Information Services & Use 42 (2022) 107–115
DOI 10.3233/ISU-210139
IOS Press

Don Lindberg and the creation of the National Center for Biotechnology Information

Daniel R. Masys[a,*] and Dennis A. Benson[b]
[a]*University of Washington School of Medicine, Seattle, Washington, USA*
[b]*U.S. National Library of Medicine*

Abstract. The highest priority new initiative resulting from the 1985–86 National Library of Medicine Long Range Planning exercise initiated by NLM Director Dr. Donald A.B. Lindberg was the creation of new information resources and services related to molecular biology and genetics, termed "biotechnology information". Beginning with existing NLM resources and research projects associated with molecular data, and with Lindberg's enthusiastic support, the institution launched a Congressionally-mandated Center that has become an essential part of 21st century biomedical science.

Keywords: Donald A.B. Lindberg, U.S. National Library of Medicine, National Center for Biotechnology Information, Human Genome Project, Genomics

1. Background

The genesis of the U.S. National Library of Medicine's (NLM) National Center for Biotechnology Information (NCBI) was intimately interwoven with the 1985–86 NLM Long Range Planning effort that was one (and perhaps the most notable) of NLM Director Dr. Donald A.B. Lindberg's signature initiatives in his first years at NLM. Shortly after he arrived as the newly appointed NLM Director in August of 1984, and with the endorsement of L. Thompson Bowles M.D., Ph.D., the chairman of the NLM Board of Regents, who had been appointed that same month, Lindberg convened more than 120 professionals in a visioning exercise that asked them to imagine what NLM's world would look like in 20 years and what its future services should be. They were also charged with identifying 10-year milestones on the path to that 20-year future, and describing the challenges and impediments to be faced in realizing that vision. Most importantly, participants were asked to identify "windows of opportunity" for the institution, for new programs and resources that could be started immediately and could be finished within five years. As part of the planning process, all involved were reminded of the humbling observation that most long range planning efforts systematically overstated what could be achieved in five years, while they fell short of actual achievements by 10 years, and completely missed key innovations and social changes that would

*Corresponding author: Daniel R. Masys, M.D., Affiliate Professor, Department of Biomedical Informatics & Medical Education, Box 358047, University of Washington School of Medicine, Seattle, WA, USA. E-mail: dmasys@uw.edu.

become the real determinants of the future two decades later. With the luxury of now more than 35 years of hindsight, those NLM long range planners can take some pride in being approximately correct far more often than they were precisely wrong.

The Long Range Planning effort under Lindberg's direction was divided into five topic areas: building and organizing the Library's collection; locating and gaining access to medical and scientific literature; obtaining factual information from data bases; medical informatics; and assisting health professions education through information technology. Seventy-seven professionals with expertise across these areas were appointed to five planning panels that each met several times over a year beginning in the Fall of 1985. The draft Long Range Plan was reviewed and approved by the NLM Board of Regents in January 1987 and became an active roadmap for the Library's major programs for the ensuing two decades and beyond [1].

Of special importance to the creation of NCBI were the discussions and recommendations of Panel 3: Obtaining factual information from data bases. The 16 individuals appointed to this panel included two current and one future Nobel laureates: Joshua Lederberg, Allan Maxam, and Richard Roberts. Lederberg had won the 1958 Nobel prize in Physiology or Medicine for "discoveries concerning genetic recombination and the organization of the genetic material of bacteria" [2]. Maxam, along with Walter Gilbert, Frederick Sanger, and Paul Berg, had shared the 1980 Nobel prize in Chemistry for devising a technique to sequence DNA [3]. Roberts shared with Phillip Sharp the 1993 Nobel prize in Physiology or Medicine for demonstrating how the RNA produced by transcription of DNA can be divided up into introns and exons, after which the exons can be joined together [4]. The background and expertise of these three researchers in molecular genetics heavily influenced the depth of the planning discussions and the vision for NLM's future.

