. Although such a scheme seems simple enough, it in fact required significant work and negotiation. For one thing, it meant making sequence data exchangeable: data either had to be in the same format, or in a similar enough format that it could be quickly converted from one format to another. The kind of work that this involved is suggested by a 1987 proposal for extending the GenBank contract:

> LANL [GenBank] and EMBL have also come to understand that sustaining and taking advantage of the collaboration requires a significant amount of time and effort… Translation of the data has not been completely susceptible to automation. Thus many entries require final intervention by the destination staff. Though on the scale of a single entry this intervention is a small percentage of the effort required to develop the information "in-house," when applied to roughly half the data coming into the database it amounts to a significant effort. Finally, the concern on the part of both database staffs that we reduce this latter effort has forced both staffs to spend a fair amount of time keeping each other informed of what the other group is planning, and discussing any plans that may cause difficulties.[8]

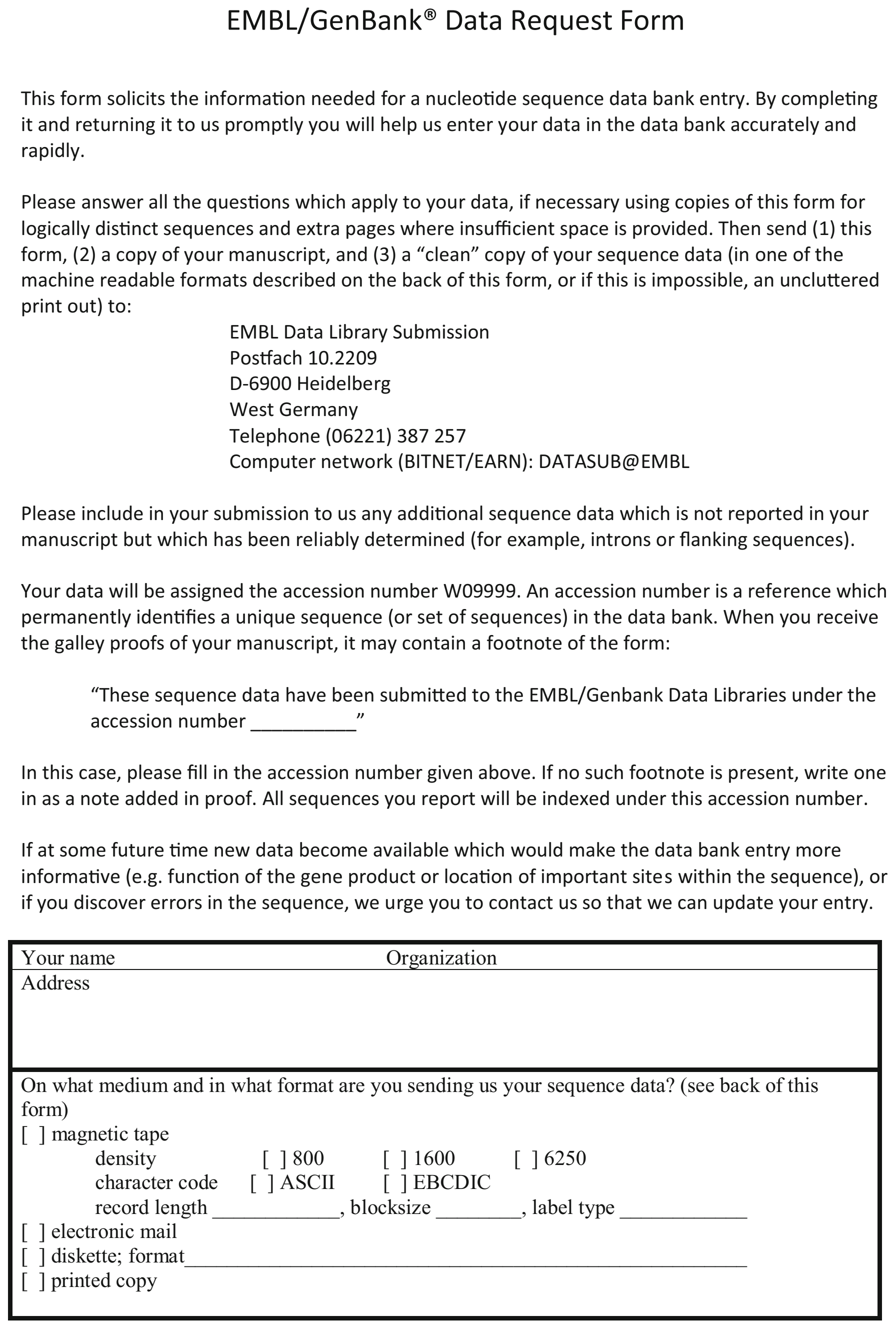
In fact, Graham Cameron, one of the database computer scientists at EMBL estimated that it would take *three months* of programming to generate the software that would convert GenBank's data into EMBL-Bank's format.[9]

Even simple issues such as the numbering of sequences within the database (so-called "accession numbers") proved to be a significant issue. If the databases numbered their entries in different ways it would

[7] James W. Fickett and Christian Burks, "Development of a Database for Nucleotide Sequences," Los Alamos National Laboratory, draft, 24 August 1986, p. 22 [BEN/01091]; "A proposal for the next five years of the GenBank Nucleic Acid Sequence Database" Response to RFP#NIH-GM-97-04, IntelliGenetics, March 1987, p. 7 [BEN/00985]. In the latter source it is labelled a "formal" agreement.

[8] "A proposal for the next five years of the GenBank Nucleic Acid Sequence Database" Response to RFP#NIH-GM-97-04, IntelliGenetics, March 1987, p. 7 [BEN/00985].

[9] Graham Cameron, "EMBL Nucleotide Sequence Data Library: Draft Development Plan," September 1984, p. 9 [CAM/03798].



EMBL/GenBank® Data Request Form

This form solicits the information needed for a nucleotide sequence data bank entry. By completing it and returning it to us promptly you will help us enter your data in the data bank accurately and rapidly.

Please answer all the questions which apply to your data, if necessary using copies of this form for logically distinct sequences and extra pages where insufficient space is provided. Then send (1) this form, (2) a copy of your manuscript, and (3) a "clean" copy of your sequence data (in one of the machine readable formats described on the back of this form, or if this is impossible, an uncluttered print out) to:

EMBL Data Library Submission
Postfach 10.2209
D-6900 Heidelberg
West Germany
Telephone (06221) 387 257
Computer network (BITNET/EARN): DATASUB@EMBL

Please include in your submission to us any additional sequence data which is not reported in your manuscript but which has been reliably determined (for example, introns or flanking sequences).

Your data will be assigned the accession number W09999. An accession number is a reference which permanently identifies a unique sequence (or set of sequences) in the data bank. When you receive the galley proofs of your manuscript, it may contain a footnote of the form:

> "These sequence data have been submitted to the EMBL/Genbank Data Libraries under the accession number __________"

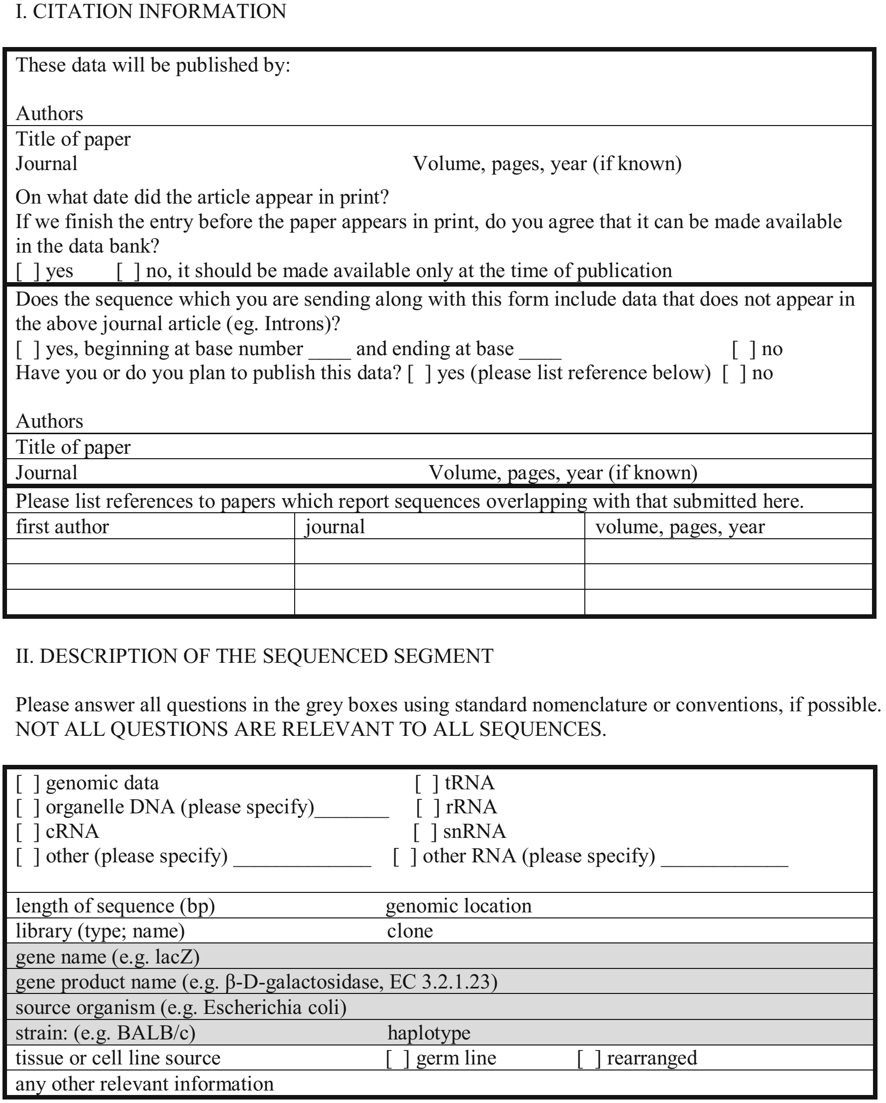
In this case, please fill in the accession number given above. If no such footnote is present, write one in as a note added in proof. All sequences you report will be indexed under this accession number.

