Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954–1965

Strasser, Bruno J.

How to cite

STRASSER, Bruno J. Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954–1965. In: Journal of the history of biology, 2010, vol. 43, n° 4, p. 623–660. doi: 10.1007/s10739-009-9221-0

This publication URL: https://archive-ouverte.unige.ch/unige:16825
Publication DOI: 10.1007/s10739-009-9221-0

Journal of the History of Biology
DOI 10.1007/s10739-009-9221-0

Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's *Atlas of Protein Sequence and Structure*, 1954–1965

BRUNO J. STRASSER
Program in the History of Science and Medicine, Section for the History of Medicine
Yale University
New Haven, CT 06511
USA
E-mail: bruno.strasser@yale.edu

Abstract. Collecting, comparing, and computing molecular sequences are among the most prevalent practices in contemporary biological research. They represent a specific way of producing knowledge. This paper explores the historical development of these practices, focusing on the work of Margaret O. Dayhoff, Richard V. Eck, and Robert S. Ledley, who produced the first computer-based collection of protein sequences, published in book format in 1965 as the *Atlas of Protein Sequence and Structure*. While these practices are generally associated with the rise of molecular evolution in the 1960s, this paper shows that they grew out of research agendas from the previous decade, including the biochemical investigation of the relations between the structures and function of proteins and the theoretical attempt to decipher the genetic code. It also shows how computers became essential for the handling and analysis of sequence data. Finally, this paper reflects on the relationships between experimenting and collecting as two distinct "ways of knowing" that were essential for the transformation of the life sciences in the twentieth century.

Keywords: bioinformatics, natural history, molecular biology, database, protein sequences, computers, ways of knowing

Introduction

Collecting, comparing, and computing protein or DNA sequences are among the most prevalent practices in contemporary biomedical research. They constitute a specific way of producing knowledge about the nature and the role of genes and proteins in inheritance, development, health, and disease, as well as the classification and evolution of species.

Published online: 24 December 2009

These practices rely crucially on the existence of extensive computerized sequence databases such as GenBank, which contains today more nucleotides than "the number of stars in the Milky Way," as the National Institutes of Health put it in a 2005 press release.[1] In this paper, I will reassess the origins of these practices before focusing on the work of Margaret O. Dayhoff, Richard V. Eck, and Robert S. Ledley. In 1965, they produced the first computer-based sequence collection, published as a book entitled the *Atlas of Protein Sequence and Structure*. In subsequent years under Dayhoff's leadership, the *Atlas* grew in size and popularity, becoming a common fixture in biomedical laboratories. It eventually served as a model for the nucleic acid sequence databases such as GenBank.[2]

In the historiography of the life sciences, the rise of sequence analysis has been tied to the development of the field of molecular evolution.[3] Indeed, in 1962, Emil Zuckerkandl and Linus Pauling suggested that differences in amino acid sequences between two species accumulated at a constant rate and could thus be used to measure evolutionary distances.[4] They considered sequences "documents of evolutionary history" explaining how entire phylogenies could be based on the comparison of protein sequences.[5] Working on these premises, the field of molecular evolution took shape in the 1960s, and its advocates sometimes clashed with the proponents of morphology-based evolution.[6] Here I argue that the key practices of molecular evolution – collecting, comparing, and computing sequences – were already well established by 1962, having developed during the previous decade in three unrelated fields: biochemical research on protein function, theoretical studies of the genetic code, and attempts to apply digital computers to the life sciences. I will show how Dayhoff, Eck, and Ledley took part in these endeavors and capitalized on their experience to create the *Atlas of Protein Sequence and Structure*, which became a

[1] NIH press release, August 22, 2005, "Public Collections of DNA and RNA Sequence Reach 100 Gigabases."

[2] Strasser, 2006a, b, c, 2008.

[3] On the history of molecular evolution, see Dietrich, 1994, 1998; Morgan, 1998; Hagen, 1999, 2001; Aronson, 2002; Suárez-Diaz, 2007, 2009; Suárez-Diaz and Anaya-Muñoz, 2008; Sommer, 2008.

[4] Zuckerkandl and Pauling, 1962.

[5] Zuckerkandl and Pauling, 1965.

[6] For different views about this episode, see Dietrich, 1998 and Hagen, 1999.

crucial tool not only for the rise of molecular evolution, but more broadly for the experimental life sciences.

This paper speaks not only to the question of the origins of these practices and of their contributions to molecular evolution, but also to the broader issue of how we are to understand the development of the life sciences in the twentieth century, and especially their transformation into an "information science."[7] This transformation, to which the ascendancy of bioinformatics seems to attest, has been linked to the "molecularization" of biology in the middle of the twentieth century,[8] and to the rise of cybernetics after World War II, which gave cultural currency and epistemic traction to an informational understanding of living processes.[9] I wish to complement this view by bringing into focus computerized biological collections[10] such as the *Atlas of Protein Sequence and Structure*, the most direct ancestor of the databases which now form the backbone of contemporary bioinformatics. While computers began in the 1950s and early 1960s to play a limited role performing calculations in such fields of the life sciences as crystallography and numerical taxonomy,[11] the creation of the *Atlas* represented one of the earliest attempt to bring computers to bear on the management and distribution of biological information.[12] In particular, it made possible the use of sophisticated algorithms to compare large amounts of data drawn from numerous species. This approach embodied the powerful idea that computers could reveal information "hidden" in empirical results by handling them comparatively, an approach that has become integral to the contemporary experimental life sciences.

A focus on biological collections such as the *Atlas* also speaks to the much-debated question of the relationship between natural history and experimentation in the twentieth century. The standard narrative set by William R. Coleman and Garland E. Allen more than four decades ago still informs that literature, namely, that from the late nineteenth

[7] Lenoir, 1999.

[8] Kay, 1993; de Chadarevian, 1998, 2002; Morange, 2000; Gaudillière, 2002; Strasser, 2006c.

[9] Kay, 2000; Keller, 2000.

[10] For the suggestion that blood collections played a role for the rise of molecular biology, see de Chadarevian, 1998.

[11] de Chadarevian, 2002, Chap. 4 and Hagen, 2001, respectively.

[12] Not including the role of computers in the distribution of bibliographic information, for example, through MEDLARS, a computerized version of the Index Medicus made available in 1964, Rogers, 1964.

century natural history was overtaken by experimental biology.[13] The idea that the experimental approach to the study of life triumphed over the natural historical – as evidenced by the rise of molecular biology, for example – is prevalent among scientists and historians alike.[14] The practices I will examine in this paper – collecting, comparing, and computing – are characteristic of the natural historical "way of knowing,"[15] and the fact of their centrality to some of the fields which have exemplified the greatest successes of the experimental tradition – biochemistry and molecular biology, for example – comes as a surprise. Furthermore, the reliance of these disciplines on data originating from a wide range of species, an approach reflecting natural history's embrace of biological diversity, contrasts sharply with the traditional reliance of the experimental sciences on a small set of model organisms. Finally, the dependence of the new fields on centralized data collections such as the *Atlas* brings them even more in line with the natural history tradition in its use of collections such as herbariums and museums. In the conclusion of this paper, I will suggest that this paradox should lead us to revisit our assumptions about the development of the life sciences in the twentieth century.