One event in particular profoundly influenced Don Lindberg and may justifiably be considered a turning point in NLM history. During one of the early face-to-face Planning Panel 3 meetings, with Lindberg present as an observer, Allan Maxam gave a spontaneous "chalk talk" on the challenges confronting researchers who were attempting to understand the molecular underpinnings of health and disease. He began by listing commonly used research databases in molecular biology and genetics, organized in a size hierarchy that went from intact cells and tissues down through individual DNA and RNA nucleotides, and included small molecules that modulate the production and functioning of genes and their protein products. The diagram he drew (Fig. 1), which with refinements was published in the Long Range Plan, came to be known within NLM as the "Tower of Babel" picture, for it highlighted the lack of naming consistency and interconnections among research databases constructed by different organizations. The incompatibility of these closely related scientific resources thwarted a researcher's ability to use similarities and insights from one database to explain findings recorded in another, and contrasted with the scientific literature where a single experiment might produce data that was then included in several disparate databases.

Maxam illustrated the promise of molecular biology computing with a story from the research literature of the time related to oncogenes, which are genes that can induce cancerous behavior in cells. He noted that databases such as GenBank that contain DNA sequences enable researchers to use computer-based analysis to calculate the similarity of genes to one another, sometimes providing powerful and unexpected insights. Such was the case with the v-sis "proto-oncogene" that was found by computer matching in the early 1980s to be nearly identical with a normal growth and development gene called "platelet derived growth factor". This finding gave rise to the key biological insight that cancer-causing genes might in some cases be normal genes simply switched on at the wrong time. Maxam noted that this ability to "reason by analogy" often depended upon findings at different levels of the biologic hierarchy depicted in his