If at some future time new data become available which would make the data bank entry more informative (e.g. function of the gene product or location of important sites within the sequence), or if you discover errors in the sequence, we urge you to contact us so that we can update your entry.

Your name Organization
Address

On what medium and in what format are you sending us your sequence data? (see back of this form)
[ ] magnetic tape
density [ ] 800 [ ] 1600 [ ] 6250
character code [ ] ASCII [ ] EBCDIC
record length ____________, blocksize ________, label type ____________
[ ] electronic mail
[ ] diskette; format____
[ ] printed copy

Figure 1. EMBL/Genbank common data request form. Source: Adapted from EMBL/Genbank Data Request Form [CAM/03511]



I. CITATION INFORMATION

These data will be published by:

Authors

Title of paper

Journal Volume, pages, year (if known)

On what date did the article appear in print?

If we finish the entry before the paper appears in print, do you agree that it can be made available in the data bank?

[ ] yes [ ] no, it should be made available only at the time of publication

Does the sequence which you are sending along with this form include data that does not appear in the above journal article (eg. Introns)?

[ ] yes, beginning at base number ____ and ending at base ____ [ ] no

Have you or do you plan to publish this data? [ ] yes (please list reference below) [ ] no

Authors

Title of paper

Journal Volume, pages, year (if known)

Please list references to papers which report sequences overlapping with that submitted here.

first author	journal	volume, pages, year

II. DESCRIPTION OF THE SEQUENCED SEGMENT

Please answer all questions in the grey boxes using standard nomenclature or conventions, if possible.
NOT ALL QUESTIONS ARE RELEVANT TO ALL SEQUENCES.

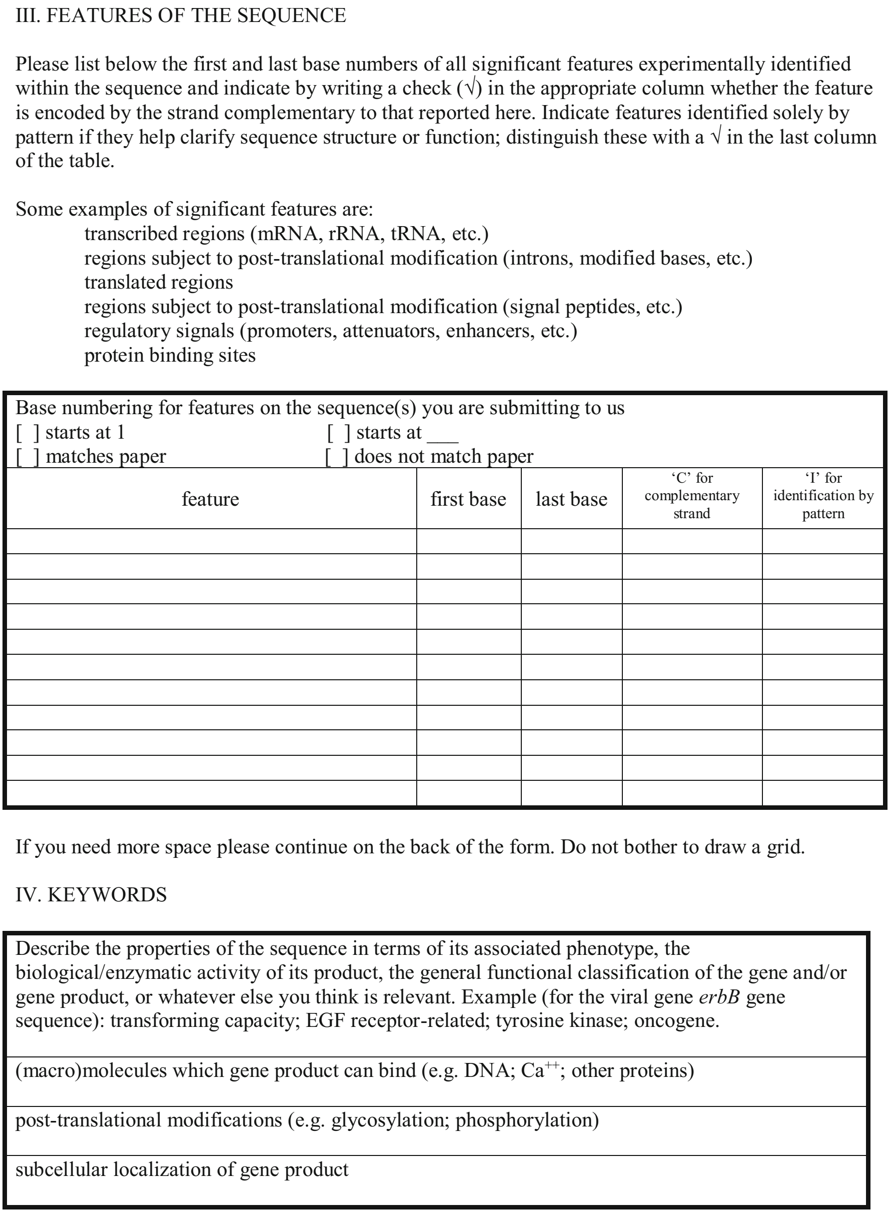
[ ] genomic data [ ] tRNA
[ ] organelle DNA (please specify)_______ [ ] rRNA
[ ] cRNA [ ] snRNA
[ ] other (please specify) _____________ [ ] other RNA (please specify) ____________

length of sequence (bp)	genomic location	
library (type; name)	clone	
gene name (e.g. lacZ)		
gene product name (e.g. β-D-galactosidase, EC 3.2.1.23)		
source organism (e.g. Escherichia coli)		
strain: (e.g. BALB/c)	haplotype	
tissue or cell line source	[ ] germ line	[ ] rearranged
any other relevant information		

Figure 1. continued

be particularly difficult to keep track of which data existed where (especially if numbers were duplicated between the two). GenBank and EMBL-Bank eventually agreed on a scheme to prevent such duplication.[10] The two databases also reached agreement on a data request

[10] "Summary: International Advisory Committee for DNA Sequence Databases," undated, p. 1 [BEN/00657].



III. FEATURES OF THE SEQUENCE

Please list below the first and last base numbers of all significant features experimentally identified within the sequence and indicate by writing a check (√) in the appropriate column whether the feature is encoded by the strand complementary to that reported here. Indicate features identified solely by pattern if they help clarify sequence structure or function; distinguish these with a √ in the last column of the table.

Some examples of significant features are:
transcribed regions (mRNA, rRNA, tRNA, etc.)
regions subject to post-translational modification (introns, modified bases, etc.)
translated regions
regions subject to post-translational modification (signal peptides, etc.)
regulatory signals (promoters, attenuators, enhancers, etc.)
protein binding sites

Base numbering for features on the sequence(s) you are submitting to us
[ ] starts at 1 [ ] starts at ___
[ ] matches paper [ ] does not match paper

feature	first base	last base	'C' for complementary strand	'I' for identification by pattern

If you need more space please continue on the back of the form. Do not bother to draw a grid.

IV. KEYWORDS

Describe the properties of the sequence in terms of its associated phenotype, the biological/enzymatic activity of its product, the general functional classification of the gene and/or gene product, or whatever else you think is relevant. Example (for the viral gene *erbB* gene sequence): transforming capacity; EGF receptor-related; tyrosine kinase; oncogene.

(macro)molecules which gene product can bind (e.g. DNA; Ca^{++}; other proteins)

post-translational modifications (e.g. glycosylation; phosphorylation)

subcellular localization of gene product

Figure 1. continued

form that would be sent to biologists to request sequence information – this was a form that could be sent to any lab known to have published a DNA or RNA sequence, requesting a clean or soft copy of the sequence information (Figure 1). Sharing this form meant that, the two databanks were at least *collecting* the same data and metadata from biologists, even if that data did not end up being represented in the same way in each database.

In the first years of their existence the nucleotide sequence databases, driven largely by the pressures imposed by data entry, worked together to resolve a number of issues that allowed them to share information. This took both technical and social work – programming software to convert data in one format to data in another, but also meeting and communicating about how each group was approaching the various problems that they encountered. ''Keeping each other informed'' and ''discussing any plans that may cause difficulties'' became a critical part not only of working together, but of working out how to build a centralized biological resource.

DNA Database of Japan

By 1980, molecular genetics – including DNA sequencing – was a large and important field in Japan (Obayashi, 1986, Uchida, 1993). The possibilities for economic development based on molecular biology and biotechnology were becoming evident and Japan was well positioned. Akiyoshi Wada's work had already established Japan's lead in automating DNA sequencing.[11] Many government agencies, including the Ministry of Education, the Ministry of Health and Welfare, the Ministry of International Trade and Industry, and the Ministry of Agriculture had an interest in developing a Japanese DNA database.[12] Many molecular biologists in Japan were also keen to see a Japanese database established, with some having established their own small databases and software tools for this purpose.[13]

Japanese scientists had also been involved in some of the earliest discussions about the creation of DNA data banks. In the summer of 1982, Tatsuo Oi (University of Kyoto) and Y. Fushimi attended the DNA database meeting organized by Walter Goad in Aspen, Colorado. The delegates discussed the possibility of Japan creating an independent database and joining the EMBL-GenBank partnership (Kanehisa and

[11] For more information on Wada's work see Kishi (2004).

[12] ''The EMBL/NIH Workshop'' DDBJ Newsletter, No. 7, November 1987 [NIG (trans.)].

[13] For example, the work of Satoru Kuhara and Katsuya Hayashi at Kyushu University on the GENAS database (Kuhara and Hayashi, 1984).