Comparing Sequences to Understand Protein Function

The practice of comparing sequences emerged as soon as protein sequences began to be determined and grew as a standard method among protein biochemists in the 1950s. The hypothesis behind the comparison of sequences from different species was that identical regions, which had

[13] Coleman, 1971; Allen, 1978, and for contemporary narrative, see for example Bowler and Morus, 2005. A number of authors have taken a more nuanced view, however. Lynn K. Nyhart for example, claimed that natural history was declining relatively and growing absolutely around 1900, due to the general expansion of biology's territory, Nyhart, 1996, p. 422. For Keith Benson, natural history remained "alive and well, primarily within museums," Benson, 1988, p. 77. Scholarship on the history of natural history has focused nearly exclusively on the period from the seventeenth to the nineteenth century, Jardine et al., 1996. When the twentieth century is considered at all, natural history practices are studied in the context of ecology, some areas of evolutionary studies, and obviously systematics, but always far from the laboratory, with the exception of Kohler, 2002. Paul Farber, taking a nuanced approach to the opposition between natural history and experimentation, noted pointedly that, from an intellectual point of view, the experimental approach (physiology) and natural history "did not have be competitors," Farber, 2000, p. 80. For a broader discussion, see Strasser, 2010.

[14] Strasser, 2010.

[15] Pickstone, 1993, 2007.

been preserved through evolution, might indicate the presence of an essential part of the molecule, such as the "active center." Variable regions, on the other hand, might indicate parts of the molecule which had not been under the pressure of natural selection, and that were thus probably of lesser functional importance. The following four examples illustrate how widespread this mode of reasoning was among biochemists in the 1950s.

By 1953, the biochemist Frederick Sanger, working at the University of Cambridge, had determined the first complete sequence of a protein – insulin – taken from an ox.[16] But from the start of his research in the late 1940s, he also examined insulin from other species, eventually sequencing pig, sheep, horse, and whale insulin.[17] After presenting an alignment of these five insulin sequences in 1956, Sanger and his coworkers noted that the differences were confined to a small portion of the molecule, the disulfide bridge. This result was puzzling because they believed this region to be important for the physiological role of the protein, perhaps even its "active center."[18] Yet they did not question the rationale behind sequence comparison; to the contrary, they called for more studies of species differences.[19]

In Vienna, the protein chemist Hans Tuppy, a student of Sanger, was pursuing similar goals by sequencing parts of the cytochrome *c* protein in horse, ox, pig, salmon, and chicken. In 1954, he was surprised to find, in view of the many physical differences of the proteins, that the sequences close to the active site of the first three proteins were identical.[20] He explained this paradoxical finding by noting that the active site of the molecule was likely to be the most conserved. A year later, however, he detected a single amino acid difference in chicken.[21] Tuppy, like Sanger, also took advantage of the first known sequence to infer the others from data on amino acid composition alone. Like Sanger, he hoped these studies would help determine how cytochromes carried out their function. "Those features which turn up invariably in all various cytochromes *c*," he argued, "are likely to be essential to the specific catalytic function, whereas structural differences will indicate points not directly concerned with catalytic activity."[22] Unlike Sanger, however,

[16] Garcia-Sancho, 2010.

[17] Sanger, 1949; Brown et al., 1955; Harris et al., 1956; de Chadarevian, 1999.

[18] Harris et al., 1956; Brown et al., 1955, p. 565.

[19] Harris et al., 1956, p. 437.

[20] Tuppy and Bodo, 1954.

[21] Tuppy and Paléus, 1955.

[22] Paléus and Tuppy, 1959, p. 2.

Tuppy did not limit himself to comparing the proteins he had sequenced, i.e. cytochrome *c*, but also took into consideration other sets of homologous sequences, including insulin, hemoglobin, and trypsin. In October 1958, for example, he gave a public lecture where he showed alignments of these different proteins sequences from various domestic organisms and reflected on how they might help determine the "active center" of the respective molecules.[23]

At the Karolinska Institute in Stockholm, the chemists Margareta and Birger Blombäck extended this approach to a much broader range of species. In the early 1950s, they embarked on a lifelong study of the clotting factor fibrinopeptide. After learning the new degradation technique developed in nearby Lund by Pehr Edman, which made protein sequencing much easier than Sanger's method, they applied it to the study of fibrinopeptides from various mammalian species. In addition to the usual domestic species studied by Sanger and Tuppy – cat, dog, ox, horse, donkey, pig, rabbit, goat and sheep – they investigated wild species – badger, bison, fox, green and rhesus monkey, llama, mink, red deer, and reindeer. In 1965, after having compared the sequence of fibrinopeptide from 22 species, they observed that certain positions in the sequence had "been stationary during mammalian evolution." These amino acids were thus likely to be "of importance for directing thrombin action,"[24] they argued.

Finally, in the United States, the biochemist Christian B. Anfinsen was pursuing a similar project using ribonuclease, and also argued that "variations from species to species may yield valuable information on the location of the site of enzymatic activity."[25] In his 1959 book, *The Molecular Basis of Evolution*, Anfinsen, drawing on the work on Sanger, Tuppy, the Blombäcks and others, presented numerous sequence alignments from insulin, ribonuclease, cytochrome *c*, adrenocorticotropin, melanotropin, vasopressin, oxytocitin and hypertensin. His main interest, like that of the other biochemists, was in the similarities which would indicate "the minimum structure which is essential for biological function."[26]

These biochemists used the diversity of nature to gain insights into the relationship between the structure and the function of proteins. They

[23] Tuppy, 1958. The conference was organized by the Gesellschaft Deutscher Naturforscher und Ärtze, The results of yeast cytochrome-*c* are presented in Tuppy and Dus, 1958.

[24] Blombäck et al., 1965, p. 1789.

[25] Anfinsen et al., 1959, p. 1118.

[26] Anfinsen, 1959, p. 143.

focused on structural similarities, assuming that natural selection had eliminated structural variations in functionally important regions of proteins. Localizing the active site of proteins was their main concern, and evolutionary considerations were a means to that end. Other biochemists, however, turned the argument on its head, trying to draw conclusions about evolution from sequence variations. As early as 1956, Sanger had suggested that "more extensive studies of species differences in amino acid sequences of polypeptide chains may lead to interesting conclusions concerning evolutionary trends in protein biosynthesis."[27] Two year later, Tuppy was much more explicit: "The more proteins differ, due the exchange of amino acids in different places of the polypeptide chain, the further away in evolution the organisms from which they originate are. The comparative search for amino acid sequence in proteins could become an aid to discover evolutionary relationships."[28] This is perhaps one of the first published statements of the idea that the quantitative comparison of amino acid sequence changes might yield information about evolutionary distances. In 1959, in his *Molecular Basis of Evolution*, Anfinsen similarly suggested that the "rate at which successful mutations [had] occurred throughout evolutionary time" may serve as "an additional basis for establishing phylogenetic relationships,"[29] yet he did not propose phylogenies himself. The comparison of protein sequences among various species was thus commonly presented as a key to evolutionary problems in the 1950s, even if protein sequences were not singled out as they would be by molecular evolutionists a decade later.[30]

The fact that these biochemists often focused their research on a single protein (or a family of similar molecules), but were keen to examine it in several species, should not come as a surprise. Indeed, a well-established tradition of comparative biochemistry (and comparative physiology) sought to shed additional light on the function and the generality of biochemical systems by comparing them among various organisms. The biochemist Ernest Baldwin, for example, one of Frederick Sanger's mentors at Cambridge,[31] wrote a popular *Introduction to*

[27] Harris et al., 1956, p. 137.

[28] Tuppy, 1959, p. 42. Originally: "sollten sich Proteine voneinander um so stärker, durch einen Austausch von Aminosäure-Resten an umso mehr verschiedenen Stellen der Polypeptidketten unterscheiden, je weiter die sie produzierenden Organismen in der Evolution voneinander entfernt sind. Die vergleichende Untersuchung der Aminosäure-Sequenzen in Proteinen könnte folglich als ein Hilfsmittel zur Aufdeckung entwicklungsgeschichtlicher Zusammenhange dienen."