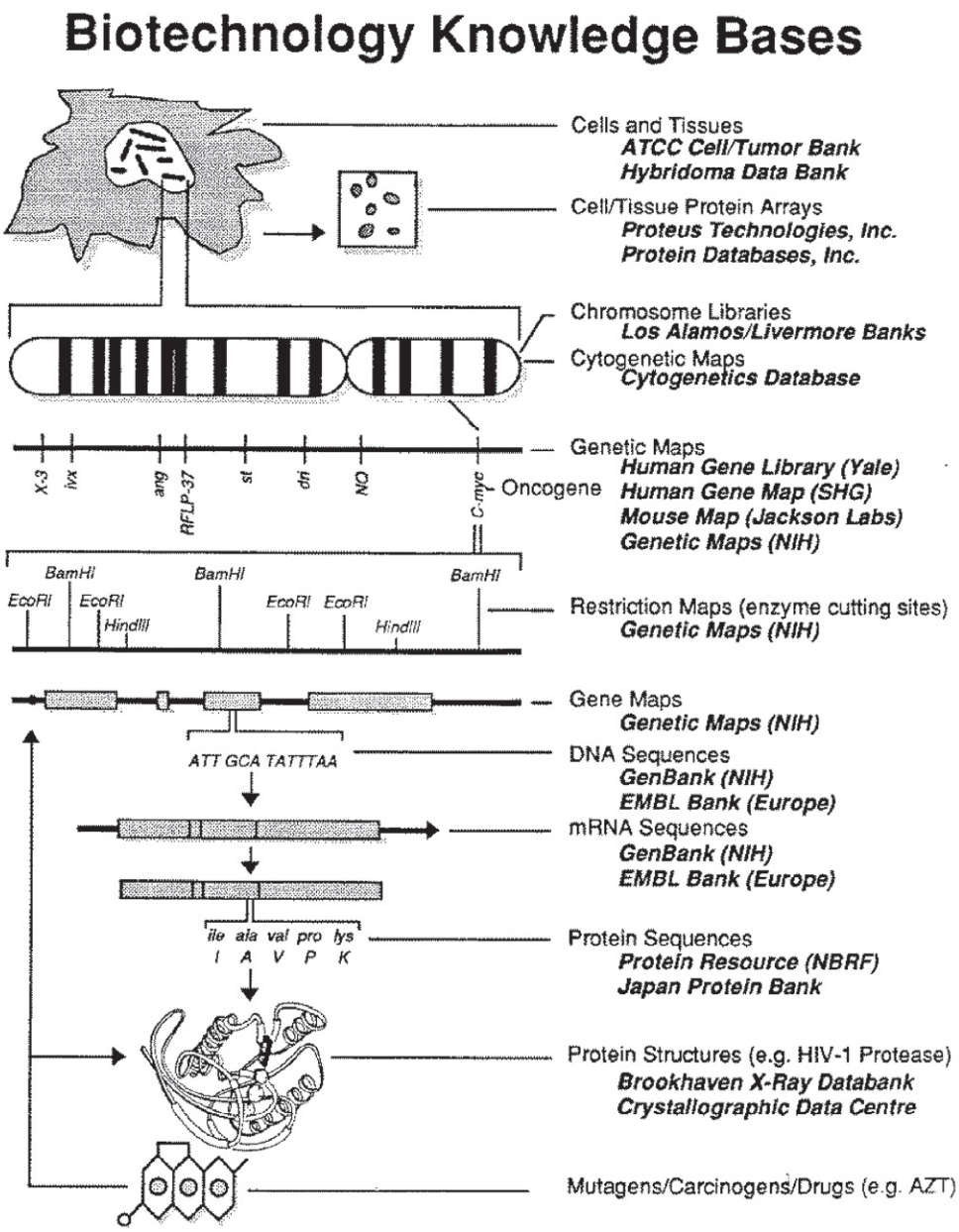

Fig. 1. Biotechnology Knowledge Bases, adapted from Allan Maxam's Long Range Plan presentation, 1987 [5].

diagram, and that there were few if any automated tools capable of finding such correlations across the dozens of databases storing molecular information and its interpretations.

Don Lindberg was immediately and enduringly impressed by this presentation and the opportunity it portrayed for NLM to help guide, structure, and link related scientific resources in pursuit of better understanding human health and disease. The "oncogene story" was included in the planning panel's report, which included the following observations [6]:

"The general area of biogenetics is moving ahead rapidly. Serious proposals have been put forward to sequence the entire human genome and to map active chromosomal regions for each tissue type in different organs systems…. The research-oriented information systems currently in place are adequate to ask low level questions: Find the degree of similarity between base-pair sequences. The next questions are: What do the differences mean? Current data bases are being used to support modeling and theory, but the tools are very primitive, and no methods exist for automatically suggesting links across levels. There is a vacuum in the area of research into ways of using information by interconnecting various levels…. Currently, no organization is taking the lead in promoting keys and standards by which the information from the related research data bases in the accompanying [Tower of Babel] figure can be systematically interlinked or retrieved by investigators."

The report went on to note "A singular and immediate window of opportunity exists for the Library in the area of molecular biology information. Because of new automated laboratory methods, biological data are accumulating far faster than they can be assimilated into the scientific literature. The problems of scientific research in the field of molecular biology are increasingly problems of information science".

The Long Range Plan included the following Recommendation:

"3.2.1. Immediately establish an intramural and extramural program for biotechnology information. The intramural component should be a National Center for Biotechnology Information, to serve both as a repository and distribution center for the growing body of knowledge and as a laboratory for developing new information analysis and communications tools essential to continued advancement in this field…. Because of the technical complexity in this scientific area and the expectation that data production will increase by a thousand times in the next five years, a major new activity is required." Lindberg's own words on the subject were included as a Preface to the Plan: "Of the numerous initiatives the plan proposes … one in particular stands out. This is the "window of opportunity" presented to the Library in the field of molecular biology and biotechnology. Attention to this opportunity - through the provision of advanced information handling services - will permit NLM to contribute significantly to discovery of new principles and treatments by health-care professionals and scientists" [1].