Oi, 1994). Together with Minoru Kanehisa (who worked with Goad at Los Alamos), Oi and Fushimi convinced Walter Goad and Greg Hamm to write letters to Wada encouraging the establishment of a Japanese DNA databank.[14] As the chairman of the scientific committee for DNA research promotion at the Japanese government's Agency of Science and Technology, Wada could help to secure funding for a Japanese databank.

Although enthusiasm for establishing an independent Japanese repository was strong, progress was slow.[15] In 1983, representatives from the Ministry of Education, the Agency of Science and Technology and the Science Council of Japan provided some funds to Haruo Ozeki to examine the databank issue. Ozeki recommended that a databank be established at a permanent research center with the capacity to conduct independent biological research. In the following year, this resulted in the setting up of a "DNA data bank steering committee" (led by Hisao Uchida at the Institute of Medical Science at the University of Tokyo) to begin planning and construction of a databank.[16] The first version of the DDBJ was established in Oi's laboratory at the Institute for Chemical Research at Kyoto University. The DDBJ committee, however, felt that a national laboratory would form a more appropriate home for the databank. In early 1984, it was decided that DDBJ would be located at the National Institute of Genetics (NIG) in Mishima as part of its newly established Genetic Information Research Center.[17]

Since DDBJ came late to the game, it aimed to closely follow the examples set by its American and European counterparts. In particular, it hoped to make as much as possible of GenBank's and EMBL-Bank's data more readily available to Japanese scientists and also provide a repository for DNA sequence published in Japanese journals and graduate theses (that could then be shared with the other databanks). The limited funding available for DDBJ also meant that it could not afford to do much more than copy GenBank and EMBL-Bank's efforts on a smaller scale.[18] In particular, "after multiple discussions with the

[14] Interview with Minoru Kanehisa, Tokyo, December 2015.

[15] "Japan is massively behind the United States in the field of basic research in biotechnology… Unfortunately, the decision to establish a DNA database in Japan has still not been approved on the national level" (Kanehisa, 1983, p. 1531).

[16] For the history of DDBJ see http://rgm22.nig.ac.jp/mediawiki-ogareport/index.php/History_of_DDBJ. In preparing this account I also had access to an unpublished account of DDBJ's history written by Machiko Itoh of NIG.

[17] Takeo Muruyama, "Preface" DDBJ Newsletter, No. 6, February 1987 [NIG (trans.)].

[18] Interview with Sanzo Miyazawa, personal communication, December 2015.

database committee," DDBJ decided to officially adopt the GenBank format for its data.[19] Even so, adapting the format to DDBJ was not entirely straightforward:

> The most time consuming process of the data entry was creating data annotations… Hence details of the coding must be provided in the form of manuals. Since the GenBank manual was deemed insufficient as coding information was lacking, special coding manuals must be created by referencing those of GenBank and EMBL.[20]

Lack of resources also meant that the formal establishment of the database was delayed for several years. The Ministry of Education finally appropriated funds for DDBJ's operating costs (as well as a professorship and an assistant professorship) in 1986.

Nevertheless, the project continued to encounter difficulties. First, Miyazawa and others involved had a hard time finding partner to which they could outsource the database work. After an extensive search, they engaged Hitachi's software engineering division to provide help with data entry.[21] Second, the NIG's structure and internal rules initially made it difficult for DDBJ to hire full time programmers or engineers. During 1986, responsibility for the database was assumed by the Evolutionary Genetics Research Department, who could hire computer personnel.[22] By 1987, DDBJ made its computer resources available for use to Japanese biologists and the first full release of DDBJ data occurred in July.[23] The released contained 66 sequences and 108970 nucleotide base pairs.[24] Although this was a small contribution, DDBJ's efforts were rapidly expanding such that within a year they were collecting three percent of the all sequences produced worldwide.

In February 1987, Japanese database scientists including Uchida, Kanehisa, and Takeo Maruyama attended a 3-day workshop on "Future Databases for Molecular Biology" sponsored by the NIH and EMBL. At this meeting, representatives of all the databases reaffirmed their commitment to international collaboration. With the accelerating

[19] Oi's original Japanese database had also used GenBank format and as such this policy also made the database consistent with his format. Sanzo Miyazawa, "The Start of DNA Data Entry," DDBJ Newsletter, No. 6, February 1987 [NIG (trans.)].

[20] *Ibid.*

[21] *Ibid.*

[22] Takeo Muruyama, "Preface," DDBJ Newsletter, No. 6, February 1987 [NIG (trans.)].

[23] Interview with Sanzo Miyazawa, personal communication, December 2015.

[24] "DNA Databank of Japan. Release 1.0," DDBJ Newsletter, No. 7, February 1987, p. 14 [NIG (trans.)].

pace of DNA sequencing, continued collaboration was not only considered a practical necessity for the databases, but also scientifically and philosophically desirable. Funders (such as the NIH, EMBL, and Japanese government ministries) saw databases as part of scientific infrastructure rather than fundamental research. This made it more desirable to foster cooperation rather than competition.[25]

This continued commitment to collaboration resulted in the establishment of an international advisory committee to oversee the cooperation between the databases.[26] This panel would advise the databases on how to further their abilities to work together. In November 1987, Sanzo Miyazawa represented DDBJ at the joint databanks meeting held at IntelliGenetics in Mountain View, California.[27] By this time, DDBJ – despite the small size of its operations – had firmly established itself as a third leg of an international DNA database consortium.

Resolving Problems

In February 1988, the International Advisory Committee – consisting of three representatives each from the US and UK and two from Japan – met for the first time.[28] At this summit, the committee praised the databases for their success in coming to agreements about accession numbers and data request forms, but noted that there were still fundamental differences between the databases that needed to be resolved. Describing these differences, and how they were tackled, in some detail will suggest the role that collaboration came to play in the constitution and re-constitution of DNA sequences themselves as electronic objects.

The most important difference between EMBL-Bank and GenBank was the so-called "features table" in each database. This table contained "annotations" of the sequence data (now usually called "metadata"). For instance, if a piece of nucleotide sequence in the database contained

[25] For discussion of funding in the USA, Europe and Japan as well as the relationship between databases and research see "EMBL/NIH Database Workshop," DDBJ Newsletter, No. 7, February 1987 [NIG (trans.)].

[26] DDBJ News Letter No. 7, 1987 [NIG]. Correspondence, Dieter Soll, Ruth Kirschstein, Lennart Philipson, and Hisao Uchida to Editor, Science, 1987 [BEN/01472].

[27] Interview with Sanzo Miyazawa, personal communication, December 2015. On the Japanese joining the database collaboration see also "Collaboration between the EMBL Data Library and GenBank: EMBL's expectations for the Future of the Collaboration," 12 March 1987, p. 5 [CAM/03773].

[28] DDBJ News Letter No. 8, 1989 [NIG]. See also Soll et al. (1988).

a gene, the features table might contain information about where that gene started and ended, what protein the gene coded for, and where to find more information about that protein. The databases differed significantly in what and how information was stored in their respective features tables (Figure 2).[29] This made it difficult to convert entries between the two databases and limited the usefulness of any ''converted'' data.

These differences also pointed towards more fundamental disagreements between the databases. In particular, GenBank and EMBL-Bank had discussions about whether information about proteins corresponding to gene-coding sequences should appear in the table: ''Should we annotate signal peptide and/or mature peptide boundaries, annotated glycosylation sites, disulphide bond sites, etc.? And even if we allow for this information, do we want to take on the responsibility of annotating it consistently and completely?''[30] The databases also worried about how closely to distinguish between different kinds of objects such as DNA, intermediate RNA, and product RNA or proteins. Should these all be labelled differently within the features table?[31] Should the table distinguish between experimentally determined and computationally-determined protein-coding regions? Should features such as alpha helices and beta sheets be annotated? Was there some general philosophy guiding what sorts of features to include or exclude? Such dilemmas proliferated. At the root of such issues were differences of viewpoint about the appropriate division of labor (who should bear responsibility for annotation – lab biologists or database curators?), the validity of computational methods (should computationally-determined genes be included?), the most important uses and users of the database (which features should be annotated?), and authority (who got to decide which features should be annotated?).

The technical and social *work* of resolving these issues was in fact the work of figuring out what a DNA sequence entry looked like in the database. This work took the form of exchanging letters, emails, and reports, as well as collaborative meetings and exchanges of personnel. An EMBL report from March 1987 noted the significance of Christian Burks's (GenBank project leader at Los Alamos) five week visit to EMBL during the fall of 1986. This was the ''start of a regular exchange scheme in which staff from each group would work with the other group

[29] G. Cameron et al. ''Feature Representation in the EMBL and GenBank Nucleotide Sequence Data Libraries,'' [draft] 20 March 1987, pp. A1, B1 [CAM/03598].