[29] Anfinsen, 1959, p. 143.

[30] *Ibid.*, Chaps. 7 and 11.

[31] Sanger, 1988, p. 3.

Comparative Biochemistry that was first published in 1937 and went into new editions through the late 1960s.[32] In line with Frederick Gowland Hopkins's programmatic vision, Baldwin's main interest was to produce generalizations about the biochemical basis of life.[33] The study of various species was a way to reach that goal, and for Baldwin "a starfish, or an earthworm, neither of which has any clinical or economic importance *per se*, is as important as any other living organism and fully entitled to the same consideration."[34] The Belgian biochemist Marcel Florkin also published an influential little book in 1944, *L'Evolution Biochimique*, translated 5 years later into English.[35] Florkin, too, reviewed the biochemistry of numerous organisms in order to stress "the unity of the biochemical plan of animal organization."[36] Unlike Baldwin, however, he suggested that biochemical characters might also serve to establish phylogenies as soon as more facts about the biochemistry of different species became known.[37]

In retrospect, one might be surprised that in the late 1950s the accumulation of sequence data from homologous protein did not lead to a more direct attempt to use them to reconstruct phylogenies. Two explanations can be offered as to why this was not the case. First, the amount of sequence data remained limited, and was often restricted to the active site of a molecule. The active site was of most interest to biochemists investigating protein function, but because it was also the portion of the molecule that was the most constant, it was the least useful for evolutionary studies. It was only when automatic amino acid analyzers became more broadly available after 1958 that larger numbers of complete protein sequences, and from somewhat more exotic organisms, came to be determined.[38] Second, the relationship between protein sequences and mutations at the DNA level was not well

[32] Baldwin, 1937, 1966.

[33] Baldwin was a student of Hopkins. On Hopkins, see Kohler, 1982.

[34] Baldwin, 1937, p. xiv.

[35] Florkin, 1944, 1949.

[36] Florkin, 1944, p. 11, translation is mine.

[37] *Ibid.*, pp. 194–196, translation is mine. The biochemist Erwin Chargaff's studies on the regularities of nucleic acid composition were also derived from the examination of material from several species, including man, ox, yeast and bacteria. Chargaff, 1955. In the United States, the comparative biochemistry tradition was also promoted by microbiologists, such as Cornelius B. van Niel, a student of Albert Jan Kluyver, from Delft, who had coined the expression "comparative biochemistry." See Spath, 1999.

[38] Following the work of William H. Stein and Stanford Moore, the instrument maker Beckman brought the automatic amino acid analyzer on the market, a Spinco Model 120; Moore et al., 1958.

understood in that period. Until around 1960, it was unclear whether DNA sequences determined protein sequences entirely or if other components of the cell intervened.[39] In 1959, for example, Christian B. Anfinsen noted: ''Many readers will not be willing to swallow, whole, the thesis that proteins represent the *direct* translation of genetic information.''[40] It was only when this question was considered unambiguously resolved in the early 1960s that protein sequences could be confidently considered to reflect directly mutations that had occurred during evolutionary history, and be safely regarded as ''documents of evolutionary history,'' as Zuckerkandl and Pauling had put it in 1965.[41]

Biochemists often built their entire careers around a single protein, for example ribonuclease in Anfinsen's case. Thus, when they collected sequences from many species, they usually focused on just one protein, or a small family of related proteins. The practice of collecting all known sequences from many different proteins and organisms grew out of a very different set of concerns: the deciphering of the genetic code.

Comparing Sequences to Crack the Genetic Code

Between 1954 and 1966, finding a solution to the problem of the genetic code was considered one of the important challenges in experimental biology. In 1954, the big-bang theorist George Gamow suggested that the genetic code could be solved as a cryptogram, and made a proposal for an overlapping code that was soon shown to be flawed.[42] He then invited a number of molecular biologists and physicists, including Francis H. C. Crick, Martynas Yčas, and Sydney Brenner, to join the RNA Tie Club, which he founded to organize the efforts to decipher the code theoretically. Lily Kay has described in great detail how these theoretical approaches borrowed, sometimes liberally, concepts from cybernetics, cryptography, and information theory.[43] Yet these attempts were not just theoretical speculations; they were also constrained by empirical data, in particular by collections of protein sequences.

The coding problem, as it was frequently stated in the 1950s, consisted of how to relate a text written with four letters (made of nucleotides) to a text written with 20 letters (made of amino acids). Had a

[39] Strasser, 2006b.

[40] Anfinsen, 1959, p. 143, emphasis in original.

[41] Zuckerkandl and Pauling, 1965.

[42] Gamow, 1954; Gamow and Metropolis, 1954.

[43] Kay, 2000.

DNA or RNA sequence and a corresponding protein sequence been known (a "Rosetta Stone"), the problem would have been relatively trivial to solve. But in the 1950s, only proteins had been sequenced. Nucleic acid sequences remained almost impossible to determine until the mid-1960s for RNA, and the mid-1970s for DNA.[44] Thus those who wished to decipher the genetic code were stuck with examining whatever protein sequences were available, a situation analogous to that in cryptanalysis when none of the content of a coded message was known. Members of the RNA Tie Club applied a typical strategy used in cryptanalysis to this case, namely the search for correlations between adjacent letters in the encrypted message. In human languages, some letters are more frequently followed by others, such as "q" and "u" for example, and similar associations in protein sequences could give clues about the underlying nucleic acid codons. For these studies, every single protein sequence, as short as two amino acids, could be used. These researchers were the first to adopt the strategy of systematically collecting sequences from different proteins and different organisms.

One key question that such a strategy might answer was whether or not the code was overlapping. If it was, then certain amino acids would preferably have certain neighbors with which they shared the overlapping part of their codons.[45] George Gamow, Alexander Rich, and Martynas Yčas, all members of the RNA Tie Club, published one such example in 1956 in one of the first extensive reviews of the "The Problem of Information Transfer from Nucleic Acids to Proteins."[46] The empirical evidence on which they relied to evaluate their different hypothetical codes consisted of four full pages listing all the proteins sequences known to that date. A year later, the biologist Sydney Brenner, another member of the Club, inferred from all the published sequences that in view of the random distribution of the amino acids, any overlapping code could be ruled out.[47]

The practice of collecting homologous sequences of different proteins and comparing them grew out of similar concerns to solve the code. As early as 1956, in their review of the coding problem, Gamow, Rich,

[44] On the history of protein sequencing, see de Chadarevian, 1996, 1999; Garcia-Sancho, 2010.

[45] If in the DNA sequence "abcd" includes two successive overlapping codons "abc" and "bcd," coding for amino acid X and Y, then X will frequently be followed by Y in protein sequences (a frequency greater than 1/20 for overlapping codes and of 1/20 in non overlapping codes).

[46] Gamow et al., 1956. On the genesis of the review, see Kay, 2000, p. 148.

