TIMOTHY LENOIR

Science and the Academy of the 21st Century

Does Their Past Have a Future in an Age of Computer-Mediated Networks?

Science and media: from Leibniz to Lederberg

Since their inception as formal institutions in the seventeenth century academies of science have been imagined above all as sites of optimal communication. Historians of the early state-recognized academies of science in Paris and London have pointed to three key features in the transition from amateur groups of aficionados and virtuosi to learned societies: Each organization had a communal mission to share information within its membership; each also had a recording secretary and a corresponding secretary who disseminated the group's collective findings in some form of journal or newsletter; each group also placed a high value on experimentation, and prided itself on its laboratory as much as on its secretary and journal. While agreeing that due emphasis should be placed on the role of the journal in helping to establish international communities of natural philosophers interacting critically with one another's work, scholars such as Elizabeth Eisenstein and Bruno Latour go even further in emphasizing the material features of scientific communications, particularly the late 15th century communications revolution in printing and improvements in postal delivery, as crucial to the entire scientific revolution itself.[1] In her work on the printing press as an agent of change, for instance, Eisenstein observes that *prior* to the innovative intellectual contributions of the Copernicuses, the Galileos, the Keplers, and the Newtons of the Scientific Revolution, there was the quiet but vitally important circulation of astronomical tables, such as the Alphonsine Tables, diagrams accompanying mathematical demonstrations and as-

[1] Bruno Latour, How to Be Iconophilic in Art, Science and Religion?, in: Carolyn A. Jones/Peter Galison, eds., Picturing Science Producing Art, New York 1998, pp. 418–440, especially pp. 424–428; Bruno Latour, Drawing Things Together, in: Michael Lynch/Steve Woolgar, eds., Representation in Scientific Practice, Cambridge, Mass. 1990, pp. 19–68; Also see, Bruno Latour, Science in Action: How to Follow Scientists and Engineers Through Society, Cambridge, Mass. 1987.

tronomical constructions, maps of all kinds, and images of plants. According to Eisenstein, in contrast to the scribal tradition of copying manuscripts the printed text became the anchor for feedback, sustained discussion, and incremental cumulative improvement of the information communicated. When early academy founders referred to the need for creating an international community and of extracting their enterprise from verbal dispute, it was the new visual language in the pictorial statements of Vesalius or the triangles of Galileo or the illustrations of the moon's surface that made it possible. They boldly proclaimed that natural philosophers should turn away from bookish learning and learn through observation and experiment to read the book of nature directly. But crucial to this perspective is that before natural philosophers could begin to read the book of nature, nature had to be inscribed in book form.

The thesis set forth by Eisenstein and transformed by Latour into a theory of "immovable mobiles" for creating actor-networks has been critiqued for its implicit technological determinism by Adrian Johns in his brilliant study, 'The Nature of the Book: Print and Knowledge in the Making'. Johns argues that before the introduction of lithography in the 19th century "fixity" was anything but a firm property of printed texts – and he in fact disputes "fixity" ever came into being. In any case, the focus on technology displaces the important issue in Johns' view. Rather than treating fixity as an inherent property of printed texts which they carry with them from place to place, we should explore the processes for calibrating local reading practices required to make a text produced in one site authoritative in another. Such an investigation calls for the recognition that fixity exists only when it is recognized and acted upon by people. Accordingly print culture is a result of various contested representations, practices and conflicts rather than a deterministic consequence of print technology.[2]

I couldn't agree more. But to be fair to Eisenstein (and Latour) the position they were arguing against paid lip service at best to the role of technology. A statement by Archer Taylor among numerous others that could be cited illustrates the position they are arguing against:

'The powers which shape men's lives may be expressed in books and type, but by and of itself printing [...] is only a tool, an instrument, and the multiplication of tools and instruments does not of itself affect intellectual and spiritual life.'[3]

The point is that print technology was not just any tool. While not succumbing to an "impact" model, we should examine the special features that distinguish printing from other innovations. Through its development, modification, and adaptation to the interests of particular groups printing technology became the material artifact around which new forms of communication and new intellectual configurations came about, one of them being natural science as we know it. As Eisenstein observed, one cannot treat printing as just one among many elements in a complex

[2] Adrian Johns, The Nature of the Book: Print and Knowledge in the Making, Chicago 1998, p. 19.

[3] Elizabeth L. Eisenstein, The Printing Press as an Agent of Change, 2 vols., Cambridge 1979, vol. 2, p. 703.

causal nexus, for the communications shift transformed the nature of the causal nexus itself. What is at stake here is the construction of the printed text *as medium*; and that story cannot be one simply of technology because the text cannot compel readers to react in certain ways. Crucial to such a program is a shift away from emphasis on the technology alone to focus on the construction of readers, and more broadly on the moral values of trust and convention in the making of knowledge, and on how readers decide what to believe when confronted with printed materials. To put it in its most general terms: *media are institutions*. The task of the media historian is twofold: On the one hand to understand how the medium is constructed in a co-evolutionary entanglement of technical artifacts, such as the various elements for producing printed works, the organizations for distributing them, and social roles such as that of author, publisher, and reader; and on the other hand the configurations of knowledge and practice for those who work within the medium.