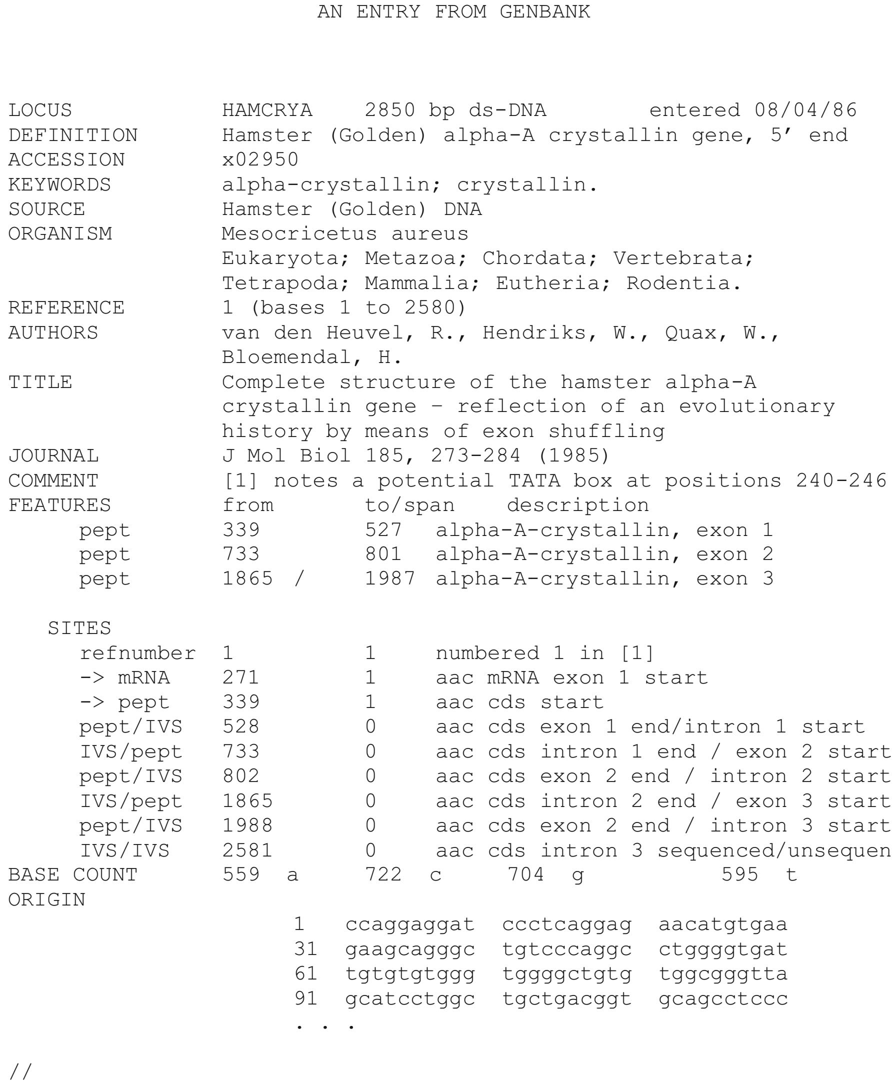
[30] *Ibid.*, pp. C2, C3[CAM/03600].

[31] *Ibid.*, p. C1 [CAM/03599].

```
                    AN ENTRY FROM THE EMBL DATA LIBRARY


ID   MAAACRY1     standard; DNA; 2850 BP.
XX
AC   x02950
XX
DT   13-NOV-1985     (first entry)
XX
DE   Hamster –A crystallin gene 5′ part (exons 1-3)
XX
KW   crystallin; alpha-crystallin
XX
OS   Mesocricetus aureus (golden hamster, hamster, Goldhamster)
OC   Eukaryota; Metazoa; Chordata; Vetebrata; Tetrapoda
OC   Mammalia; Eutheria; Rodentia
XX
RN   [1]  (bases 1-2580; enum. 1 to 2580)
RA   van den Heuvel R., Hendriks W., Quax W., Bloemendal H.;
RT   “Complete structure of the hamster alphaA crystallin gene –
RT   reflection of an evolutionary history by means of exon
RT   shuffling”;
RL   J. Mol. Biol. 185: 273-284 (1985)
XX
FH   Name       Key        Location
FH
FT   aacry.prm  TATA       (240, 246)
FT   aacry.m1   mRNA       (271, 527)
FT   aacry.iv1  IVS        (528, 732)
FT   aacry.m2   mRNA       (733, 801)
FT   aacry.iv2  IVS        (802, 1864)
FT   aacry.m3   mRNA       (1865, 1987)
FT   aacry.iv3  IVS        (1988, >2850)
FT   aacry.c1   CDS        (339, 527) /note=“alpha A crystallin”
FT   aacry.c2   CDS        (733, 801) /note=“alpha A crystallin”
FT   aacry.c3   CDS        (1865, 1987)/note=“alpha A crystallin”
XX
SQ   Sequence 2850 BP; 559 A; 722 C; 704 G; 595 T;
     ccaggaggat  ccctcaggag  aacatgtgaa  gaagcagggc  tgtcccaggc
     ctggggtgat  tgtgtgtggg  tggggctgtg  tggcgggtta  gcatcctggc
     . . .
//
```

Figure 2. Flat file formats for EMBL-Bank (above) and GenBank (overleaf). The features tables for each are shown towards the bottom of each entry, above the sequence en-try itself. Source: Adapted from G Cameron et al. “Feature Representation in the EMBL and GenBank Nucleotide Sequences Data Libraries,” 20 March 1987, pp. A1, B1 [CAM/03589]



```
                         AN ENTRY FROM GENBANK


LOCUS            HAMCRYA    2850 bp ds-DNA       entered 08/04/86
DEFINITION       Hamster (Golden) alpha-A crystallin gene, 5' end
ACCESSION        x02950
KEYWORDS         alpha-crystallin; crystallin.
SOURCE           Hamster (Golden) DNA
ORGANISM         Mesocricetus aureus
                 Eukaryota; Metazoa; Chordata; Vertebrata;
                 Tetrapoda; Mammalia; Eutheria; Rodentia.
REFERENCE        1 (bases 1 to 2580)
AUTHORS          van den Heuvel, R., Hendriks, W., Quax, W.,
                 Bloemendal, H.
TITLE            Complete structure of the hamster alpha-A
                 crystallin gene - reflection of an evolutionary
                 history by means of exon shuffling
JOURNAL          J Mol Biol 185, 273-284 (1985)
COMMENT          [1] notes a potential TATA box at positions 240-246
FEATURES         from       to/span    description
     pept        339        527   alpha-A-crystallin, exon 1
     pept        733        801   alpha-A-crystallin, exon 2
     pept        1865  /    1987  alpha-A-crystallin, exon 3

   SITES
     refnumber   1          1     numbered 1 in [1]
     -> mRNA     271        1     aac mRNA exon 1 start
     -> pept     339        1     aac cds start
     pept/IVS    528        0     aac cds exon 1 end/intron 1 start
     IVS/pept    733        0     aac cds intron 1 end / exon 2 start
     pept/IVS    802        0     aac cds exon 2 end / intron 2 start
     IVS/pept    1865       0     aac cds intron 2 end / exon 3 start
     pept/IVS    1988       0     aac cds exon 2 end / intron 3 start
     IVS/IVS     2581       0     aac cds intron 3 sequenced/unsequen
BASE COUNT       559  a     722  c     704  g          595  t
ORIGIN
                      1   ccaggaggat  ccctcaggag  aacatgtgaa
                      31  gaagcagggc  tgtcccaggc  ctggggtgat
                      61  tgtgtgtggg  tggggctgtg  tggcgggtta
                      91  gcatcctggc  tgctgacggt  gcagcctccc
                      . . .

//
```

Figure 2. continued

for more extended periods."[32] Such exchanges, meetings, and discussions were crucial for resolving the complicated issues concerning the representation of features in the databases.

[32] "Collaboration between the EMBL Data Library and GenBank: EMBL's expectations for the Future of the Collaboration," 12 March 1987, p. 3 [CAM/03773]. DDBJ also participated in joint staff meetings from November 1987. Sanzo Miyazawa and Hidenori Hayashima, "DDBJ Activity Report 1989" DDBJ Newsletter, No. 9, May 1990 [NIG (trans.)].

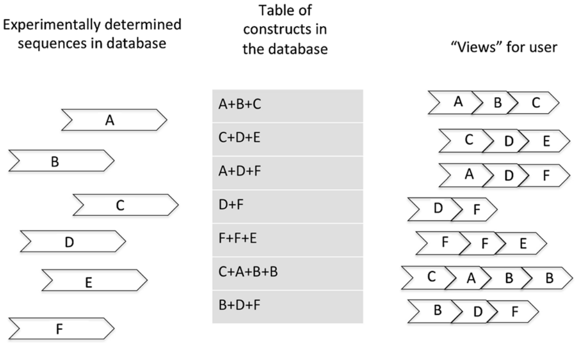


Figure 3. Representation of sequences in databases. As nucleotide sequence database evolved, the experimentally determined sequences assumed less and less importance. Rather, for biologist-users, data was represented as "views" of chromosomes of genomes constructed from multiple pieces sequence

In a long appendix to a document "attempting to draw together the ideas for a common feature table as discussed in a two week meeting in May 1986" Burks summarized his team's technical work to develop an common features table.[33] But the tone and substance of document makes clear that this was the product of intense discussion and negotiation between the two database teams.[34] The EMBL team responded in a further appendix: "EMBL is still interacting with GenBank about [feature table] keys… We agree pretty much with the subject matter that Christian [Burks] has covered with the keys, though we might like to suggest some modifications to the actual terms used and to the family organization… [For example,] We don't understand the proposed qualifiers /alpha, /number, /label, and /text. The /note qualifier proposed at Tyson's [Corner meeting] seems enough."[35] These kinds of discussions suggest how these negotiations were carried on via both face-to-face meetings and in more formal communications. What constituted an "entry" or a "sequence" in the nucleotide databases was not obvious or given, but had to be constituted through technical and social work *between* EMBL-Bank and GenBank.

Relations

The discussions over the features table were just a first step towards making the databases fully interoperable and compatible. As the data-

[33] G. Cameron et al. "Feature Representation in the EMBL and GenBank Nucleotide Sequence Data Libraries," [draft] 20 March 1987 [CAM/03589].

[34] *Ibid*., pp. C2–C5 [CAM/03600 and CAM/03601].

[35] *Ibid*., p. D1 [CAM/03607].