[47] Brenner, 1957. In fact Brenner only showed that codes overlapping by 2 nucleotides out of 3 were impossible.

and Yčas listed all known protein sequences and then presented alignments of six different sets of homologous proteins in order to test their hypothetical code.[48] The same year, in another paper on the code, Yčas presented twelve sets of aligned homologous proteins.[49] Alignments of protein sequences from different natural strains and mutants of a single organism, the tobacco mosaic virus (TMV), or closely related viruses, came to play a particularly important role in the cracking of the code after August 1961. Heinz Fraenkel-Conrat in Wendell M. Stanley's laboratory at the University of Berkeley and Heinz G. Wittmann in Georg Melchers's Max Planck Institut für Biologie in Tübingen pursued this approach most directly.[50] After the possibility of inducing mutations in TMV using nitrous acid was demonstrated in 1958, they produced numerous TMV mutants, sequenced their polypeptides, and compared their amino acid composition and sequences, an approach which had been facilitated by the development of the automatic amino acid analyzer that same year. Both groups believed that this approach could be key to solving the genetic code. Heinz G. Wittmann, for example, reported in 1960 that most, but not all, of 26 natural and chemically induced mutants had altered amino acid sequences, and noted that this information would become essential "to solve the coding problem."[51] Similarly, Tsugita and Frankel-Conrat, after presenting a sequence alignment from a TMV protein and a chemically induced mutant, noted three amino acid differences and concluded that this information would be of "considerable interest in connection with the mechanism of coding the genetic properties."[52]

These assessments proved correct. The collection of amino acid changes became crucial after August 1961, when Marshall W. Nirenberg and J. Heinrich Matthaei announced the discovery of some codons as the result of experiments with synthetic polynucleotides.[53] Indeed, assuming that a mutation from one amino acid to another involved a single nucleotide change, once a few codons were known, a collection of amino acid changes would drastically simplify the determination of the

[48] Gamow et al., 1956.

[49] Yčas, 1958, 1961.

[50] Creager, 2002, pp. 303–311; Kay, 2000, pp. 179–192; Brandt, 2004, Chap. 6.

[51] Wittmann, 1960, p. 610.

[52] Tsugita and Fraenkel-Conrat, 1960, p. 641.

[53] Lily Kay points out that this information was used to confirm the code, but she does not raise the point that it was *produced* with the coding problem in mind (Kay, 2000, pp. 187–189); however, Angela Creager makes this point (Creager, 2002, pp. 303–311).

remaining codons.[54] The biochemist Severo Ochoa relied extensively on this reasoning, and on the data about TMV mutants to confirm his codon assignments and infer new ones.[55]

Although the importance of TMV mutants for the resolution of the genetic code has been recognized by historians, the broader relevance of sequence comparisons and alignments has been overlooked. At the same time that Severo Ochoa and collaborators were using TMV substitutions, the biochemist Emil L. Smith was relying on the large body of sequence data of cytochromes *c*, insulin, hemoglobin, and other proteins that had been taken from organisms as different as pigs and bacteria. In 1962, Smith used sequence alignments to gather information about amino acid replacement and confirmed the genetic code assignments made by Ochoa and others.[56] He also speculated on the evolution of protein function and hoped that this approach might provide "a new tool for the study of species relationships."[57] The same year, the biologist Thomas H. Jukes used 48 known amino acid changes from an equally wide range of species to suggest new codon assignments.[58]

The biologist Richard V. Eck (1922–2006), who would become a co-author of the *Atlas of Protein Sequence and Structure*, also began to collect sequences when he worked on the genetic code. After studying chemical engineering, and then plant biology at the University of Maryland, Eck joined the National Cancer Institute in 1954. There he developed mathematical models to evaluate complications from cancer surgery, until, in 1960, he turned to the theoretical study of the genetic code. In 1961, Eck published a paper in *Nature* in which he compared all the sequences of hemoglobin variants, such as sickle cell hemoglobin, and all the sequences of homologous proteins, such as insulin, from different species. He suggested that "the published data on amino acid sequences can be sorted, tabulated and arranged in a great variety of ways [and] any such manipulation will produce some sort of pattern."[59] Indeed, he noted numerous amino acids substitutions between various sets of homologous proteins, some of which occurred more often than others. Contradicting Brenner, he concluded that these frequencies

[54] For example, if UUU coded for the amino acid phenylalanine, as Nirenberg and Matthaei had established, and phenylalanine was replaced by another amino acid in a mutant, one could deduce that this amino acid was coded by one of only nine different codons (all including two Us), and thus excluding 44 other possible combination.

[55] Lengyel et al., 1961, 1962; Speyer et al., 1962a, b; Basilio et al., 1962.

[56] Smith, 1962a, b.

[57] Smith, 1962b, p. 863.

[58] Jukes, 1962a, b, c.

[59] Eck, 1961, p. 1285.

reflected the fact that the code was at least partially overlapping. As Eck pointed out, this hypothesis of an overlapping code had "several attractive features – one of which is the theoretical possibility of solving it."[60] Soon after, he prepared a more extensive treatment of his analysis for the *Journal of Theoretical Biology*. After "compiling the published sequences," he presented 61 protein sequences aligned with their homologous sequences, the largest published collection of sequences to date. He then proposed a complete solution to this "protein Cryptogram."[61] Eck's papers were composed before, but appeared in print just after, Nirenberg and Matthaei's August 1961 announcement that they had solved the first codon of the genetic code experimentally – thus providing much more compelling evidence for their solution than the theoretical approaches pursued by Eck and others could do for theirs.

The three complete solutions to the genetic code which had been proposed by 1962, by Smith, Jukes, and Eck, as well as the later codon assignments derived by the biochemist Walter M. Fitch, for example,[62] relied extensively on the comparison of many homologous sequences from a variety of organisms, including humans, pigs, sheep, oxen, horses, sperm whales, finback whales, humpback whales, seals, salmon, chickens, turkeys, silkworms, frogs, rabbits, bacteria, and viruses. These results had been obtained in the context of studies on the relationships between the structure and function of proteins and afterwards assembled to solve questions related to the genetic code. In the following years, they became an essential part of the nascent field of molecular evolution as theorized by Linus Pauling, Emil Zuckerkandl, and many others. Interestingly, Smith, Jukes, Eck, and Fitch, after their work on the code, all became involved in the study of molecular evolution. The comparative perspective on protein sequences which they had adopted to solve the genetic code transferred easily to the determination of phylogenies in the context of molecular evolution.

In 1965, when Dayhoff and Eck published their *Atlas of Protein Sequence and Structure*, the practice of collecting and comparing sequences was thus already well established, and in fields other than molecular evolution. However, the *Atlas* differed in one crucial way from previous collections of protein sequences. It was the first presentation of homologous sequences that was not tied to a specific research question. The *Atlas* was an open-ended tool. What made it particularly powerful for addressing numerous scientific problems was the fact that

[60] *Ibid.*, p. 1285.

[61] Eck, 1962b.

[62] Fitch, 1964, 1966b.

it was created as a computerized collection of data, probably the earliest in the life sciences.[63]

Computing Sequences: Dayhoff, Ledley and the Computer Revolution

Margaret O. Dayhoff (1925–1983) played the leading role in turning sequence data scattered through the printed literature into a computerized collection and bringing computers to bear on problems of sequence analysis and molecular evolution. She obtained a PhD in quantum chemistry in 1948 under George E. Kimball at Columbia University, after obtaining a BA in mathematics and an MA in chemistry.[64] As a fellow at the Watson IBM Computing Laboratory in 1947–1948, she used punch card machines to calculate resonance energies in small molecules.[65] After obtaining her degree, she worked at the Rockefeller Institute (now Rockefeller University) as a research assistant on problems of theoretical chemistry and then at the University of Maryland. She joined the National Biomedical Research Foundation (NBRF) in 1960,[66] and eventually became professor of physiology and biophysics at Georgetown University and president of the Biophysical Society (1980–1981).

The NBRF was a unique environment in which computers and biology were brought into close proximity. This private non-profit institution had been founded in 1960 just outside of Washington D.C. by Robert S. Ledley to explore the possible uses of electronic computers in biomedical research.[67] It was created as a place where computing and "biology or medicine could be combined intimately."[68] Born in 1926, Ledley was trained as a dentist before obtaining an MA in theoretical physics from Columbia University and becoming interested in digital computers.[69]

[63] The introduction of computers in X-ray crystallography and systematics has been explored in de Chadarevian, 2002, Chap. 4 and Hagen, 2001, respectively. The only broad account on the topic is November, 2006. Hagen provided the first historical perspective of the role of Margaret O. Dayhoff in the birth of bioinformatics, Hagen, 2000.