By the time the Long Range Plan was published, Lindberg was already taking steps to implement its key recommendations. One of these steps was appointing a new director of the Lister Hill National Center for Biomedical Communications (LHNCBC), which since its 1968 creation had been the intramural research and development division of the Library. The previous director, Dr. Richard B. Friedman, had left the NLM for a faculty appointment at the University of Wisconsin in 1984, shortly before Don Lindberg's arrival.

Lindberg recruited Daniel Masys M.D. to the LHNCBC director position in the spring of 1986 by inviting him to an informal visit to the NLM Director's office. He presented the proposition that "we cannot make the progress needed in biotechnology information without you". Masys was at that time the chief of the International Cancer Research Databank branch of the National Cancer Institute (NCI). He had participated in the Long Range Plan exercise as an appointed member of the Factual Data Bases panel. As a physician trained in hematology and medical oncology, Masys was familiar with the nascent and rapidly evolving science of "molecular medicine" and had come to NIH to help design the NCI's Physician Data Query (PDQ) computer system. PDQ was a continuously updated online resource of cancer information for both physicians and patients [7]. The offer of the LHNCBC directorship was readily accepted. The personnel action was straightforward because Masys was a commissioned officer in the U.S. Public Health Service and his NLM appointment was simply a re-assignment.

2. Building on existing programs and staff

The Long Range Plan recommended a $9.7 million annual budget increase devoted to biotechnology information services and 34 additional full time equivalent NLM personnel [1]. Such a significant expansion could not be achieved by reallocating existing budgets, and would necessarily involve supplemental appropriations by the U.S. Congress. That notwithstanding, the work needed to begin immediately. The obvious and immediate path forward was to organize current staff and existing information resources and research projects to become a platform for future growth and enhancement.

The largest and most widely used of NLM's resources was the MEDLINE database of bibliographic citations and their associated index terms and author abstracts. Essentially all factual databases in molecular biology included citations to the published articles that reported the data included in each factual database record. This feature gave NLM a powerful mechanism for linking disparate research databases. At Don Lindberg's direction, an early enhancement to MEDLINE added the "reverse pointer" of the external database name (e.g., GenBank) and the external database record unique identifier to MEDLINE citation records. With this linking data, both human users and computer programs could begin with a search of the scientific literature in MEDLINE and then navigate to the actual data reported by the article in the externally linked data base.

NLM had the good fortune of existing intramural research projects conducted by LHNCBC staff that had a focus in clinical and molecular genetics. The most notable was the Online Reference Works project whose goal was to produce electronic authoring systems for complex publications such as biomedical textbooks. The model envisioned a "scholar's workstation" that helped an author write and maintain a large and evolving corpus of knowledge, output phototypesetting files that would generate the printed copy of the monograph, and serve also as a searchable database of full text [5].

At the same time, the Welch Medical Library at Johns Hopkins had contacted NLM for assistance on behalf of Victor McKusick M.D., the author of *Mendelian Inheritance in Man* (MIM), an 1,800 page, comprehensive compendium of information of human genes and genetic phenotypes that was in its sixth edition in the early 1980s [8]. Dr. McKusick was heroically attempting to edit and maintain the text personally by reading the newly published literature every day and using manual, paper-based notations to update the monograph. Using MIM as its test case, in 1984 the Online Reference Works project created a multimedia system called IRx (Information Retrieval Experiment) that contained both text and images such as gene maps and clinical photos of genetic diseases. This work preceded the Long Range Planning exercise and provided a fertile research and development environment for new methods of linking genetic-related data [9].

Systems such as the IRx prototype were possible only because of the technical expertise of LHNCBC staff, and the project benefitted from the leadership of an individual who would become one of NCBI's senior leaders. Dennis Benson Ph.D. was a neuroscientist working as a Research Associate at Johns Hopkins School of Medicine Department of Biomedical Engineering, conducting research in auditory neurophysiology. His work involved extensive programming of laboratory instrumentation. He came to NLM's Lister Hill Center in January of 1980 and was the technical lead for the IRx project when the request for help with MIM came. As a result, he was already immersed in the world of computers and genetics when the Long Range Plan was completed.