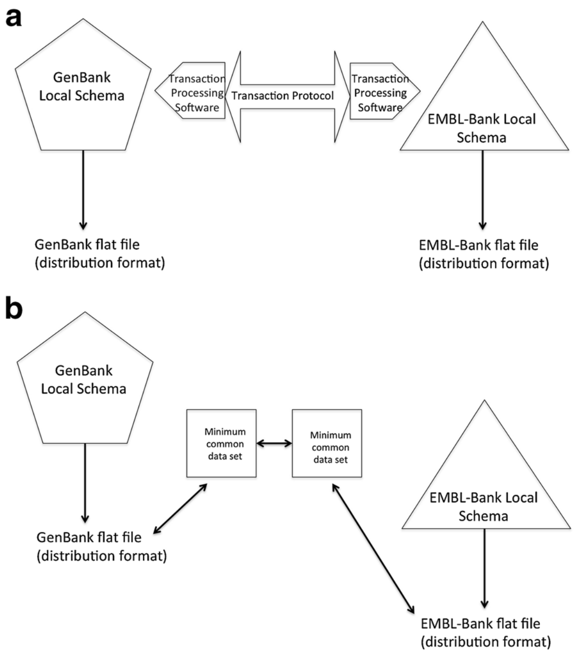

a
GenBank Local Schema
Transaction Processing Software
Transaction Protocol
Transaction Processing Software
EMBL-Bank Local Schema
GenBank flat file (distribution format)
EMBL-Bank flat file (distribution format)
b
GenBank Local Schema
Minimum common data set
Minimum common data set
EMBL-Bank Local Schema
GenBank flat file (distribution format)
EMBL-Bank flat file (distribution format)

Figure 4. (a) Transaction protocol. The transaction protocol was designed to coordinate the data in the two databases. Each database maintained its own local schema. The data in this schema would be translated into the transaction protocol by the "transaction processing software." Data would then be communicated from one database to another using the transaction protocol. (b) Common schema. As an alternative to the transaction protocol, data from each data's flat file distribution could be abstracted into a "minimum common data set." This data could then be exchanged between the two databases

bases became more sophisticated and the amount of data contained within them increased, there was increasing need for automatic procedures that would coordinate updates between the two databases. Solving these problems entailed converting each database to a "relational database management system" (RDBMS) and developing a "transaction protocol" for exchanging data between them.

From their inception, both GenBank and EMBL-Bank had utilized so-called "flat file" databases. In a flat file, each entry (corresponding to a piece of DNA sequence) is appended one after another in a long list (the entries shown in Figure 2 are in flat file format). This kind of database is easily read by humans, but inefficient for retrieving and updating data. For example, searching the database for a single piece of data requires searching through the database from beginning to end.

Likewise, updating a flat file database to include a new piece of information involves modifying every single entry in the database one by one.[36] By 1987, as GenBank transitioned to a new contract and new management, moving from flat file to relational database was one of the most important tasks at hand.[37]

Such a move was necessitated by both the ever-increasing volume of submitted sequence data as well as the specific uses to which sequence data was being put by biologists. Increases in the speed and efficiency of laboratory sequencing methods meant that the amount of nucleotide sequence that needed to be databased continued to increase exponentially. Biologists most often used nucleotide databases by searching for matches between sequences found in the lab and sequences found in the database. Such matches could not only determine the novelty of lab-determined sequences, but also suggest important information about sequence function (for example, if a new human DNA sequence largely matched a known mouse gene, it was likely to be a human gene). This meant that the ability to rapidly and efficiently search – using specialized bioinformatics software – for nucleotide sequences in a database was critical for the biological community (Stevens 2011). An RDBMS could offer increased sophistication and speed in such searches.

Although both GenBank and EMBL-Bank agreed that converting their databases to a RDBMS was desirable, they disagreed about how exactly this should be achieved. Moreover, staff at both databases realized that making their relational database schemas *compatible* was going to be critical to future collaboration. Indeed, further collaboration and data-sharing was one of the main drivers for implementing a relational system. "The goal of autoconvertability," EMBL staff argued, "will be rendered more approachable by anticipated changes in data management procedures…"[38] Reworking nucleotide data into a new (relational) format meant a new chance to collaborate on developing "underlying data structures" that would facilitate compatibility.[39]

In particular, implementing an RDBMS would allow the databases to disentangle some tricky issues that had hampered collaboration.

[36] For more on the relationship between flat files and relational databases see Haigh (2004); for more on the transition in the context of GenBank see Stevens (2013).

[37] The initial contract for GenBank ran from 1982 to 1987. This contract was serviced by BBN in partnership with Los Alamos National Laboratories. The 1987 contract renewal was won by IntelliGenetics, Inc., again in collaboration with Los Alamos. See Stevens (2013) for more details.

[38] "Collaboration between the EMBL Data Library and GenBank: EMBL's expectations for the Future of the Collaboration," 12 March 1987, p. 4 [CAM/03772].

[39] *Ibid.*

Under the flat file scheme, the representation of the database in the computer system was identical to the way in which the data was distributed to its users (biologist users were simply mailed a magnetic tape containing the flat file). An RDBMS, however, meant that the representation in the database (the tables where data was stored and manipulated) could be conceptually distinct from the "distribution format."[40] The flexibility afforded by an RDBMS meant that each database could keep (and continue to produce) its own "distribution format" for its users while modifying the underlying relational schema.

This opened up new ways of organizing and representing sequences in the databases. A report of a joint NIH/EMBL meeting from 1987 outlined the existing practice:

> It has been taken for granted in the past that we should attempt to reflect the sequence data as they occur in nature. This implies, among other things, that where we encounter publications that present overlapping stretches from the same genome we should 'merge' the sequence entries in the data library to produce one large entry. While at first sight this appears logical, it is not the only approach.[41]

Graham Cameron, in particular, was thinking about alternative ways to represent sequence in the databases that would allow greater interoperability. Sharing data between databases required some form of standardization, but attempting to coordinate the underlying structures of databases had proved too difficult. Instead, Cameron suggested, "Database designers should more concentrate their energies on the definition of standards about operations that can be performed on the databases. That is, we should define standards about *messages to* the databases…"[42] The result of this thinking was a system in which "views" of the database – what users could see when they accessed the data – were distinct from the ways in which the data was actually stored in the database. Data was stored as a "chunks" of sequence submitted by experimenters, but genes, chromosomes, genomes, and other objects of biological interest could be constructed "on top" of these via "machine processable instructions"

[40] "Collaborative Meeting: DDBJ: The EMBL Data Library: GenBank: Report" EMBL, Heidelberg, 5–15 September 1988, p. 6 [NHGRI/0141-009, p. 50].

[41] "EMBL/NIH Workshop: Future Databases for Molecular Biology (Some preparatory thoughts from the EMBL Data Library)," EMBL Heidelberg, 25–27 February 1987, p. 9 [CAM/03914].

[42] *Ibid.*, p. 5 [CAM/03609]. My emphasis.

(Figure 3).[43] "Sequences" – in the sense that biologists usually thought of them and used them – were represented not as discrete elements in the database, but rather as a set of instructions for building the "sequence" from overlapping smaller chunks. Sharing demanded increased flexibility in the databases, but that flexibility meant that "sequences" – as objects stored in database – became conceptually distinct and abstracted from "sequence" as usually used or understood in the lab. The need to create more mobile sequence objects not only made an RDBMS desirable to database managers, but also changed how sequences were represented within the nucleotide databases.

Making sequences global

Even more importantly, an RDBMS opened up the possibility of using a "transaction protocol" or a distinct language that could be used to communicate data bewteen the databases. By 1988, the key issue for collaboration between the databases became settling the issue of a suitable mode of communication.[44] Each database could maintain its own underlying structure and its own distribution format, so long as they agreed on a transaction protocol (Figure 4a). As one EMBL report put it:

> Even if at one instant we were to have identical copies of the data at all sites, the problem would not be solved. We would still have to communicate all new data and changes to all sites. Thus far data have been exchanged by exchanging normal releases of the database every three months. Relatively smooth automatic procedures have been developed for including all the new data at each release, but no automated systems for transmitting updates exist at present. This inability to propagate updates coupled with the delays of the three monthly release cycle render the present system unsatisfactory. All groups recognize this, and agree that the solution to this problem should be through a *transaction protocol* – a common language to communicate changes to the database in real time.[45]

Although GenBank and EMBL-Bank agreed that a transaction protocol was necessary, they disagreed strongly about the form that such a protocol should take. The Americans proposed that a language (for the

[43] *Ibid*., p. 9 [CAM/03914].

[44] Graham Cameron, "The GenBank Transaction Protocol – The EMBL View" (EMBL Data Library Discussion Document), undated [CAM/03543].

[45] *Ibid*., p. 1 [CAM/03543].

transaction protocol) needed to be developed from scratch – to be built uniquely for the biological database. The Europeans argued that an existing database language, such as SQL, should be implemented.

In an email to Christian Burks, Graham Cameron at EMBL wrote: "The reason we are bitching about TP design as it stands is that we (or maybe you should read I) actually don't think it can be made to work and keep working without the application of resources out of all proportion to the utility it provides. We think the argument applies both sides of the [A]tlantic. We don't want to turn this into an EMBL/GenBank battle, but I guess the bottom line is that we think it is ill conceived and will fail."[46] Part of the problem was resources: GenBank was better resourced than its counterpart and able to invest more in developing their own software and tools. But more deeply than this, the disagreement suggests that EMBL saw the problem of database communication as a specific instance of a more general type of problem for which there was no need to invent specific new tools. GenBank, on the other hand, believed that biological data was mismatched with existing database and informatic tools and needed to create something more specifically "biological."

In September 1988, David Hazledine (at EMBL) wrote an email to GenBank and DDBJ with the subject: "Transaction Protocol – Have we got it wrong?": "The argument in favour of a network schema [GenBank's model] (I think) is that it is a more abstract representation of the data than a relational schema, and it can thus be structured to correspond more closely to the biologist's notion of reality…."[47] For GenBank, the database representation of sequences should necessarily be shaped by some connection to biological concepts. But Hazledine went on: "WHICH biologist's notion of reality? And what grounds do we have for believing that OUR notion (which is, after all, what the schema would reflect) is the same as their's?"[48] This exchange suggests how resolving differences between two databases involved confronting fundamental issues about *what a sequence was* for biologists.

EMBL-Bank and GenBank had very different philosophies about *how* to represent sequences in a database and again these differences had

[46] Email correspondence, Graham Cameron to Christian Burks, 13 December 1988 [CAM/03537].