[64] Margaret O. Dayhoff, "Biographical sketch Margaret Oakley Dayhoff," 1965, National Biomedical Research Foundation Archives, currently processed at the National Library of Medicine, Bethesda (NBRF Archives hereafter).

[65] Oakley and Kimball, 1949.

[66] Robert S. Ledley, "Memorandum," November 16, 1960, NBRF Archives.

[67] Robert S. Ledley to Harvey E. Saveley, June 29, 1960, NBRF Archives. The NBRF eventually moved to Georgetown University Medical Centre, Washington, DC.

[68] Margaret O. Dayhoff to Naomi Mendelsohn, June 28, 1966, NBRF Archives.

[69] On the early career of Ledley, see November, 2006, pp. 59–76.

From 1952, he worked at the National Bureau of Standards programming the SEAC, one of the first stored-program electronic computers in the United States. In 1965, he published a 900-page monograph entitled *Use of Computers in Biology and Medicine*.[70] It constituted an introduction to the principles and methods of digital computing and an exploration of their possible application in a number of fields of biology and medicine. The publication of this book was only one example of Ledley's lifelong commitment to promote the use of digital computers in biomedicine, from the automated recognition of chromosome images to computer-assisted medical diagnostics, and in the analysis of molecular sequences.

Of particular significance in understanding how computers came to be applied to sequence analysis by Dayhoff is the fact that Ledley was invited by George Gamow in 1954 to become one of the twenty members of the RNA Tie Club.[71] Gamow believed that Ledley's expertise in digital computers and symbolic logic would be useful in solving the genetic code. Ledley's first, and only, contribution resulting from this participation in the Club was to outline a very general "system of digitalized computational methods" to be applied to practical problems in "science, industry, and government." He gave as an example the evaluation of overlapping codes by analyzing amino acid sequences.[72] Ledley noted that it "should take a computer no more than a hundred hours" to work out a solution, whereas if all possible solutions had to be tested, "a computer put to work in the days of the Roman Empire, at a rate of one million solutions per second, 24 h a day, all year round, would not yet be close to finishing the job."[73]

After his initial contribution, rather unsuccessful in view of the absence of tangible results and the complete neglect of the method by other researchers,[74] Ledley envisioned another application of computers to sequence analysis. This time, he suggested that computers could assist biochemists in their efforts to determine protein sequences. A standard experimental method consisted in cutting the polypeptide chain into several overlapping fragments and establishing the sequences of each.

[70] Ledley, 1965. On the genesis of this volume, see November, 2006, Chap. 2.

[71] Georges Gamow to James Watson, December 6, 1954, reproduced in Watson, 2001, Annex 12.

[72] Ledley, 1955. The paper was communicated by George Gamow.

[73] *Ibid.*, p. 511. Similarly, at Los Alamos, George Gamow was using the MANIAC computer to make Monte Carlo simulations to produce a randomly ordered protein sequence and compare them with the available empirical data in his study of the genetic code. Gamow and Yčas, 1955. On this episode, see Kay, 2000, p. 141.

[74] Ledley's paper was almost never cited, except by Ledley himself. ISI Web of Science.

The problem was then to reassemble these partial sequences into the complete sequence of the original protein. In the 1960 draft of his book,[75] which was only published in 1965, Ledley outlined a method to solve this problem using a computer.[76] He invited Dayhoff to join the NBRF in 1960 to continue investigating this question under an NIH grant.[77] In their reports published between 1962 and 1964, they described a set of FORTRAN programs they had devised for the IBM 7090, a mainframe computer located at Georgetown University, that could assemble partial sequences in the right order in less than 5 min.[78] One of the programs searched the peptide sequences for particular characteristics, while another compared all peptide sequences in search of overlaps.[79] These two practices – searching and comparing – would later become essential to computing sequences in molecular evolution and other fields. Simultaneously, a very similar approach to sequences analysis was being pursued by Richard V. Eck at the nearby National Cancer Institute in Bethesda, where he tested his algorithm in a "paper experiment" designed from published sequences.[80]

Ledley and Dayhoff made clear that their computer programs would not downgrade the protein chemist to a simple technician, but that the computer would merely serve as an aid: "These routines may be thought of as analogous to the staff of a laboratory. Each routine has a function to perform just as a laboratory has people each with a job to perform, cleaning people, technicians, senior research workers, a librarian, a machinist, etc. The programmer and protein chemist have been upgraded to the chief of the computer staff."[81] In pressing this analogy, where the protein chemists and the programmer were in charge, Ledley perhaps wanted to avoid the outraged reactions he had just faced from physicians in response to his suggestion of using computers to make medical diagnoses.[82] He thus made clear that computers would not replace humans, but only assist them. Indeed, the computer programs he designed would print out intermediate results "for examination by

[75] As cited in Dayhoff and Ledley, 1962.

[76] Ledley, 1965, p. 373.

[77] Robert S. Ledley, "Memorandum," November 16, 1960, NBRF Archives.

[78] "Summary Progress Report of GM-08710," January 15, 1963, NBRF Archives.

[79] *Ibid.*

[80] Eck, 1962a.

[81] "Summary Progress Report of Grant Sequences of Amino Acids in Proteins by Computer Aids," January 15, 1963, NBRF Archives.

[82] Ledley and Lusted, 1959; Ledley, 1959a and the reactions in Ledley, 1960.

the biochemist'' and the process thus reflected ''a close cooperative effort between the computer and the biochemist.''[83]

Ledley, Dayhoff, and Eck hoped that these computer methods would be used by the increasing number of biochemists sequencing proteins. Yet these computational methods seem to have had no visible impact on sequencing practices.[84] Many biochemists did not have access to computers in the early 1960s, and when they did, they often lacked the programming skills to use them.[85] More importantly perhaps, even when they could have secured the help of a programmer, they seem to have been resistant to the use of computers, which they perceived as particularly foreign to the culture of the ''wet lab.'' In 1966, Dayhoff warned a student in search of a job in a laboratory where she could use her expertise in programming and in biochemistry to make sure ''that the biochemists are sympathetic to the computer.''[86]

Dayhoff and Eck's early attempts to use computers for sequence analysis led them to compile published data on amino acid sequences, a compilation which eventually became the *Atlas of Protein Structure and Sequence*. It also brought them to think about the best ways of handling sequences with a computer. For example, they adopted a one-letter notation for amino acids, instead of the usual three letter code, in order to save computer memory and to make alignments more readable on fixed-space printers. Most earlier sequence comparisons, using the three letter code, failed to present the data in an easily comparable way due to the different typographic length of the three-letter amino acid notation (compare ''Ile'' to ''Asn,'' for example).[87]

Other research projects carried out at the National Biomedical Research Foundation also played a role in the computerization of sequence analysis. For example, Ledley and Dayhoff devised computer programs to draw contour maps and density maps from X-ray diffraction data.[88]

[83] Dayhoff and Ledley, 1962, p. 267. In the same paper, Dayhoff and Ledley suggest using the same approach for DNA and RNA sequencing once the experimental data becomes available, p. 274. See also Bernhard et al., 1963 for another computer approach to the same problem.

[84] The articles written by Eck, Dayhoff and Ledley were hardly ever cited, except by themselves in the 1960s and 1970s. See however the discussions between Margaret O. Dayhoff and Marvin Shapiro, NBRF Archives, December 1962, and Shapiro et al., 1965.

[85] By the time, most campuses in the United States had central computing facilities, but there is no evidence that biochemists used them, Anonymous, 1962.