Drs. Masys and Benson, with Dr. Lindberg's boundlessly enthusiastic support, became the early "biotechnology information project" team as soon as Masys arrived in July 1986. They viewed their principal tasks as outreach to better understand the needs of researchers, and increasing the visibility of NLM's newfound interest in advancing molecular genetics and biotechnology-related research. This

mission led to much travel and participation in domestic and international scientific meetings, where they would present NLM's current and planned services.

At one of these events, another fortuitous connection occurred that would mold NCBI's programs. Masys and Benson gave a presentation at a 1986 workshop on "Genes and Computers" held in Waterville Valley, New Hampshire. At the participant lunch that followed, graduate student James Pustell introduced himself and expressed an interest in NLM's plans. Pustell was studying for his Ph.D. in molecular biology with Dr. Fotis Kafatos at Harvard. An energetic self-starter and practitioner of what would become widely known as "bioinformatics", Pustell had written a suite of molecular sequence analysis programs for microcomputers that were marketed by International Biotechnologies, Inc. [10]. A natural "meeting of the minds" occurred that day, which eventuated in Pustell (later to become James Ostell on the occasion of his marriage to Kate Oster) becoming one of the founding senior leaders and guiding lights of NCBI.

As noted in the NLM Long Range Plan, there was a growing conviction among scientists that a project to sequence the entire human genome would soon become both feasible, due to new automated sequencing technologies, and immensely valuable in understanding human health and disease. In October 1986, NIH Director James Wyngarden convened an NIH Director's Advisory Committee meeting on the role that NIH should play in a federally-supported Human Genome Project [11]. NLM was represented by Drs. Lindberg (meeting co-chair), Benson, and Masys, who presented a vision of NLM assuming a central role in managing the data arising from such a project, including the possibility of NLM taking over management of the GenBank DNA sequence database. The response from the director of the National Institute for General Medical Sciences, at that time the GenBank sponsor, was emblematic of the organizational challenges to be faced: "Of course, you don't want librarians taking care of gene sequences." It was an example of the common lack of appreciation that NLM was no ordinary medical library. On the contrary, many on NLM's staff were doctoral level scientists who were trained in the fields of medical and biological information sciences.

The first molecular biology research tools implemented at NLM were hosted on a server computer within the Lister Hill Center. It provided a simple text-based menu interface for searching several databases shown in Fig. 1, and analysis tools for nucleotide sequence comparisons and alignments that could be used on the record sets retrieved from those databases. Improving researchers' awareness of and skills in using computerized analysis was (and remains) an important component of the overall NLM biotechnology information program. To this end, NLM began collaborating with other NIH institutes and hosting onsite hands-on workshops in LHNCBC's Educational Technology Branch classrooms.

One series of these educational programs was taught by David Lipman M.D., a researcher in the Mathematical Research Branch of the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). As these NLM programs were starting, Lipman was already an internationally prominent developer of methods for sequence database searching and determination of molecular sequence similarity, including the Wilbur–Lipman algorithm in 1983 and FASTA search in 1985 [12,13]. With Lipman's participation in designing and implementing analytic programs, the senior leadership team was now in place to establish NCBI as a global resource for molecular biology information.

3. The legislative road to creation of NCBI

Don Lindberg and NLM Deputy Director Kent Smith were well equipped to confront the challenges of creating new programs whose size and resource needs would require additional authorization and

appropriations by the U.S. Congress. Lindberg had a natural talent for communicating complex scientific issues in ways that members of Congress and their staffs easily understood, and an ability to give public testimony in congressional hearings that conveyed the public benefit to be gained by new programs, while adhering to federal agency restrictions that prohibited employees from advocating for larger appropriations.

Every piece of new legislation also needs legislative champions, and NLM benefitted from the longstanding friendship between one of its employees, Frances Humphrey Howard (sister of Hubert Humphrey, U.S. Vice President from 1965 to 1969), and Claude Pepper (D-FL), who was a U.S. Senator (1936–1951) and a Congressman (1963–1989). A vocal advocate for medical research and particularly the health and welfare of the elderly, Pepper learned of the proposed Human Genome Project and viewed it as a "Manhattan Project" for health. The legislative process also benefitted from the outreach and educational efforts of the newly-formed "Friends of the National Library of Medicine".