[47] Email correspondence, David Hazledine to "NUCORE, LANL, IG, DDBJ, DATALIB," 28 September 1988 [CAM/03538].

[48] *Ibid.*

to be worked through in order to make the international collaboration possible.[49] Smith (1998) has argued that making decisions about the representations and relations of objects inside software and databases is doing more than merely computational work. Making such decisions about how to structure objects inside computers, Smith says, has important implications for what those objects actually are *in the world*. Programming and databasing have *ontological* implications. Here, it was precisely the international exchanges between different databases that brought these issues to the forefront, opening up debates about what objects *should* look like and *how* they should be represented.

These differences were resolved, at least in part, by the development of a so-called "common schema." At issue here was the fact that none of the databases wanted to give up their own internal representation of the data – that is EMBL did not wish to literally copy or convert their database to GenBank and GenBank did not want to simply reproduce EMBL-Bank at a different site. Each believed that their own database had some superior (or at least importantly unique) features. This issue was raised explicitly in the joint database meeting held in 1989 where different models of collaboration were discussed. The options of having one database hold the "definitive copy" of the data or holding a definitive copy at a "remote site" were considered "politically unrealistic." The only realistic solution was the "multiple copies" model (in which identical copies were held at different locations) but this was hampered by problems of "making sure that every update is propagated to and installed in each database."[50]

The eventual solution to this problem was to agree on a "higher level" representation of the data that included some subset of "entities" that both databases could agree were important. Such a "common schema" could be distinct from, but easily convertible to, the "local schema" for each database. In the "common schema," sequence data

[49] These difficulties are suggested by the following kinds of problems: "Extracting a unique, but complete set of information from the two databases is a non-trivial task. Where references A, B, and C report contiguous sequences, they may appear in one database with an entry composed of the data from A and B merged, and in the other with B and C merged. Even where a simple one-to-one correspondence can be identified, it is not simply a matter of taking the "best" version, one database may have a more up-to-date version of the sequence, while the other has annotated the features of the sequence much more thoroughly." "Collaborative Meeting: DDBJ: The EMBL Data Library: GenBank: Report," EMBL, Heidelberg, 5–15 September 1988, p. 11 [NHGRI/0141-009, p. 55].

[50] "Report on GenBank/EMBL/DDBJ Collaborative Meeting," Mishima, Japan, 19–23 June, 1989, p. 3 [NHGRI/0141-008, p. 117].

was represented in a way that was divorced from any one group's notion of what sequence data should be or should look like; rather it consisted of the minimum information (or "minimum common data set") required to make the databases consistent with one another (Figure 4b). Although this data may not have provided *all possible* information about a piece of sequence, the purpose of this common data set was to "ensure that all scientists throughout the world have access to the same complete set of sequence information."[51]

Although GenBank and EMBL-Bank agreed to disagree about the transaction protocol, the notion of the "common schema" allowed them to find a way of converting between the two databases. In practice, this meant that GenBank, EMBL-Bank, and DDBJ continued to exchange flat files that could be used to automatically update critical information in each database and make it consistent with the others. This was achieved through creating a complicated system of internal version numbers for each database.[52] In practice, the data elements included in this "common schema" came to define what a sequence was a global object. Sequences were less what was represented at GenBank, or EMBL-Bank, or DDBJ, but rather more *what was exchanged in common between them*. "The same complete set of sequence information" that the databases were trying to provide to scientists everywhere in the world was defined by the common schema that was worked out between the databases.

The collaboration between the databanks had important implications for how each databank worked and what it contained. As Graham Cameron wrote to Burks: "That we routinely (and largely automatically) transfer data between the two collections is now taken for granted. Some years ago our formats and philosophies differed enough to make even data transfer difficult. The work savings achieved by this data transfer are enormous. One should not, however, allow these very tangible benefits to lead one to ignore other, extremely important, benefits of our collaboration. The joint design work has not just brought the two databases closer together, it has resulted in great improvements to both."[53] In other words, the collaboration significantly changed how sequence was represented in the databases and therefore how it was understood by biologists on both sides of the Atlantic. The difficulties of

[51] "Collaborative Meeting: International Nucleotide Sequence Database: DDBJ/EMBL/GenBank," National Institute of Genetics, Mishima, Japan, 18–22 May, 1992, p. 8 [ASH/04798].

[52] "Report from the collaborative meeting: EMBL/DDBJ/Genbank," EMBL, Heidelberg, 24–28 June 1991, pp. 13, 17 [ASH/04627 and ASH/04631].

[53] Correspondence, Graham Cameron to Christian Burks, 21 January 1987 [NIH/01135].

exchanging information forced both databanks to closely confront and refine *how* they represented sequence objects.

Coordination between the databases also meant that sharing data in a timely manner became a serious issue. At first, data could be shared on magnetic tapes and then imported into local databases; this might work well enough if you were just concerned with new entries (new sequences). But what about if sequences needed to updated (due to, for example, an experimental error)? This posed a bigger problem since database entries might be edited in inconsistent ways in the different databases. This would create conflicts when the entries were later merged. This is similar to the problem encountered if you edit Dropbox files while offline; if you edit the same file on your work computer and your home computer without synching in between, you end up with two incompatible versions of the same file.

This problem could be mitigated by updating the databases more frequently. Just as with Dropbox, the more frequent the updates, the less likely it would be that the databases would become inconsistent with each other. In practice, this meant that sharing database updates via electronic networks became more important. In 1988, EMBL-Bank set up a Fileserver through which updates could be received by email.[54] Although EMBL-Bank still produced "releases" of its most up-to-date sequence data every three months (distributed by either magnetic tape or, later, CD-ROM), this system allowed the database to distribute *daily* updates to a set of computer nodes around Europe connected by EMBNET (European Molecular Biology Network). Users with connections to these nodes could download the latest versions of sequences remotely. Likewise, in 1989, GenBank began to produce weekly "packages" of new data "available for public access over the Internet by anonymous FTP from the computer host *genbank.bio.net*."[55] When DDBJ joined the international database consortium it sought to network its computers via the US Department of Energy's Energy Science Network. By March 1990, however, it had managed to secure a direct Internet connection via the University of Tokyo Faculty of Science and the University of Hawai'i.[56] Prior to the emergence of the World Wide

[54] "Report of the meeting of the European Advisory Panel with staff of the EMBL Data Library," Heidelberg, 23 September 1988, p. 1 [NIH/01735].

[55] "Proposal to modify the GenBank contract to enhance GenBank on-line services, provide for the entry of United States Patent sequence data, and provide for increased collaboration with the National Center for Biotechnology Information," 31 August 1989, p. 2 [NHGRI/0141-008, p. 58].

[56] Sanzo Miyazawa and Hidenori Hayashida, "DDBJ Activity Report for 1989," DDBJ Newsletter, No. 9, May 1990 [NIG (trans.)].

Web, the databases struggled to make the latest updates to their data as widely available as possible via Internet FTP, BITNET, EARN, NETNORTH, JANET, email, and direct dial-up.

This increased network connectivity of the databases allowed the databases to exchange data more frequently and more rapidly. By 1990, the three databases were exchanging data updates on a daily basis by email of flat files.[57] Although this was not quite the fully automated database-to-database "transaction protocol" that had been planned, daily updates allowed the databases to move closer to their goal of "functional equivalence." This entailed that GenBank/EMBL-Bank/DDBJ would effectively become "one database" providing the same data to scientific researchers in any part of the world.[58] By the time the National Center for Biotechnology Information assumed full control over GenBank in 1992, the goal of having a uniform, shared resource was well on its way to being achieved.

Although the expansion of the World Wide Web made sharing data between databases considerably easier, sharing data more efficiently remained a primary concern into the 1990s, especially improving EMBL's connections to European networks.[59] Although this was partly driven by a desire to make the data accessible to users, the fact that consistency across databases required frequent updates meant that networking was also critical to making international collaboration feasible. Collaboration between databases influenced both the representation of DNA sequences (as standardized objects) and increased their circulation and over electronic networks. In other words, the need to collaborate made DNA sequences into the kinds of widely-distributed, widely-circulated, and standardized objects that biologists now imagine them to be.

Towards the Human Genome Project

Although DDBJ played a relatively minor role in negotiations between the databanks in the 1980s, as major genome sequencing projects got

[57] "The European Bioinformatics Institute: A Draft Plan for the Working Group of the EMBL Council," EMBL, June 1992, p. 71 [CAM/03669].

[58] "Fourth Annual Meeting, International Advisors for Nucleotide Sequence Databases" (Summary Recommendations) Washington, DC, 21–23 March, 1991, p. 2 [ASH/04605].

[59] In 1992, EMBL-Bank was experiencing serious problems connecting to EMBNET nodes. See "European Members of the International Advisory Committee for Nucleotide Sequence Databases," October 1992 meeting, draft 13 October 1992, p. 3 [ASH/04580].

underway in the 1990s, Japan's role became more significant. During 1991, Takashi Gojobori replaced Sanzo Miyazawa as the administrative director of DDBJ. Under Gojobori's leadership the database managed to attract additional funds from the Ministry of Education in Japan, to convert significant parts of GenBank's software to make it compatible with Japanese computers, and to begin planning computer and database upgrades.[60] By 1992, DDBJ's data releases contained all the sequences from both EMBL-Bank and GenBank, totalling more than 65000 entries and 85 million nucleotides.[61]

By this time, too, the transition to direct author input of sequence data was taking effect. Working together, the consortium of databases had managed to convince a growing number of journals editors that database submission should be a criterion of acceptance for publication. Since the determination of nucleotide sequences was no longer considered novel science, many editors were keen to have sequences deposited in databases rather than printed on the journal page.[62] This allowed the three databases to move to a system of requiring direct submission from authors, reducing the time spent keying data and metadata from journal pages and increasing accuracy.