[86] Margaret O. Dayhoff to Naomi Mendelsohn, June 28, 1966, NBRF Archives.

[87] See Table 1 in Hunt, 1958.

[88] NIH GM 8710 Reports, 1962–1965, NBRF Archives.

In this field, unlike that of sequence analysis, Ledley and Dayhoff were building on a long tradition in crystallography of using computers to assist in the determination of protein structure.[89] It led them to investigate further the question of the relationship between a protein sequence and the structure of its active site, another field of protein science which would become important in the *Atlas*.[90] This attempt illustrated once again the belief that computers could produce meaningful results through the analysis of empirical data, the key premise on which the production of the *Atlas* rested.

The role of the computer in the creation of the *Atlas* was not only to analyze the sequences, but also to store, tabulate, and print them. All sequences, and their related information, were entered on punch cards. Each card constituted an entry, and the collection of all the cards was regarded as an "Amino Acid Sequence Library,"[91] which could be subjected to the increasingly sophisticated computing techniques that were being developed in the field of library science.[92] Ledley took part in these developments as well, which provided another important resource for the creation of the *Atlas*. In 1958, for example, he developed a new system for coordinating the indexing of book-format bibliographies, which he called TABLEDEX. His method, another application of symbolic logic, allowed the user to search for entries containing several keywords, instead of a single one as in most indexes. The National Science Foundation, which was actively promoting computing in American universities,[93] supported Ledley's attempt to utilize "a digital computer to assist in the automatic preparation of a bound book form bibliographical index."[94] Similarly, in 1961, Ledley proposed to the National Library of Medicine a method for using digital computers for the publication of the Index Medicus,[95] which would include programs to search the Index. The primary reasons for Ledley and others' concerns with the organization of scientific information in the 1950s was the perception that the amount of published information was "exploding." In 1957, Ledley claimed that the "rate of doing research" had doubled since 1950 and that it was

[89] de Chadarevian, 2002, Chap. 4.

[90] Dayhoff, 1964.

[91] Richard V. Eck, "Appendix to Progress Report," July 1965, NBRF Archives.

[92] Miles, 1982, Chap. 13.

[93] Aspray and Williams, 1994.

[94] Robert S. Ledley to James B Wilson, "Report on 'A Tabledex Computer Program'," March 1, 1961, NBRF Archives.

[95] Robert S. Ledley, "Final Report on SAph 71251," January 1961, NBRF Archives.

continuing its exponential growth.[96] This observation would become central to the argument of Dereck J. de Solla Price's 1963 book *Little Science, Big Science*.[97] The same perception of an "explosion of information"[98] in the field of protein sequencing would prompt the use of computers to organize and make sense of information in this field.

Ledley had long believed that computers were ideal tools not just for calculation, but for "data processing" and the analysis of "large amounts of detailed experimental results." In 1957, when surveying the possible uses of computers in biology and medicine, he gave equal weight to calculation ("numerical solutions to partial differential equations" and "simulation of biological systems") and to data processing ("bio-medical processing and reduction" and "bio-medical information retrieval").[99] In this, his vision of the field and its future was atypical. Other surveys on the uses of computers in biology and medicine mainly emphasized calculation (equation solving and numerical simulations).[100] Ledley, in contrast, highlighted the promise of computers for data processing. He went so far as to outline a vision in which scientific data would be published electronically. Instead of "publishing articles in journals, research results might be transmitted to a central information centre," he suggested. The creation of the *Atlas* represented a first step towards accomplishing this vision. Given Ledley, Dayhoff, and Eck's backgrounds in using computers to analyze sequences and to organize information, it is not surprising that the *Atlas* was created as a computerized system.

The use of computers also brought Eck and Dayhoff independently to consider questions of evolution. In 1964, for example, at a conference on Engineering in Biology and Medicine, Eck presented a "cryptogrammic" method to trace the "evolution of proteins."[101] As he had done with his earlier speculations about the genetic code, he now suggested that "the publication of the amino acid sequences of many proteins" made it possible to "treat the whole of evolution … as a cryptogram."[102] He used

[96] Robert S. Ledley "Functional criteria for biomedical digital electronic computer design," March 1957, NBRF Archives.

[97] de Solla Price, 1963.

[98] Margaret O. Dayhoff to Kendrew, January 28, 1966, NBRF Archives.

[99] Robert S. Ledley "Functional criteria for biomedical digital electronic computer design," March 1957, NBRF Archives. This manuscript formed the basis of the influential piece published 2 years later in Science, Ledley, 1959b.

[100] See for example, Stacy and Waxman, 1965; Sterling and Pollack, 1965; Medical Research Council, 1965.

[101] Eck, 1964.

[102] *Ibid.*

data from several hundred amino acid substitutions in homologous proteins to calculate with a digital computer the probability that one amino acid was replaced by another. Using this result, he suggested that one could calculate the "degree of relatedness of each protein" with reference to its ancestors, and from there draw "a family tree of proteins … to *scale*," the distances between the branches of the tree representing a "numerical measure of relatedness."[103] Even though he did not actually present a phylogenetic tree, he outlined the possibility of constructing "a detailed phylogenetic tree of the vertebrates,"[104] provided that enough protein sequence data became available.

Like Eck, who became involved in evolutionary studies through cryptanalysis, Dayhoff entered evolutionary biology through one research agenda of the Cold War, the search for life in space.[105] In the post-Sputnik era, investigations into the physicochemical conditions that could have led to the creation of organic compounds on earth, and eventually to life, were actively supported by NASA. This pursuit easily captured the public's imagination and helped to legitimize NASA's use of taxpayer money. Dayhoff climbed on the bandwagon in collaboration with the chemist Ellis R. Lippincott and the astronomer and science popularizer Carl Sagan.[106] Together with Eck, they used the IBM 7090 available at Georgetown University to simulate the evolution of the prebiological atmosphere and examine under which conditions "biologically interesting compounds, such as amino acids, were generated."[107] This work followed up on Stanley L. Miller and Harold C. Urey's discovery of 1959 that amino acids could form spontaneously from chemical compounds believed to have been present on earth before the appearance of life.[108]

Dayhoff's interests in chemical evolution and in amino acid sequences converged and became mutually reinforcing.[109] The early chemical conditions on earth suggested that certain proteins such as ferredoxin might have played an important role in the origins of life, and indicated that certain amino acids, because comparatively stable under these conditions, might have been present in the ancestral

[103] *Ibid.*

[104] *Ibid.*

[105] On the history of exobiology, Wolfe, 2002 and Strick, 2004.

[106] Dayhoff et al., 1967.

[107] Dayhoff et al., 1964.

[108] Miller and Urey, 1959.

[109] "Final Report to the Office of the Life Sciences Programs, October 1, 1965 to September 1, 1965," NBRF Archives.

sequences of protein. Eck and Dayhoff, using a computer program that "matched the sequence of ferredoxin against itself in all combinations," found that the sequence had evolved through the duplication of a very short primitive protein.[110] This compelling demonstration of how computers could reveal evolutionary information was published in *Science* in 1966. In a letter to her NASA sponsor, Dayhoff pointed that "the biochemists who published the sequence missed the evolutionary implications entirely."[111]

The *Atlas of Protein Sequence and Structure*

The publication of the *Atlas of Protein Sequence and Structure* in 1965 resulted from the growing interest in collecting, comparing, and computing sequences outlined above. It was meant as a tool to produce knowledge about the structure, function, and evolution of proteins. As Dayhoff later put it in a letter to a colleague, "there is a tremendous amount of information regarding evolutionary history and biochemical function implicit in each sequence and the number of known sequences is growing explosively. We feel it is important to collect this significant information, correlate it into a unified whole and interpret it."[112]

The first edition, authored (or edited) by Dayhoff, Eck, and two collaborators at the NBRF was just under 100 pages and contained the sequences of around 70 proteins, mainly cytochromes *c*, hemoglobins, and fibrinopeptides from various species. Each page gave the name of a protein and an organism ("Hemoglobin beta – gorilla," for example), followed by the protein's amino acid sequence symbolized both in the three letter abbreviation and in a custom one letter abbreviation system. Each page also listed the amino acid composition, any remarks on how the data had been obtained, and a reference to the source of the data, usually a bibliographic reference. The *Atlas* also included alignments of sequences of a same protein, such as haemoglobin, taken from several organisms. This presentation allowed the users to grasp in a single look where conserved regions resided along the protein, giving an essential clue to the presence of an functionally essential part of the molecule.