The controversies around print culture and the history of the book provide a point of departure for considering the topic that interests me today: namely the transformation in scholarly practices and indeed in the reconfiguration of knowledge and knowledge-production being brought about by the communications revolution of our own day, computer-mediated communication. I want to concentrate on fields of biomedical research where computer-mediated communication is having some of its most dramatic effects, particularly through the emergence of genomics and post-genomic science. If we think of the work of Watson, Crick, Monod and Jacob, Gamow, Nirenberg, and others as having launched biology into an era of exploring the structures for encoding and replicating biological information, then the human genome project and its ancillary tools for reading, writing, indexing and searching (analogous in many ways to the "knowledge industry" described by Eisenstein) massive libraries of biological data are creating its second (science-based) industrial revolution. More importantly, as these new tools from information science have been adopted within biology, the conceptual terrain and indeed many of the material practices of biologists have been and are continuing to be radically transformed. It hasn't happened by simply booting up a computer. Controversy, labor, and the sustained contributions of numerous groups of scientists and engineers over two decades have gone into shaping these social and technical networks into a new medium. The development of bioinformatics, which can be generally defined as the study of how information technologies are used to solve problems in biology, offers a useful window for illustrating those changes. Today I want to suggest that the outcome of these changes will be the fusion of the communication and experimentation functions – the merging of the journal and the lab – in the post-modern academy.

A New Biology for the Information age

The Holy Grail of many biologists since the 1960s has been the construction of a mathematized, predictive theoretical biology. An early vision of how to get there

was directed by the notion that the information for the three dimensional folding and structure of proteins is uniquely contained in the linear sequence of their amino acids and that understanding the dynamics of protein folding would provide the basis for computational biology.[4] The molecular dynamics approach assumed that if all the forces between atoms in a molecule, including bond energies and electrostatic attraction and repulsion, are known, then it is possible to calculate the three-dimensional arrangement of atoms that requires the least energy. While theoretically elegant, the determination of protein structure from chemical and dynamical principles has been hobbled with difficulties. In the abstract, analysis of physical data generated from protein crystals, such as x-ray and nuclear magnetic resonance data, should offer rigorous ways to connect primary amino acid sequences to 3D structure. But the problems of acquiring good crystals and the difficulty of getting NMR data of sufficient resolution are impediments to this approach. These and several related difficulties have contributed to the slow rate of progress in registering atomic coordinates of macromolecules.[5] Moreover, while quantum mechanics provides a solution to the protein folding problem in theory, the computational task of predicting structure from first principles for large protein molecules containing many thousands of atoms has proven impractical. Shortcuts have been developed that combine molecular dynamics computations, artificial intelligence and interac-

[4] See Christian B. Anfinsen, Principles that Govern the Folding of Protein Chains, Science (1973). 181 (Number 4096), pp. 223–230 discusses the work for which he was awarded the Nobel Prize for Chemistry in 1972: "This hypothesis (the 'thermodynamic hypothesis') states that the three-dimensional structure of a native protein in its normal physiological milieu [...] is the one in which the Gibbs free energy of the whole system is lowest; that is, that the totality of interatomic interactions and hence by the amino acid sequence, in a given environment" (p. 223).

[5] An indicator of the difficulty of pursuing this approach alone is suggested by the growth of databanks of atomic coordinates for proteins. The Protein Data Bank (PDB) was established in 1971 as a computer-based archival resource for macromolecular structures. The purpose of the PDB was to collect, standardize, and distribute atomic co-ordinates and other data from crystallographic studies. In 1977 the PDB listed atomic coordinates for 47 macromolecules. In 1987 that number began to increase rapidly at a rate of about 10 percent per year due to the development of area detectors and widespread use of synchrotron radiation, so that by April 1990 atomic coordinate entries existed for 535 macromolecules. Commenting on the state of the art in 1990, Holbrook et al. noted that crystal determination could require one or more man-years. (F. C. Bernstein/T. F. Koetzle, et al. (1977), The Protein Data Bank: A computer based archival file for macromolecular structure, Journal of Molecular Biology 112, pp. 535–542). Currently (1999) the Biological Macromolecule Crystallization Database (BMCD) of the Protein Data Bank contains entries for 2526 biological macromolecules for which diffraction quality crystals have been obtained. These include proteins, protein:protein complexes, nucleic acid, nucleic acid:nucleic acid complexes, protein:nucleic acid complexes, and viruses. (S. R. Holbrook/S. M. Muskal, et al. (1993). Predicting Protein Structural Features with Artificial Neural Networks, Artificial Intelligence and Molecular Biology, ed. by L. Hunter, Menlo Park, CA, pp. 161–194.)

tive computer graphics in deriving protein structure directly from chemical structure. But the task is still daunting.

While structure determination was moving at a snail's pace, beginning in the 1970s another stream of work contributed to the transformation of biology as an information science. In the mid-1970s new tools of molecular biology, such as recombinant DNA techniques, gene cloning, restriction enzymes, protein sequencing, and gene product amplification began to emerge. Biologists were suddenly awash in a sea of new data. They deposited this data in large and growing electronic databases of genetic maps, atomic coordinates for chemical and protein structures, and protein sequences. Indeed more than 140,000 genes were cloned and sequenced in the twenty years from 1974 to 1994, of which more than 20 percent were human genes.[6] By the early 1990s, at the beginning of the Human Genome Initiative, the NIH GenBank database (release 70) contained more than 74,000 sequences, while the Swiss Protein database (Swiss-Prot) included nearly 23,000 sequences. Protein databases were doubling in size every 15 months, and some were predicting that as a result of technological impact of the Human Genome Initiative by the year 2000 ten million base pairs a day would be sequenced.