As more and more journals and authors adopted direct submission, the previous division of responsibilities by journal (the so-called "journal split") became obsolete. Instead, the INSDC adopted a "geographical split" – data submitted by researchers could be submitted to any database, regardless of the journal of publication. However, US researchers were encouraged to submit GenBank, European researchers to EMBL-Bank, and Asian researchers to DDBJ.[63] This was a significant step towards creating global uniformity in sequence data – data

[60] "Fourth Annual Meeting: International Advisors for Nucleotide Sequence Databases" (Minutes) Washington, DC, 21–23 March 1991, p. 2 [ASH/04649]. "The effort to translate PC Gene to the NEC operating system took a year using some of Intelligenetics' best programmers, a task far more complicated than people realized", *Ibid.*, p. 5 [ASH/04652]. See also: "Collaborative Meeting: International Nucleotide Sequence Database: DDBJ/EMBL/GenBank," National Institute of Genetics, Mishima, Japan, 18–22 May 1992 [ASH/04791].

[61] "Collaborative Meeting: International Nucleotide Sequence Database: DDBJ/EMBL/GenBank," National Institute of Genetics, Mishima, Japan, 18–22 May, 1992, p. 2 [ASH/04792].

[62] For more on the role of databases in changing the perceived value of sequence as publishable objects see Stevens (2011).

[63] "DDBJ News Letter No. 13" (English Version), February 1993, pp. 33–37 [ASH/04770 and ASH/04472]. On this transition and its difficulties see also Sanzo Miyazawa and Hidenori Hayashida, "DDBJ Activity Report," DDBJ Newsletter No. 9, May 1990 [NIG (trans.)].

could now be submitted to any of the databases with the expectation that they be treated in the same way, and that it would all end up in same place (namely, all three databases).

This method of direct submission, however, generated another significant problem. Databases provided authors with the option to keep their submitted sequences private within the database until a future date. This mode of "hold until publication" ensured that no-one else could publish or analyse an author's data until they themselves had had a chance to analyse and publish it themselves (that is, they could not be "scooped by their own data"). In practice, however, authors did not let databases know when their data would be published and much data languished in the "hold until publication" state, inaccessible to the public and other researchers.[64] This problem was extensively discussed by the database managers and their international advisors. One solution was to educate authors: "We propose that the databanks embark on a path which help the scientific community to perceive as the authors' responsibility to notify the databases when data can be released."[65] More technical solutions, such as automatically scanning journals to find "held" sequences, proved technically infeasible.

Concerns about "holds" overlapped, too, with concerns about proprietary information. Could submission to a database be considered publication in itself and therefore conflict with patent claims or journal publication?[66] Were some authors delaying sharing of their data in order to make patent claims?

One solution proposed by the databases was to hold data for a "specified fixed time" after receipt by the database.[67] At the International Advisory Committee meeting in March 1992, the advisors worried explicitly about the decision of the European Yeast Project to hold onto its all of its data until the complete sequence of a chromosome had been determined:

> Although the advisors appreciate the legitimate right of the worker who has originally determined a new sequence to have initial pri-

[64] "European Members of the International Advisory Committee for Nucleotide Sequence Databases – Sixth Meeting," 1 October 1993 [ASH/04719].

[65] Email correspondence, Robert T. Sauer to Michael Ashburner, 26 March 1994 [ASH/04717].

[66] "International Advisory Committee for Nucleotide Sequence Databases (Minutes)," EMBL International Seminar and Guest House, Heidelberg, 9–10 March 1992, p. 5 [ASH/04600].

[67] "Fourth Annual Meeting: International Advisors for Nucleotide Sequence Databases" (Minutes) Washington, DC, 21–23 March 1991, p. 11 [ASH/04658].

> ority on its analysis, it is in the interest of the general scientific community to move the results of genome sequencing initiatives into the public databases as quickly as possible. A total period of one year reserved for data evaluation… seems sufficient to the Advisors to preserve this priority. They recommend that the databases point out to the funding agencies that stringent and explicit rules for the public release of sequence data have to be written into the contracts of future projects, and that these rules have to be strictly enforced.[68]

As the genome projects ramped up their activities, the international database collaboration became important actors in encouraging rapid sharing and release of sequence information. This was both because the databases' mission was to rapidly make data available as widely as possible, but also because the structure of the collaboration meant that "hold until publication" raised all sorts of technical and practical difficulties that made the databases' job more difficult. It was in the interests of the databases to dissociate sequences from publication and from any kinds of proprietary claims. The database managers and their advisors worked to achieve this, both through technical means (making direct submission easy and fast) and social means (putting pressure on journal editors and experimenters to submit sequences rapidly and without holds).

Another related aspect of the databases' work in the 1990s was the development of mechanisms for large-scale sequence submissions. In 1991, the data from the University of Cambridge's worm sequencing project was transferred *en masse* to EMBL-Bank. This process was not straightforward or automatic, but relied on negotiation between the database and the sequencers: "This was a result of detailed discussion in advance with the informatics experts from that group and a protocol for data exchange, which is largely based on the EMBL flat file."[69] By 1992, in addition to the worm data, EMBL-Bank had received almost one million base pairs of sequence from the Genethon Genexpress project, as well as large data depositions from the yeast project, the French *Arabidopsis* project, and the Munich Genexpress project.[70]

These large-scale submissions were the direct result of the increasing

[68] "International Advisory Committee for Nucleotide Sequence Databases (Minutes)," EMBL International Seminar and Guest House, Heidelberg, 9–10 March 1992, p. 9 [ASH/04602].

[69] Mary Shimoyama, "A Summary of the Collaborative Meeting," National Institute of Genetics, Mishima, Japan, 18–22 May 1992, p. 39 [ASH/04773].

[70] "EMBL Data Library Report – 1992," p. 3 [ASH/04730].

automation of DNA sequencing from the late 1980s onwards and the increased funding for large-scale sequencing projects. Applied Bioscience's first automatic sequencing machine (based on the work of Leroy Hood) became commercially available in 1987 (Smith et al., 1986). Concurrently, by the late 1980s, the National Institute of Health and the Department of Energy began to award large grants for the mass-sequencing of human and model-organism DNA (National Center for Human Genome Research, 1989). Whereas small-scale DNA sequencing had previously aimed at identifying genes or other features of special interest, the ramping up of the genome projects meant that sequencing increasingly became an end in itself.[71] Within this context, the role of DNA sequence databases gradually changed from the collection and management of small sequences from individual investigators to the management and coordination of large submission.

At DDBJ, throughout the 1990s significant effort was invested in developing and updating submission systems for larger and larger chunks of data. The Japanese database received large-scale submissions from universities and institutes in Japan sequencing *Synechocystis*, *E. coli*, *Arabidopsis*, and worm: "For each of those projects, we at DDBJ have formed a team to discuss with the project people the submission, processing, and release of data, before submissions are made" (Tateno et al., 1998). This eventually led to the development of a "large-scale" or "mass" submission system made available over the World Wide Web. Such systems performed a large number of automatic checks for errors, consistency, and formatting, allowing large submissions without requiring extensive intervention by databases curators (Sugawara et al., 1999; Tateno et al., 2000).

In both the push towards immediate release of data and in creating systems for the large-scale submission of sequences, the database collaboration played a critical role in enabling the work of the genome projects. After the Bermuda meeting in 1996, rapid data release became a hallmark of genomic work.[72] However, the social and technical work by the database collaboration meant that the structures for this rapid sharing were already in place. Likewise, the collaboration between the

[71] These developments also involved changes in the scientific value and moral economy of DNA sequence. See Strasser (2011) and Stevens (2011).

[72] The Bermuda Meetings were convened between the major genome sequencing centers in order to agree upon guidelines for the public release of newly determined DNA sequence in the human genome project. The outcome included the agreement that any sequence greater than 1 kilobase would be uploaded to one of the databases (EMBL-Bank, GenBank, or DDBJ) within 24 h. For more detail see Ankeny et al. (this volume).

databases and sequencing teams allowed the kind of rapid and large-scale submission that the genome projects required. The sharing of data that had already been taking place between the databases meant that standards and infrastructure for mass data sharing were already developed and ready to use for moving sequences between producers and databases.

Conclusions

The point I have hoped to establish is that international collaboration had an important impact on both the organization of the databases and the construction of "sequence" as a digital objects. The transnational nature of the database collaboration meant that sequence had to become increasingly *abstracted* from its points of origin, they had to become increasingly *sharable*, and they had to become *mobile* via electronic media (and networks). This did not occur automatically, but was the product of intensive work by databases managers at GenBank, EMBL-Bank, and at DDBJ. It involved both negotiation, compromise, politics, but also the technical work of establishing how databases could actually talk to one another via electronic networks, translation, and standardization.

The HGP relied on sequence being exactly the kind of abstract, sharable, mobile object that the database collaborations had made it into. Just as no one database could tackle the problem of archiving all sequence data, no one lab could tackle the problem of sequencing the human genome. For the human genome to become the human genome, sequence data had to be shared rapidly not only between databases but also between labs in different parts of the world. For this sharing to work, it had to be a standardized object – sequence from the UK had to be the same as sequence from China.

Many of the accounts of the HGP and genomics take for granted the circulation of sequence data as a mobile and consistent object, but this mobility was hard won. This story suggests that international database collaboration played an important role in doing the work of making sequences mobile. Again this was not merely a practical or technical feat of working out ways to transmit this data, but also involved the social feat of figuring out how to make sequence something that could be *represented the same way* in different parts of the world. This is a story that necessarily has a transnational dimension – this work took place across and between different geographical sites and involved elements of translation as well as standardization. The mobility of sequences was a direct product of the transnational context in which they were developed as informatic objects.

The significance of the INSDC not only suggests the value of a transnational approach to the history of genomics, but also suggests ways in which we might enrich the history of genomics by widening the frame beyond Anglo-America. Although the vast majority of sequencing was done in the US and UK, scientists not only from Japan, China, France, and Germany, but also the former Soviet states, Australasia, and other parts of Europe and Asia contributed to technology development, informatics, standardization, networking, and myriad other activities that made the genome projects possible. The significance of the "global genome" and of genomics as a global enterprise will only be understood when the story of the HGP is placed within this more international context.