[110] Eck and Dayhoff, 1966a, p. 365.

[111] Margaret O. Dayhoff to George Jacobs, January 12, 1966, NBRF Archives. This was perhaps true, but Dayhoff was not the only one to find internal duplication in proteins, as at least two other groups published the same conclusion in 1966 (Doolittle et al., 1966; Fitch, 1966a).

[112] Margaret O. Dayhoff to Carl Berkley, February 27, 1967, NBRF Archives.

The authors paid tribute to those who had determined the sequences included in the *Atlas* by dedicating it "to all the investigators who have developed the techniques necessary for the grand accomplishments represented by this tabulation, and to all those who have spent so much tedious effort in their application."[113] They firmly positioned their work in the experimental tradition by claiming that the *Atlas* "voluminously illustrates the triumph of experimental technique over the secretiveness of nature," and cited as the goal of their collection to make apparent the information that was "hidden in the amino acid sequence."[114] That information was important to studies of the conformation of proteins, the sequence of the underlying genes, and "the record of the many thousand mutational steps by which we can quantify a phylogenetic tree." The editors asserted that basing phylogenies on sequences would be far superior to traditional taxonomic criteria that were deemed to be "extremely vague and uncertain,"[115] thus siding – unsurprisingly – with the new molecular evolutionists and taxonomists against their organismic counterparts.[116]

The authors of the *Atlas* also invited their readers to cooperate with the project by submitting additional sequences and corrections. They hoped to base their collecting efforts on a gift economy, where researchers would contribute unpublished sequences and receive a copy of the *Atlas* in return. In a move that would have dire consequences for the future of their project, the authors made clear that they did not want to become involved in questions of "history or priority."[117] If the publication of sequences in the *Atlas* did not establish priority, then authors would not get credit for their work, a major incentive for making their results public. Another feature of the *Atlas* made some users uncomfortable was the fact that is was copyrighted and could thus not be redistributed. The proprietary status of the *Atlas* ran against the idea that experimental results, once published, should be freely available.[118]

Dayhoff sent out the *Atlas* to more than 70 scientists in the United States, Canada, Japan, and Europe. The list of recipients included all the researchers who had determined sequences which had been included

[113] Dayhoff et al., 1965, unnumbered page.

[114] *Ibid.*, p. 2.

[115] *Ibid.*, p. 2. On the issue of precision and objectivity, see Suárez-Diaz and Anaya-Muñoz, 2008.

[116] Dietrich, 1998; Hagen, 1999.

[117] Eck and Dayhoff, 1966b, p. xiv.

[118] On this issue, see Strasser, 2006a, 2008.

in the *Atlas*, those who had analyzed sequences, editors of major scientific journals, and Nobel Prize-winning scientists.[119] The reactions of the recipients were generally enthusiastic. The Nobel Prize-winning chemist Melvin Calvin, for example, believed the *Atlas* would "ultimately prove to be a veritable dictionary of biological activity."[120] Another Nobelist, the geneticist Joshua Lederberg stressed that the *Atlas* would become "an important contribution to the next stage of molecular biological architecture" and would be a crucial tool in the "computer search for active site configurations" in proteins.[121] Understandably, the biochemists who had been trying to keep up with all the known sequences were the most pleased by the *Atlas*. Emanuel Margoliash, for example, stressed the role of the *Atlas* as a repository of sequences: "It will clearly become a most valuable compilation, particularly as this sort of information accumulates and one's memory begins to be overburdened with all the details."[122]

But the reactions also reflected a prevailing view that the *Atlas*, while useful, represented a mere compilation of known sequences. According to this view, the "compilers," as the Nobel Prize-winning chemist Richard L. M. Synge addressed Dayhoff and her team,[123] had simply gathered data which were available in the published literature and reprinted them, something that could hardly qualify as a scientific contribution. When Dayhoff applied to become a member of the American Society of Biological Chemists, the biochemist John T. Edsall answered, a bit embarrassed: "Personally I believe that you are the kind of person who should become a member of the American Society of Biological Chemists. [However, the candidate must] demonstrate that he or she has done research which is clearly his own. The compilation of the Atlas of Protein Sequence and Structure scarcely fits into this pattern."[124] Another potential supporter of Dayhoff's application, the biochemist William H. Stein, also discouraged her because she did "not do experimental work."[125] The idea that the compilation of sequences, unlike their experimental determination, did not count as a scientific contribution, would plague the development of sequence databases for the decades to come, and explains a great deal of the resistance to their support displayed

[119] Correspondence index card set, 1965, NBRF Archives.

[120] Melvin Calvin to Margaret O. Dayhoff, February 11, 1966, NBRF Archives.

[121] Joshua Lederberg to Margaret O. Dayhoff, March 12, 1964, NBRF Archives.

[122] Emanuel Margoliash to Richard V. Eck, February 2, 1966, NBRF Archives.

[123] Richard Synge to "Compilers," April 7, 1966, NBRF Archives.

[124] John T. Edsall to Margaret O. Dayhoff, November 4, 1969, NBRF Archives.

[125] William H. Stein to Margaret O. Dayhoff, December 4, 1969, NBRF Archives.

by science funding agencies. Almost two decades later, after the NIH had turned down one of her grant requests for the *Atlas*, Dayhoff complained once again, "databases do not inspire excitement."[126]

Compiling the *Atlas* required far more expertise (and effort) than most of the users were ready to admit. First, the scientific literature had to be systematically surveyed, either manually or using bibliographic indexing systems such as MEDLARS and the American Chemical Society Abstracts search service.[127] One of the challenges, after having located an article, was careful proofreading of the sequence it reported. This task was necessary because, as Dayhoff complained, "there is scarcely a paper published that doesn't have at least one typographical error in the data."[128] Because the errors were not immediately obvious and there was no dictionary of sequences available for comparison, the proofreading process actually required a careful evaluation of the experimental data from which the proposed sequence had been deduced. And as biochemist Christian B. Anfinsen had admitted, a "certain amount of personal judgment is frequently involved" in the production of sequencing data.[129] In addition to collecting and evaluating sequences, Dayhoff and her team made several decisions about the kind of data to be included and how best to represent it in a way that would be useful to researchers. These decisions reflected a serious engagement with the scientific research based on protein sequences, and not the passive collection of existing data.[130]

A clear intellectual contribution of the *Atlas* that should have been recognizable as such from the outset was the information that accompanied the sequences. Indeed, beginning with the second edition, published in 1966 by Dayhoff and Eck, the *Atlas* included introductions to the current knowledge about the structure of proteins, new methods to analyze them, and inferences that could be drawn using these methods.[131] Most of these contributions concerned the evolution of proteins and the evolutionary relationships between species. These methods became part of a series a computer programs developed by Eck and Dayhoff to analyze the data contained in the *Atlas*.

[126] Margaret O. Dayhoff to DM Moore, September 24, 1981, NBRF Archives.