In 1986 and 1987, Pepper introduced bills to create a National Center for Biotechnology Information located within NLM. With the help of colleagues in the House and Senate, his bill was incorporated into the NIH reauthorization legislation known as the Health Omnibus Extension Act (Public Law 100-607), which was signed into law by President Ronald Reagan on November 4, 1988 [14].

The mission given to NCBI in its founding legislation was the following:

"(1) design, develop, implement, and manage automated systems for the collection, storage, retrieval, analysis, and dissemination of knowledge concerning human molecular biology, biochemistry, and genetics;

"(2) perform research into advanced methods of computer-based information processing capable of representing and analyzing the vast number of biologically important molecules and compounds;

"(3) enable persons engaged in biotechnology research and medical care to use systems developed under paragraph (1) and methods described in paragraph (2); and

"(4) coordinate, as much as is practicable, efforts to gather biotechnology information on an international basis."

Another key meeting took place during this time. With Dr. Lindberg's concurrence and blessing, Drs. Masys and Benson walked over to the intramural offices of the NIDDK Mathematical Research Branch, and made the same proposition to David Lipman that Lindberg had made to Masys several years earlier: "We cannot make the progress needed in molecular biology and bioinformatics without you." At that point, Lipman agreed only to discuss the opportunity with Lindberg. But the rest, as the saying goes, is history.

Lipman accepted the newly created position of NCBI director in 1989. His personnel action was as straightforward as for Masys, since he too was a commissioned officer in the U.S. Public Health Service, and could simply be reassigned with the concurrence of both NIH institutes. With Lindberg's support and encouragement, Lipman formed his initial senior leadership team: Dennis Benson, James Ostell, and David Landsman, Ph.D., who was at the time an intramural research scientist at NCI.

NCBI became not only a global resource for molecular biology and genetics, but also a brain trust for the redesign and modernization of NLM's other services, including MEDLINE, its flagship literature resource. NCBI began with the subset of MEDLINE records linked to factual databases in biotechnology such as GenBank, and transformed the MEDLINE unit record design into a relational data model that enabled use of highly scalable relational database management systems. This redesign was then extended to the entire MEDLINE citation collection, and with a web-compatible search interface became the basis for PubMed, a system that beginning in 1997 provided free public access to MEDLINE. As it gained technical expertise, NCBI's Information Engineering Branch, led by Jim Ostell, created and deployed

internet accessible systems that could process thousands of simultaneous queries *per second*. This made the staff highly valued consultants for essentially all of NLM's online information services.

For 28 years, until retiring in 2017, Lipman and his fellow NCBI leaders translated NLM's interest in advancing molecular science into tangible and widely used resources and tools for researchers worldwide. By all measures, the organization has exceeded the goals originally envisioned by Don Lindberg and the Long Range planners, and its services have become woven into the fabric of 21st century science, continuing to catalyze biomedical research on a global scale.

Upon Lipman's retirement, James Ostell was appointed NCBI's second director, taking over a staff that had grown from less than a dozen when Lipman, Benson, Ostell, and Landsman started, to more than 700. Ostell retired in 2020 after 32 years at NCBI. As of this writing, Dennis Benson, in his capacity as NCBI Deputy Director and with 41 years of NLM service, and David Landsman, with 32 years as Chief of the NCBI Computational Biology Branch, are the organization's most experienced leaders.

Choosing to make the creation of NCBI the highest priority of the original NLM Long Range Plan was distinctively Don Lindberg: Recognize the right idea at the right time, and create programs and organizations that grow and are sustainable over time. The long tenure and sustained institutional loyalty of NCBI's founding leaders has also been a testament to Don Lindberg's ability to recognize talented individuals, recruit them for important institutional missions, and then do his best to provide them with both the resources and the freedom they needed to manage their programs in whatever way they found most effective. Lindberg's style of "trust and delegate" informed his interactions with NLM senior staff throughout his 31 years as NLM Director.