Such an explosion of data and its registration in databases encouraged the development of a second approach to determining function and structure of protein sequences: namely, prediction from sequence data alone by applying artificial intelligence, expert systems and developing search tools to identify structures and patterns in their data. That is the vision of bioinformatics, a discipline less than a decade old. It studies two important information flows in modern biology. The first is the flow of *genetic information* from the DNA of an individual organism up to the characteristics of a population of such organisms (with an eventual passage of information back to the genetic pool, as encoded within DNA). The second is the flow of experimental information from observed biological phenomena to models that explain them, and then to new experiments in order to test these models.

MOLGEN, SUMEX-AIM and GENET

A key project illustrating the ways in which computer science and molecular biology began to merge in the formation of bioinformatics was the MOLGEN project at Stanford and events related to the formation and subsequent development of BIONET. MOLGEN was a continuation of projects in artificial intelligence and knowledge engineering begun at Stanford with DENDRAL during the 1960s. MOLGEN was started in 1975 as a artificial intelligence project in the Heuristic Programming Project with Edward Feigenbaum as principal investigator directing the thesis pro-

[6] D. Brutlag, Understanding the Human Genome, Scientific American Introduction to Molecular Medicine. P. Leder/D. A. Clayton/E. Rubenstein, eds., New York 1994, p. 159.

jects of Mark Stefik and Peter Friedland.[7] The aim of MOLGEN was to model the experimental design activity of scientists in molecular genetics.[8] Before an experimentalist sets out to achieve some goal, he produces a working outline of the experiment, guiding each step of the process. The central idea of MOLGEN was that in designing a new experiment scientists rarely plan from scratch. Instead they find a skeletal plan, an overall design that has worked for a related or more abstract problem, and then adapt it to the particular experimental context. Similar to DENDRAL this approach is heavily dependent upon large amounts of domain-specific knowledge in the field of molecular biology and especially upon good heuristics for choosing among alternative implementations.

MOLGEN's designers chose molecular biology as appropriate for the application of artificial intelligence because the techniques and instrumentation generated in the 1970s seemed ripe for automation. The advent of rapid DNA cloning and sequencing methods had had an explosive effect on the amount of data that could be most readily represented and analyzed by a computer. Moreover, it appeared that very soon progress in analyzing information in DNA sequences would be limited by the appropriate combination of the available search and statistical tools. MOLGEN was intended to apply rules to detect profitable directions of analysis and to reject unpromising ones.[9]

Peter Friedland was responsible for constructing the knowledge-base component of MOLGEN, and though not himself a molecular biologist, he made a major contribution to the field by assembling the rules and techniques of molecular biology into an interactive, computerized system of analytical programs. Friedland worked with Stanford molecular biologists Douglas Brutlag, Laurence Kedes, John Sninsky, and Rosalind Grymes, who provided expert knowledge on enzymatic methods, nucleic acid structures, detection methods, and pointers to key references in all areas of molecular biology. Brutlag, Kedes, Sninsky, and Grymes were interested in having a battery of automated tools for sequence analysis, and they contracted with Friedland and Stefik – both gifted computer program designers – to build them in exchange for contributing their expert knowledge to the project.[10] This collaboration

[7] E. A. Feigenbaum and N. Martin, Proposal: MOLGEN – a computer science application to molecular genetics, Heuristic Programing Project, Stanford University, Technical Report No: HPP-78-18, 1977.

[8] P. Friedland, Knowledge-Based Experiment Design in Molecular Genetics. Ph.D. Thesis, Computer Science, Stanford University, Stanford, 1979.

[9] E. A. Feigenbaum/B. Buchanan, et al., A Proposal for Continuation of the MOLGEN Project: A Computer Science Application to Molecular Biology, Computer Science Department, Stanford University, Heuristic Programming Project, Technical Report No. HPP-80-5, April, 1980, Section 1, p. 1.

[10] Douglas Brutlag, personal communication. Peter Friedland, personal communication. After his work on MOLGEN and at IntelliGenetics (discussed below) Friedland went on to become chief scientist at the NASA-Ames Laboratory for Artificial Intelligence in 1987.

of computer scientists and molecular biologists helped biology along the road to becoming an information science.

An example of the programs Friedland and Stefik created for MOLGEN was SEQ, an interactive self-documenting program for nucleic acid sequence analysis which had 13 different procedures with over 25 different sub-procedures, many of which could be invoked simultaneously to provide different analytical methods for any sequence of interest. SEQ brought together in a single program methods for primary sequence analysis described in the literature.[11] SEQ also performed homology searches and specified the degree of homology and dyad symmetry (inverted repeats) searches on DNA sequences.[12] Another feature of SEQ was its ability to prepare restriction maps with the names and locations of the restriction sites marked on the nucleotide sequence in addition to having a facility for calculating the length of DNA fragments from restriction digests of any known sequence.