Finally, the history presented here suggests how we might enrich our accounts of scientific databases by approaching them as transnational institutions. Building on previous work that has shown the importance of negotiation across multiple sites, we have seen here how not only nucleotide sequence databases, but also nucleotide sequences themselves, emerged through a process of intensive international negotiation and transnational work. Sequences, as objects, were made transnationally. Taking such a viewpoint can help to explain the apparent "automatic" globalization of objects such as nucleotide sequences and also some of the ease with which the HGP became an "international" project. The sorts of transnational dynamics described here form the kind of "behind the scenes" work necessary for establishing smooth global flows of data.

Acknowledgements

The research and writing of this paper was funded by a Tier 1 grant from the Ministry of Education, Singapore (RG 56/13). Sponsorship from the National Institute for Human Genome Research allowed me to present an early version of this paper at that institute's "Archiving the Human Genome Project" workshop in April 2015. Thanks to the NHGRI's Christopher Donahue for working to make NHGRI sources available for this paper. The National Institute of Genetics in Mishima, Japan also provided support that enabled me to present a version of this paper there. At the NIG, I received the generous help of Machiko Itoh, who provided many documents related to DDBJ, and also helped me to organize interviews with those involved with the Japanese database. Thanks to Arita Masanori, Asao Fujiyama, Kazuho Ikeo, Kosaku Okubo, Minoru Kanehisa, and Sanzo Miyazawa for agreeing to be interviewed for this project. I would also like to acknowledge the assistance of Elite Translations Asia for their translation from Japanese of documents made available by DDBJ, including DDBJ newsletters. Michael Ashburner, Dennis Benson, and Graham Cameron also generously made records in their personal possession available for me to consult.

Open Access

Archival sources

The full details of the archival collections appearing in the footnotes are as follows:

ASH – Records of Michael Ashburner (International Advisor, International Nucleotide Sequence Database Consortium). Digital files in the possession of the author.

BEN – Records of Dennis Benson (National Institutes of Health). Digital files in possession of the author.

CAM – Records of Graham Cameron (European Bioinformatics Institute). Digital files in possession of the author.

NHGRI – NHGRI History Archive Resource, National Human Genome Research Institute Archive, nih.sharepoint.com/site/NHGRIHistoryArchive.

NIG – Sources obtained from the National Institute of Genetics, Mishima, Japan. Newsletters can be found at: https://www.ddbj.nig.ac.jp/. Newsletters No. 1–9 were in Japanese only. The author had these translated by a professional translation service and these sources are listed as [NIG (trans.)]. Newsletters numbered 10 and above were published in both Japanese and English.

References

Ankeny, R. A., Maxson Jones, K. and Cook-Deegan, R. this volume. "The Bermuda Triangle: the politics, principles, and pragmatics of data sharing in the history of the Human Genome Project, 1963–2003." *Journal of the History of Biology*.

Bowker, Geoffrey. 2005. *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.

Bowker, Geoffrey C. and Star, Susan L. 1999. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Chow-White, Peter A. and García-Sancho, Miguel. 2011. "Bidirectional Shaping and Spaces of Convergence: Interactions Between Biology and Computing from the First DNA Sequencers to Global Genome Databases." *Science, Technology, and Human Values* 37(1): 124–164.

Cook-Deegan, Robert. 1996. *The Gene Wars: Science, Politics, and the Human Genome*. New York: W.W. Norton.

Davies, G., Frow, E. and Leonelli, S. 2013. "Bigger, Faster, Better? Rhetorics and Practice of Large-Scale Research in Contemporary Bioscience." *Biosocieties* 8(4): 386–396.

Farquhar, Judith and Rajan, Kaushik S., eds. 2014. "Knowledge/Value: Information, Archives, Databases [special issue]." *East Asian Science, Technology, and Society: An International Journal* 8(4): 383–478.

García-Sancho, Miguel. 2012. *Biology, Computing, and the History of Molecular Sequencing: From Proteins to DNA, 1945–2000*. London: Palgrave Macmillan.

Haigh, Thomas. 2004. "'A Veritable Bucket of Facts': Origins of the Data Base Management System." M. E. Bowden and B. Rayward (eds.), *Proceedings of the 2nd Conference of the History and Heritage of Scientific Information Systems*. Medford, NJ: Information Today Press, pp. 73–78.

Hilgartner, Stephen. 1995. "Biomolecular Databases: New Communications Regimes for Biology?' *Science Communication* 17(2): 240–263.

—— 2004 "Making Maps and Making Social Order: Governing American Genome Centers, 1988–1993." J.-P. Gaudillière and H.-J. Rheinberger (eds.), *From Molecular Genetics to Genomics: The Mapping Cultures of Twentieth-Century Genetics*. New York: Routledge, pp. 113–135.

—— 2013 "Constituting Large-Scale Biology: Building a Regime of Governance in the Early Years of the Human Genome Project." *Biosocieties* 8(4): 397–416.

Hine, Christine. 2006. "Databases as Scientific Instruments and Their Role in the Ordering of Scientific Work." *Social Studies of Science* 36(2): 269–298.

Howlett, Peter and Morgan, Mary S. 2011. *How Well Do Facts Travel? The Dissemination of Reliable Knowledge*. Cambridge, UK: Cambridge University Press.

Kanehisa, Minoru. 1983. "DNA Databases and Computer Analysis." *Cellular Engineering* 2(13): 1520–1532 [translated from Japanese].

Kanehisa, Minoru and Oi, Tatsuo. 1984. "DNA Databank." *Biophysics* 24(1): 51–54 [translated from Japanese].

Kishi, Nobuhito. 2004. *Genomu haiboku* [A Defeat in the Genome Project]. Diamond.

Kuhara, Satoru and Hayashi, Katsuya. 1984. "Q&A: Genetic Database System." *Cellular Engineering* 3(2): 170–179 [translated from Japanese].

Leonelli, Sabina. 2009. "Centralising Labels to Distribute Data: The Regulatory Role of Genomic Consortia." P. Atkinson, P. Glasner and M. Lock (eds.), *The Handbook for Genetics and Society: Mapping the New Genomic Era*. London: Routledge, pp. 469–485.

—— 2010 "Packaging Small Facts for Re-use: Databases in Model Organism Biology." P. Howlett and M. S. Morgan (eds.), *How Well Do Facts Travel? The Dissemination of Reliable Knowledge*. Cambridge, UK: Cambridge University Press, pp. 325–348.

Mackenzie, Adrian, et al. 2015. "Post-archival Genomics and the Bulk Logistics of DNA Sequences." *Biosocieties* 11(1): 82–105.

McCray, W. Patrick. 2014. "How Astronomers Digitized the Sky." *Technology and Culture* 55(4): 908–944.

McElheny, Victor. 2012. *Drawing the Map of Life: Inside the Human Genome Project*. New York: Basic Books.

National Center for Human Genome Research. 1989. Human Genome Program Center Grants (P30). NIH Guide for Grant and Contracts 18, no. 25 (21 July). https://grants.nih.gov/grants/guide/historical/1989_07_21_Vol_18_No_25.pdf. Accessed 10 July 2017.

National Human Genome Research Institute. 2000. Press Release, The White House, Office of the Press Secretary, 26 June. https://www.genome.gov/10001356/. Accessed 26 July 2016.

Obayashi, M. 1986. "Origins of Molecular Biology in Japan." *Journal of the University of Occupational and Environmental Health* 8(2): 251–256.

Smith, Brian Cantwell. 1998. *On the Origin of Objects*. Cambridge, MA: MIT Press.

Smith, L. M., Sanders, J. Z., et al. 1986. "Fluorescence Detection in Automated DNA Sequence Analysis." *Nature* 321: 674–679.

Soll, D., Kirschstein, R. L., Philipson, L. and Uchida, H. 1988. "DNA Databases Monitored." *Science* 240(4851): 375.

Stevens, Hallam. 2011. "Coding Sequences: A History of Sequence Comparison Algorithms." *Perspectives on Science* 19(3): 263–299.

—— 2013 *Life Out of Sequence: A Data-Driven History of Bioinformatics*. Chicago, IL: University of Chicago Press.

Strasser, Bruno. 2011. "An Experimenter's Museum: GenBank, Natural History, and the Moral Economies of Biomedicine." *Isis* 102(1): 60–96.

Sugawara, H., et al. 1999. "DNA Data Bank of Japan Dealing with Large-Scale Data Submission." *Nucleic Acids Research* 27(1): 25–28.

Sulston, John and Ferry, Georgina. 2002. *The Common Thread: A Story of Science, Politics, Ethics, and the Human Genome*. London: Joseph Henry Press.

Tateno, Y., et al. 1998. "DNA Data Bank of Japan at Work on Genome Sequence Data." *Nucleic Acids Research* 26(1): 16–20.

—— 2000 "DNA Data Bank of Japan (DDBJ) in Collaboration with Mass Sequencing Teams." *Nucleic Acids Research* 28(1): 24–26.

Thacker, Eugene. 2006. *Global Genome: Biotechnology, Politics, and Culture*. Cambridge, MA: MIT Press.

Turchetti, Simone, Herran, Nestor and Boudia, Soraya. 2012. "Introduction: Have We Ever Been 'Transnational'? Towards a History of Science Across and Beyond Borders." *British Journal for the History of Science* 45(3): 319–336.

Uchida, Hisao. 1993. "Building a Science in Japan. The Formative Decades of Molecular Biology." *Journal of the History of Biology* 26(3): 499–517.

Walgate, Robert. 1982. "Europe Leads on Sequences." *Nature* 296: 596.

Wang, Zouyue. 2010. "Transnational Science During the Cold War: The Case of Chinese/American Scientists." *Isis* 101(2): 367–377.