References

[1] National Library of Medicine. Long range plan/report of the Board of Regents [Internet], U.S. Dept. of Health and Human Services, National Institutes of Health, Bethesda, MD, January 1987 [cited 2021 May 17]. Available from: http://resource.nlm.nih.gov/101646837.

[2] NobelPrize.org. The Nobel Prize in Physiology or Medicine 1958 [Internet]; Nobel Media AB 2021 [cited 2021 May 17]. Available from: https://www.nobelprize.org/prizes/medicine/1958/summary/.

[3] NobelPrize.org. The Nobel Prize in Chemistry 1980 [Internet]; Nobel Media AB 2021 [cited 2021 May 17]. Available from: https://www.nobelprize.org/prizes/chemistry/1980/summary/.

[4] NobelPrize.org. The Nobel Prize in Physiology or Medicine 1993 [Internet]; Nobel Media AB 2021 [cited 2021 May 17]. Available from: https://www.nobelprize.org/prizes/medicine/1993/summary/.

[5] National Library of Medicine. Obtaining factual information from data bases/report of Panel 3, National Library of Medicine Long Range Plan [Internet], U.S. Dept. of Health and Human Services, National Institutes of Health, Bethesda, MD, 1986 [cited 2021 May 17]. Available from: http://resource.nlm.nih.gov/8706413.

[6] National Library of Medicine. Obtaining factual information from data bases/report of Panel 3, National Library of Medicine Long Range Plan [Internet], U.S. Dept. of Health and Human Services, National Institutes of Health, Bethesda, MD, 1986 [cited 2021 May 17], p. 18. Available from: http://resource.nlm.nih.gov/8706413.

[7] S.M. Hubbard, N.B. Martin, L.W. Blankenbaker, R.J. Esterhay Jr, D.R. Masys, D.E. Tingley et al., The Physician Data Query (PDQ) cancer information system, *J Cancer Educ* **1**(2) (1986), 79–87. doi:10.1080/08858198609527818. PMID: 3079208.

[8] V.A. McKusick, *Mendelian Inheritance in Man: Catalogs of Autosomal Dominant, Autosomal Recessive, and X-linked Phenotypes*, 6th ed. Johns Hopkins University Press, Baltimore, 1983.

[9] C.M. Goldstein, Online reference works (ORW), *Bull Am Soc Inf Sci* **15**(5) (1989), 9–11, PMID: 10293125.

[10] J. Pustell and F.C. Kafatos, A convenient and adaptable package of DNA sequence analysis programs for microcomputers, *Nucleic Acids Res* **10**(1) (1982), 51–59. doi:10.1093/nar/10.1.51. PMID: 6278412; PMCID: PMC326113.

[11] NIGMS Director memo to NIH Director on Sequencing the Human Genome [Internet], July 2, 1986 [cited 2021 May 17]. Available from: https://repository.library.georgetown.edu/bitstream/handle/10822/556971/78%20Kirschstein%201986%20Memo%20to%20NIH%20Director.pdf.

[12] W.J. Wilbur and D.J. Lipman, Rapid similarity searches of nucleic acid and protein data banks, *Proc Natl Acad Sci USA* **80**(3) (1983), 726–730. doi:10.1073/pnas.80.3.726. Bibcode:1983 PNAS...80..726W. PMC 393452, PMID 6572363.

[13] D. Lipman and W. Pearson, Rapid and sensitive protein similarity searches, *Science* **227**(4693) (1985), 1435–1441. doi:10.1126/science.2983426. Bibcode:1985Sci...227.1435L. PMID 2983426.

[14] K. Smith, A brief history of NCBI's formation and growth. in: *The NCBI Handbook*, 2nd ed. National Center for Biotechnology Information (US), Bethesda, MD, 2013, [Internet]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK148949/.