In its first phase of development (1977–1980) MOLGEN consisted of such programs described above and a knowledge base containing information on about 300 laboratory methods and 30 strategies for using them. It also contained the best currently available data on about 40 common phages, plasmids, genes, and other known nucleic acid structures. The second phase of development beginning in 1980 scaled up both analytical tools and knowledge base. Perhaps the most significant aspect of the second phase was making MOLGEN available to the scientific community at large on the Stanford University Medical Experimental national computer resource, SUMEX-AIM. SUMEX-AIM, supported by the Biotechnology Resources Program at NIH since 1974, had been home to DENDRAL and several other programs. The new experimental resource on SUMEX, comprising the MOLGEN programs and access to all major genetic databases, was called GENET. In February 1980 GENET was made available to a carefully limited community of users.[13]

MOLGEN and GENET were immediate successes with the molecular biology community. In their first few months of operation in 1979 more than 200 labs (with several users in each of those labs) accessed the system. By November 1, 1982 more than 300 labs from 100 institutions accessed the system around the clock.[14]

[11] L. J. Korn, C. L. Queen, and M. N. Wegman, Computer Analysis of Nucleic Acid Regulatory Sequences, Proceedings of the National Academy of Sciences 74 (1977), pp. 4516–4520; R. Staden, Sequence Data Handling by Computer, Nucleic Acids Research 4 (1977), pp. 4037–4051; R. Staden, Further Procedures for Sequence Analysis by Computer, Nucleic Acids Research 5 (1978), pp. 1013–1015; R. Staden, A Strategy of DNA Sequencing Employing Computer Programs, Nucleic Acids Research 6 (1979), pp. 2602–2610.

[12] P. Friedland, D. L. Brutlag, J. Clayton, and L. H. Kedes, SEQ: A Nucleotide Sequence Analysis and Recombinant System, Nucleic Acids Research 10 (1982), pp. 279–294.

[13] T. Rindfleisch, P. Friedland, and J. Clayton, The GENET Guest Service on SUMEX, SUMEX-AIM Report, 1981: Stanford University Special Collections, Friedland Papers, Fldr. GENET.

[14] Doug Brutlag, Personal Communication, 6/19/99. Also discussed in the official site review for BIONET conducted by the NIH Special Study Section, March 17–19, 1983, BIONET, National Computer Resource for Molecular Biology, Stanford University Special Collections,

Traffic on the site was so heavy that restrictions had to be implemented and plans for expansion considered. In addition to the academic users a number of biotech firms, such as Monsanto, Genetech, Cetus, and Chiron, used the system heavily. In order to insure that the academic community had unrestricted access to the SUMEX computer and that the NIH would be satisfied commercial users were not getting unfair access to the resource, Feigenbaum, principle investigator in charge of the SUMEX resource, and Thomas Rindfleisch, facility manager, decided to exclude commercial users.[15]

To provide commercial users with their own unrestricted access to GENET and MOLGEN programs, Brutlag, Feigenbaum, Friedland, and Kedes formed a company, IntelliGenetics, which would offer the suite of MOLGEN software for sale or rental to the emerging biotechnology industry. With 125 research labs doing recombinant DNA research in the US alone and a number of new genetic engineering firms starting up, opportunities looked outstanding. Numerous firms were being formed with staffs of molecular biologists exceeding 50 individuals, and several were planning to hire over 1,000 Ph.D.s in molecular biology. No one was currently supplying software in this rapidly growing genetic engineering marketplace. With their exclusive licensing arrangement with Stanford for the MOLGEN software, IntelliGenetics was poised to lead a huge growth area.[16] The resource that IntelliGenetics eventually offered to commercial users was BIONET. Like GENET, its prototype, BIONET combined in one computer site databases of DNA sequences with programs to aid in their analysis.

Prior to the startup of BIONET, GENET was not the only resource for DNA sequences. Several researchers were making their databases available. Margaret Dayhoff had created a database of DNA sequences and some software for sequence analysis for the National Biomedical Research Foundation that was marketed commercially. Walter Goad, a physicist at Los Alamos National Laboratory, collected DNA sequences from the published literature and made them freely available to researchers. But by the late 1970s the number of bases sequenced was already approaching 3 million and expected to double soon. Some form of easy communication between labs and effective data handling was considered a major priority in the biological community. While experiments were going on with GENET a number of

Brutlag Papers, p. 2. Also discussed in R. Lewin, National Networks for Molecular Biologists, Science 223 (1984), pp. 1379–1380.

[15] This was announced to the GENET community by Allan Maxam, the chairman of the national advisory board. See: Allan M. Maxam to GENET Community. Subject: Closing of GENET: August 23, 1982. Stanford University Special Collections, Peter Friedland Papers, Fldr. GENET.

[16] Business Plan for IntelliGenetics, May 8, 1981, p. 5. Stanford Special Collections, Brutlag Papers, Fldr. IntelliGenetics. Emphasis in the original. Details of the software licensing arrangement and the revenues generated are discussed in a letter to Niels Reimers, Stanford Office of Technology Licensing on the occasion of renegotiating the terms. See: Peter Friedland to Niels Reimers. Subject: Software Licensing Agreement: April 2, 1984. Stanford University Special Collections, Fldr. IntelliGenetics.

nationally prominent molecular biologists had been pressing to start a NIH-sponsored central repository for DNA sequences. An early meeting organized by Joshua Lederberg was held in 1979 at Rockefeller University. The proposed NIH initiative was originally supposed to be coordinated with a similar effort at the European Molecular Biology Laboratory (EMBL) in Heidelberg, but the Europeans became dissatisfied with the lack of progress on the American end and decided to go ahead with their own databank. EMBL announced the availability of its Nucleotide Sequence Data Library in April 1982, several months before the American project was funded. Finally, in August, 1982 the NIH awarded a contract for $ 3 million over 5 years to the Boston-based firm of Bolt, Berenek, and Newman (BB&N) to set up the national database known as GenBank in collaboration with Los Alamos National Laboratory.

Although GenBank launched a formal national DNA sequence collection effort, the need for computational facilities voiced by molecular biologists was still left unanswered. In September 1983 after a review process that took over a year, the NIH division of research resources awarded IntelliGenetics a $ 5.6 million five year contract to establish BIONET, in part to address the need for a national center for computational analysis of DNA.[17] The contract started on March 1, 1984 and ended on February 27, 1989.

BIONET first became available to the research community in November 1984. The fee for use was $ 400 per year per laboratory, and remained at that level throughout its first five years. BIONET's use grew impressively. Initially the IntelliGenetics team set the target for user subscriptions at 250 labs. However the annual report for the first year's activities of BIONET in March, 1985 listed 350 labs with nearly 1132 users. By August 1985 that number had increased dramatically to 450 labs and 1500 users.[18] By 1989 900 laboratories in the U.S., Canada, Europe, and Japan (comprising about 2800 researchers) subscribed to BIONET, and 20 to 40 new laboratories joined each month.[19]

BIONET was intended to establish a national computer resource for molecular biology satisfying three goals. A first goal was to provide a way for academic biologists to obtain access to computational tools to facilitate their nucleic acid (and possibly protein) related research. In addition to giving researchers ready access to national databases on DNA and protein sequences, BIONET would provide a library of sophisticated software for sequence searching, matching, and manipulation. A second goal was to provide a mechanism to facilitate research into improving such tools. The BIONET contract provided research and development support for

[17] Lewin noted that this was the largest award of its kind by NIH to a for-profit organization. See fn. 14, p. 1380.

[18] Minutes of the Meeting of the National Advisory Committee for BIONET, March 23, 1985 (Final version prepared 1 August 1985), p. 4. In Stanford University Special Collections, Brutlag Papers, Fldr. BIONET.

[19] Joel Huberman, BIONET: Computer Power for the Rest of Us, (1989), p. 1.

additional software. A third goal of BIONET was to enhance scientific productivity through electronic communications.

The stimulation of collaborative work through electronic communication was perhaps the most impressive achievement of BIONET. BIONET was much more than the Stanford GENET plus the MOLGEN-IntelliGenetics suite of software. Whereas GENET with its pair of ports could accommodate only two users at any one time, BIONET had 22 ports providing an estimated annual 30,000 connect hours.[20] All subscribers to BIONET were provided with email accounts. For most molecular biologists this was something entirely new, since most university labs were just beginning to be connected with regular email service. At least 20 different bulletin boards on numerous topics were supported by BIONET. In an effort to change the culture of molecular biologists by accustoming them to the use of electronic communications and more collaborative work, BIONET users were required to join one of the bulletin board groups.

BIONET subscribers had access to the latest versions of the most important databases for molecular biology, including (i) GenBank[tm]; (ii) EMBL, the European Molecular Biology Laboratory nucleotide sequence library; (iii) NBRF-PIR, the National Biomedical Research Foundation's protein sequence database;(iv) SWISS-PROT; (v) VectorBank[tm], a database of cloning vector restriction maps and sequences; (vi) Restriction Enzyme Library, a complete list of restriction enzymes and cutting sites provided by Richard Roberts at Cold Spring Harbor; and (vii) Keybank, IntelliGenetics' collection of predefined patterns or "keys" for database searching. Several smaller databases were also available, including a directory of molecular biology databases, a collection of literature references to sequence analysis papers, and a complete set of detailed molecular biological laboratory protocols (especially for *E. coli* and yeast work).[21]

Perhaps the most important contribution made by BIONET to establishing molecular biology as an information science was negotiated at the renewal of the contract to manage GenBank in 1987. BB&N was 2 years behind in publishing and disseminating sequence data it had received, making GenBank about 20 % out of date compared to other commercially available databases.[22] Concerned that re-

[20] Peter Friedland, BIONET Organizational Plans, 27 April 1984, Company Confidential Memo. Stanford University Special Collections, Brutlag Papers, Fldr. BIONET, p. 1. A published version of these objectives appeared as: Dennis H. Smith/Douglas Brutlag/Peter Friedland/Laurence H. Kedes, BIONET[tm]: national computer resource for molecular biology, Nucleic Acids Research, 14(1) (1986), pp. 17–20.

[21] IntelliGenetics, Introduction to Bionet[tm]: A Computer Resource for Molecular Biology, User manual for Bionet subscribers, Release 2.3, Mountain View, CA, IntelliGenetics, 1987, p. 23, Databases available on BIONET.

[22] Douglas Brutlag, Personal communication, June 19, 1999. Steve Boswell, Los Alamos Workshop – Exploring the Role of Robotics and Automation in Decoding the Human Genome, IntelliGenetics trip report, January 9, 1987, In: Stanford Special Collections, Brutlag Papers, Fldr. BIONET.

searchers would turn to other data sources, the NIH insisted that IntelliGenetics solve the problem.[23] IntelliGenetics proposed to solve this problem by automating the submission of gene and protein sequences. Instead of laboriously searching the published scientific literature for sequence data, rekeying them into a GenBank standard electronic format, and checking them for accuracy, which was the standard method employed at that time, IntelliGenetics automated the submission procedure with an online submission program, XGENPUB (later called "AUTHORIN").[24]

Creating a new culture requires both the carrot and the stick. Making the online programs available and easy to use was one thing. Getting all molecular biologists to use them was another. In order to doubly encourage molecular biologists to comply with the new procedure of submitting their data online, the major molecular biology journals agreed to require evidence that the data had been submitted before they would consider a manuscript for review. 'Nucleic Acids Research' was the first journal to enforce this transition to electronic data submission.[25] With these new policies and networks in place, BIONET was able to reduce the time from submission to publication and distribution of new sequence data from two years to 24 hours. As noted above, just a few years earlier, at the beginning of BIONET, there were only 10 million base pairs published, and these had been the result of several years' effort. The new electronic submission of data generated 10 million base pairs a month.[26] Walter Gilbert may have angered some of his colleagues at the 1987 Los Alamos Workshop on Automation in Decoding the Human Genome when he stated that, "Sequencing the human genome is not science, it is production".[27] But he surely had his finger on the pulse of the new biology.

The matrix of biology

The explosion of data on all levels of the biological continuum made possible by the new biotechnologies and represented powerfully by organizations such as BIONET was a source of both exhilaration and anxiety. Of primary concern to many biolo-

[23] Barbara H. Duke, Contracting Officer, NIH, to IntelliGenetics, Inc., Request for Revised Proposal in Response to Request for Proposals RFP No. NIH-GM-87-04 entitled "Nucleic Acid Sequence Data Bank", June 3, 1987, Letter with attachment, Stanford Special Collections, Brutlag Papers, Fldr. BIONET.

[24] Douglas L. Brutlag, and David Kristofferson, BIONET: An NIH Computer Resource for Molecular Biology, Biomolecular Data: A Resource in Transition, ed. by R. R. Colwell, Oxford 1988, pp. 287–294. Also see, Automatic Data Submission to GenBank, EMBL, and NBRF-PIR, BIONET News, Vol. 1, No. 1, April 1988.

[25] Ibid.

[26] Douglas Brutlag, Personal Communication, June 19, 1999. See nomination for Smithsonian-Computerworld Award in Stanford Special Collections, Brutlag Papers, Fldr. Smithsonian Computerworld Award.

[27] Quoted from Steve Boswell, "Los Alamos Workshop – Exploring the Role of Robotics and Automation in Decoding the Human Genome" (fn. 22).

gists was how best to organize this massive outpouring of data in a way that would lead to deeper theoretical insight, perhaps even a unified theoretical perspective for biology. The National Institutes of Health were among those most concerned about these issues, and they organized a series of workshops to consider the new perspectives emerging from recent developments. The meetings culminated in a report chaired by Harold Morowitz entitled "Models for Biomedical Research: A New Perspective" (1985). The panelists foresaw the emergence of a new theoretical biology "different from theoretical physics, which consists of a small number of postulates and the procedures and apparatus for deriving predictions from those postulates". The new biology was far more than just a collection of experimental observations. Rather it was conceived as a vast array of information gaining coherence through organization into a multi-dimensional matrix of biological knowledge,[28] the complete data base of published biological experiments structured by the laws, empirical generalizations, and physical foundations of biology and connected by all the interspecific transfers of information.[29] Moreover this matrix of biological knowledge would be tied to the use of computers, which would be required to deal with the vast amount and complexity of the information.[30]

In its Long Range Plan of 1987 the Board of Regents of the National Library of Medicine further elaborated on the notion of the matrix of biological knowledge explicitly in terms of fashioning the new biology as an information science.[31] In the view of the panel, the field of molecular biology was opening the door to an era of unprecedented understanding and control of life processes, particularly through the "automated methods now available to analyze and modify biologically important macromolecules".[32] Due to the complexity of biological systems, basic research in the life sciences would be increasingly dependent on automated tools to store and manipulate the large bodies of data describing the structure and function of important macromolecules. According to the NIH because of new automated laboratory methods, genetic and biochemical data are accumulating far faster than they can be assimilated into the scientific literature. The problems of scientific research in biotechnology, the NIH stated, are increasingly problems of information science.[33]

To support and promote the entry into the new age of biological knowledge the NIH recommended building a National Center for Biotechnology Information to serve as a repository and distribution center for this growing body of knowledge and as a laboratory for developing new information analysis and communications tools

[28] H. Morowitz, Models for Biomedical Research: A New Perspective. Washington, D.C. 1985, p. 21.
[29] Ibid., p. 65.
[30] Ibid., p. 67.
[31] Board of Regents, NLM Long Range Plan (Report of the Board of Regents), Bethesda, MD, National Library of Medicine, (1987).
[32] Ibid., p. 26.
[33] Ibid., p. 29.

essential to the advance of the field. The proposal recommended $ 12.75 mil per year for 1988–1990, with an additional $ 10 mil per year for work in medical informatics.[34] The program would emphasize collaboration between computer and information scientists and the biomedical researcher. In addition the NIH would support research in the areas of molecular biology database representation, retrieval-linkages, and modeling systems, while examining interfaces based on algorithms, graphics and expert systems. The recommendation also called for the construction of online data delivery through linked regional centers and distributed database subsets.

Brave new theory

The recent explosive growth of sequencing data that began to become available in university and company databases, and more recently publicly through the Human Genome Initiative has produced a paradigm shift in both the intellectual and institutional structures of biology. According to some of the central players in this transformation, at the core is biology's switch from having been an observational science, limited primarily by the ability to make observations, to being a data-bound science limited by its practitioner's ability to understand large amounts of information derived from observations. To understand the data the tools of information science have not only become necessary handmaidens to theory: they have also fundamentally changed the picture of biological theory itself. To use this flood of knowledge, which will pour across the computer networks of the world, biologists not only must become computer-literate, but also change their approach to the problem of understanding life. Walter Gilbert characterizes the situation sharply:

> "The next tenfold increase in the amount of information in the databases will divide the world into haves and have-nots, unless each of us connects to that information and learns how to sift through it for the parts we need."[35]

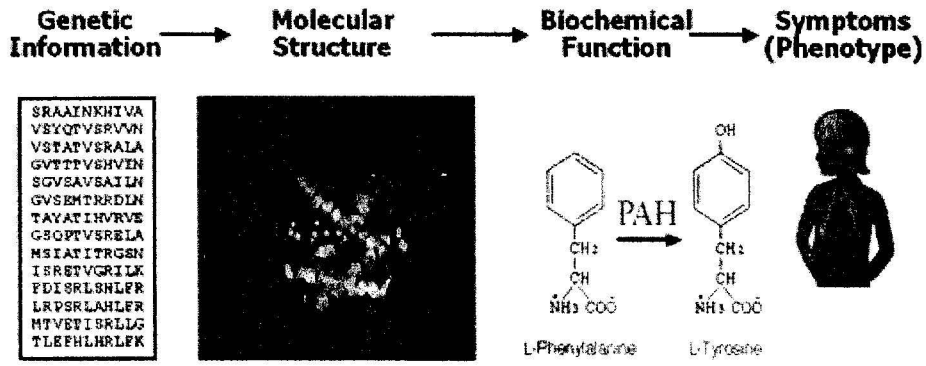
Gilbert goes on to describe the newly forming genomic view of biology:

> "The new paradigm now emerging is that all the 'genes' will be known (in the sense of being resident in databases available electronically), and that the starting point of a biological investigation will be theoretical. An individual scientist will begin with a theoretical conjecture, only then turning to experiment to follow or test that hypothesis. The actual biology will continue to be done as 'small science' – depending on individual insight and inspiration to produce new knowledge – but the reagents that the scientist uses will include a knowledge of the primary sequence of the organism, together with a list of all previous deductions from that sequence."[36]

[34] Ibid., pp. 46–47. The figures for Medical Informatics were $ 7.4, $ 9.9, and $ 13 Mil for 1988–1990.

[35] Walter Gilbert, Towards a Paradigm Shift in Biology, Nature, 349 (1991), p. 99.

[36] Ibid.



Genomics, computational biology, and bioinformatics restructure the playing field of biology, bringing a substantially modified toolkit to the repertoire of molecular biology skills developed in the 1970s. Along with the biochemistry components new skills are now required, including machine learning, robotics, databases, statistics and probability, artificial intelligence, information theory, algorithms, and graph theory.[37]

Proclamations of the sort made by Gilbert and other promoters of genomics may seem like hyperbole. But the Human Genome Initiative and the information technology that enables it has changed molecular biology in fundamental ways, and indeed, may suggest similar changes in store for other domains of science. The online DNA and protein databases I have described have not just been repositories of information for insertion into the routine work of molecular biology, and the software programs discussed in connection with IntelliGenetics and GenBank are more than retrieval aids for transporting that information back to the lab. As a set of final reflections, I want to look in more detail at some ways this software has been used to address the problems of molecular biology in order to gain a sense of the changes taking place.

Biology in silicio

A dramatic illustration of how sequence alignment tools can be brought to bear on determining function and structure is provided by the case of cystic fibrosis. Cystic fibrosis is caused by aberrant regulation of chloride transport across epithelial cells in the pulmonary tree, the intestine, the exocrine pancreas, and apocrine sweat

[37] These are the disciplines graduate students and postdocs in molecular biology in Brutlag's lab at Stanford are expected to work with. Source: Douglas Brutlag, Department Review: Bioinformatics Group, Department of Biochemistry, Stanford University, 1998, personal communication.

glands. This disorder was identified as due to defects in the cystic fibrosis transmembrane conductance regulator protein (CFTR). The CFTR gene was isolated in 1989, and subsequently identified as producing a chloride channel whose activity depends on phosphorylation of particular residues within the regulatory region of the protein. Using computer-based sequence alignment tools of the sort described above, it was established that a consensus sequence for nucleotide binding folds that bind ATP are present near the regulatory region and that 70 percent of cystic fibrosis mutations are accounted for by a 3 base-pair deletion that removes a phenylalanine residue within the first nucleotide binding position. A significant portion of the remainder of cystic fibrosis mutations affect a second nucleotide-binding domain near the regulatory region.[38]

In working out the folds and binding domains for the CFTR protein Hyde, Emsley, Hartshorn, et al. (1990) used sequence alignment methods similar to those available in early models of the IntelliGenetics software suite.[39] In 1992 IntelliGenetics introduced BLAZE, an even more rapid search program running on a massively parallel computer. As an example of how computational genomics can be used to solve structure-function problems in molecular biology, Brutlag repeated the CFTR case using BLAZE.[40] A sequence similarity search compared the CFTR protein to more than 26,000 proteins in a protein database of more than 9 million residues, resulting in a list of 27 top similar proteins, all of which strongly suggested the CFTR protein is a membrane protein involved in secretion. Another feature of the comparison result was that significant homologies were shown with ATP-binding transport proteins, further strengthening the identification of CFTR as a membrane protein. The search algorithm identified two consensus sequence motifs in the protein sequence of the cystic fibrosis gene product that corresponded to the two sites on the protein involved in binding nucleotides. The search also turned up distant homologies between the CFTR protein and proteins of *E. coli* and yeast. The entire search took three hours. Such examples offer convincing evidence that tools

[38] S. C. Hyde/P. Emsley, et al. (1990). Structural Model of ATP-binding Proteins Associated with Cystic fibrosis, Multidrug Resistance and Bacterial Transport, Nature 346, pp. 362–365; B. S. Kerem/J. M. Rommens, et al. Identification of the Cystic Fibrosis Gene: Genetic Analysis, Science 245 (1989), pp. 1073–1080; B. S. Kerem/J. Zielenski, et al., Identification of Mutations in Regions Corresponding to the Two Putative Nucleotide (ATP)-Binding Folds of the Cystic Fibrosis Gene, Proceedings of the National Academy of Sciences 87 (1990), pp. 8447–8451; J. R. Riordan/J. M. Rommens, et al., Identification of the Cystic Fibrosis Gene: Cloning and Characterization of Complementary RNA, Science 245 (1989), pp. 1066–1073.

[39] Hyde/Emsley, et al. used the Chou-Fasman algorithm (1973) for identifying consensus sequences and the Quantatm modeling package produced by Polygen Corp., Waltham, Mass. for modeling the protein and its binding sites. See S. C. Hyde/P. Emsley, et al. (1990). Structural Model of ATP-binding Proteins Associated with Cystic fibrosis, Multidrug Resistance and Bacterial Transport, Nature 346, pp. 362–365.

[40] D. Brutlag, Understanding the Human Genome, Scientific American Introduction to Molecular Medicine, ed by P. Leder/D. A. Clayton/E. Rubenstein, New York 1994: pp. 164–166.

of computational molecular biology can lead to the understanding of protein function.

The methods for analyzing sequence data discussed above were just the beginnings of an explosion of database mining tools for genomics that is continuing to take place.[41] In the process biology is becoming even more aptly characterized as an information science.[42] Advances in the field have led to large-scale automation of sequencing in genome centers employing robots. In order to keep pace with this flood of data emerging from automated sequencing, genome researchers have in turn looked increasingly to artificial intelligence, machine learning, and even robotics in developing automated methods for discovering patterns and protein motifs from sequence data. Many molecular biologists who welcomed the Human Genome Initiative with open arms undoubtedly believed that when the genome was sequenced everyone would return to the lab to conduct their experiments in a business-as-usual fashion, empowered with a richer set of fundamental data. The developments in automation, the resulting explosion of data, and the introduction of tools of information science to master this data have changed the playing field forever: in the words of genome scientist Hans Lehrach, there may be no "lab" to return to. In

[41] See for instance the National Institute of General Medical Science, (NIGMS), Protein Structure Initiative Meeting Summary, April 24, 1998, at: http://www.nih.gov/nigms/news/reports/protein_structure.html

[42] I have focused on the development of software in this discussion. But a further crucial stimulation to the takeoff of bioinformatics, of course, are hardware and networking developments. The growth of databases and complexity of the searches that were to be undertaken stimulated the demand for faster algorithms, more powerful computer systems, and network bandwidth. At the beginning of this "bioinformatics revolution" in the 1970s, for example, a search on a DNA sequence of typical size would be performed by a computer capable of performing one million instructions per second (one MIP) and would take approximately 15 minutes. Throughout the late 1970s and 1980s mini-computers and personal computer workstations continued to increase in power at about the same rate as the growth of the databases, so that a typical search still took around 15 minutes. By the end of the 1980s, however, the growth in sequence data – now hundreds of megabytes in size – had overtaken the ability of computers to search it with acceptable turnaround time. Shortcut search methods and more efficient code helped, but the most rigorous and sensitive searches began to require hours of computing time to align and score even a single query sequence against a database of sequences. The NIH and NSF responded to the challenge by supporting research and development of new computer architectures, regional supercomputer centers and several large-scale computing initiatives. (See Thomas P. Hughes, et al., ed., Funding a Revolution: Government Support for Computing Research, Washington, D.C. 1999.) Commercial vendors such as DEC, SUN Microsystems, Cray Computers, and MasPar Computer Corporation tried to meet the large-scale computing needs of geneticists with, for example, massively parallel computers, such as the MasPar MP-1 computer. In early 1992, the MasPar MP-1104 with 4,096 processors could search the entire Swiss-Protein database in 30 seconds with a query of 100 amino acids, and a query of 1000 amino acids could be executed on the GenBank database (74,000 sequences) in 15 minutes. (See IntelliGenetics, Inc., and MasPar Computer corporation, BLAZE: A Massively Parallel Sequence Similarity Search Program for Molecular Biologists, Product Information Bulletin, May 1992.)

its place is a workstation hooked to a massively parallel computer, producing simulations by drawing on the data streams of the major databanks and carrying out "experiments" *in silico* rather than *in vitro*. Elizabeth Eisenstein, Bruno Latour, and Adrian Johns have argued that a pre-condition for science as we know it is the elaborate apparatus and organization of practice for transcribing nature into a form compatible with institutions of the letter; and in his own work on a center for molecular biology research, Latour argued – now famously – that modern scientific laboratories are in effect inscription devices and that the din of their operation is carefully hidden and ultimately silenced in the production of scientific facts.[43] I have argued that the fusion of the laboratory and contemporary forms of computer-mediated communication offers a new – perhaps final – twist to this position by the erasure of the wet lab from the academy altogether. The result of biology's metamorphosis into an information science just may be the relocation of the lab to the industrial park and the dustbin of history.

[43] Bruno Latour/Steve Woolgar, Laboratory Life: The Construction of Scientific Facts, Princeton 1979, second ed. 1986.