[127] Medlars was the computerized version of the Index Medicus which had become available in 1964. Margaret O. Dayhoff and Richard V. Eck to the NSF, "Application for publication support," 1966, p. 5, NBRF Archives.

[128] Margaret O. Dayhoff to D. Haas, April 11, 1969, NBRF Archives.

[129] Anfinsen, 1959, p. 146.

[130] Dayhoff would hold to this position all her life, often in face of fierce opposition from funding agencies who believed the two activities should be kept separate.

[131] Eck and Dayhoff, 1966a, b.

One of these programs addressed the key problem for those measuring evolutionary relations through protein sequence differences, namely to determine the real number of amino acid changes that had occurred in a protein over time, as opposed to the apparent number of changes, which could be lower due to multiple changes in a single position. By analyzing all the data contained in the *Atlas* with a program named ALLELE, Dayhoff and Eck constructed a matrix of probabilities of amino acid exchanges which they could then use to estimate, from the observed number of changes, the actual number of replacements which had occurred over time in a given protein. In the case of cytochrome *c*, for example, two sequences showing 52 differences were estimated to have actually undergone as many as 92 changes. This information made it possible to draw a phylogenetic tree to scale with the length of the branches representing the actual number of changes that had occurred between two protein since their divergence.[132] Dayhoff refined this approach over the years; it became known as the "Dayhoff matrix" or PAM ("point accepted mutation") matrix, and was widely used by molecular evolutionists.[133]

Dayhoff and Eck also attempted to find a computerized way to determine the topology of a phylogenetic tree, not just the length of its branches, by inferring ancestral sequences.[134] In the 1966 edition of the *Atlas*, the authors noted that a mathematical procedure had not yet been presented "in the detail necessary for a computer program," but that "arguments based on this approach [had] been used by Pauling, Zuckerkandl, and others,"[135] referring probably to the idea of minimizing the number of mutations in a topology, or of considering a constant mutation rate. The computer program developed by Dayhoff and Eck compared numerous topologies which were compatible with the data and picked the most likely one. However, as the authors noted, "since there are usually several millions of possible topological configurations, it is impracticable to try them all in the search for a minimum, even on a high-speed computer."[136] The program thus proceeded stepwise, beginning with three sequences, and suggesting a possible ancestral

[132] Eck and Dayhoff, 1966a, p. 163.

[133] Felsenstein, 2004, Chap. 10.

[134] For a useful comparison of early tree building methodologies, see Felsenstein, 2004, Chap. 10.

[135] Eck and Dayhoff, 1966a, p. 165.

[136] *Ibid.*, p. 165. Two years later, the NBRF would acquire its own mainframe computer, an IBM 360 Model 44, instead of using the shared IBM 7090. NBRF Archives, "Application (renewal), NASA 21-003-002," 1968.

sequence from which they had derived. Then it considered one additional sequence and repeated the operation.[137] At each step, researchers could decide whether to continue the path suggested by the computer or to make manual adjustments in the inferred sequences based on physicochemical or other considerations. As with other programs developed at the NBRF, computers did not replace humans, they assisted them. Once a topology was chosen and the number of mutations minimized, another computer program determined the length of the branches in geological time, assuming a constant rate of mutation for each set of homologous proteins, and taking the elapsed time of be proportional to the number of inferred mutational changes. In multi-page foldouts included in the *Atlas*, Dayhoff and Eck presented the results of their calculations as phylogenetic trees and provided the alignments of the sequences they had used to produce them.

The 1966 edition included, for example, a phylogeny based on cytochrome *c*, comprising organisms ranging from yeast to man and including tuna fish and kangaroo. It was topologically identical to that published a year later by Walter M. Fitch and Emanuel Margoliash,[138] which is often mistaken as the first phylogeny based on protein sequences produced using a computer. The methodology of the two groups differed in a significant way, however. Whereas Fitch and Margoliash weighted the amino acid differences with a value of one to three, depending on the number of nucleotide changes that were required according to the genetic code, Dayhoff and Eck, using all the data contained in their *Atlas*, weighted the amino acid differences based on the empirically observed frequency of amino acid change, thus obtaining a more realistic estimate of the actual numbers of mutations.[139] These different methodologies may have reflected different theoretical assumptions about the effects of natural selection at the molecular level, but they also reflected the possibilities offered to Dayhoff and Eck, and not to others, to use a computerized collection of protein sequences to derive statistical regularities essential for reconstructing the course of evolution from molecular data. Indeed, the punch cards that were used to store the data and print the *Atlas* could be

[137] For a later description, see Dayhoff, 1969.

[138] Fitch and Margoliash, 1967, Figure 2.

[139] Russell F. Doolittle and Birger Blombäck, in a phylogeny published in 1964 based on variations in fibrinopeptide sequences, used an even more crude method than Fitch and Margoliash, simply counting the number of amino acid changes, without any weighting. Doolittle and Blombäck, 1964. Margoliash had done the same a year earlier Margoliash, 1963.

directly fed into the computer and analyzed with the programs developed by Dayhoff and her collaborators.

The sometimes disparaging attitude of biochemists toward the intellectual, as opposed to the practical, value of the *Atlas* reflected their general attitude toward theoretical approaches in science. As Thomas H. Jukes would put it in 1963, a theoretical approach "is not acceptable to many biochemists, being inductive, rather than deductive."[140] The German biochemist Gerhard Braunitzer, whose many hemoglobin sequences were included in the *Atlas*, told Dayhoff bluntly: "I am not a theorizer," but he nevertheless valued the data compiled in the *Atlas*.[141] The American biochemist John T. Edsall made the same point when he wrote to Dayhoff that he had used the *Atlas* "primarily as a source of data and [had] not read very much of [the] interpretative material."[142] Frederick Sanger would later express his preference for the empirical even more clearly: "'Doing' for a scientist implies doing experiments"[143] – not, he might have added, collecting facts determined by others or engaging in theoretical speculations.

Gender probably also played a part in the valuation of Dayhoff's *Atlas*. The work for the *Atlas* was almost exclusively carried out by women. Indeed, besides Richard V. Eck, all of Dayhoff's collaborators were women. The microbiologist Minnie R. Sochard (1931–) and the applied mathematician Marie A. Chang (1937–), both co-authors of the first edition of the *Atlas*, the biologist Lois D. Hunt (1933–) and four other women assisted in the project. The *Atlas* project thus squarely fit the model of earlier scientific endeavors where groups of women were employed to perform repetitive tasks, as "computers" or as surveyors, for example.[144] Dayhoff, advising a young female colleague about her career, warned her about participating in "the 'masculine' scientific world," but argued that those who eventually decided for a scientific career brought "to this desert a range of feminine concerns that have been completely overlooked."[145] As an example, she took her work with the *Atlas*:

[140] Jukes, 1963, p. 2. On theory as a dividing line between biochemist and molecular biologists, see Abir-Am, 1992.

[141] Gerhard Braunitzer to Margaret O. Dayhoff, April 18, 1968, NBRF Archives.

[142] John T. Edsall to Margaret O. Dayhoff, December 3, 1968, NBRF Archives. The interpretive material was presented in a discursive format and had not been peer reviewed, two factors that may have predisposed other scientists to ignore it.

[143] Sanger, 1988, p. 1.

[144] Light, 1999.

[145] Margaret O. Dayhoff to S. Tideman, October 18, 1968, NBRF Archives.

> I realized that the answers people were giving to social problems were very shallow and naïve – often only palliative in nature. Our knowledge of ourselves is quite medieval […] I like to think that the *Atlas* and related research are going to help in the gigantic endeavor to solve these vexing problems. Species differences, race differences, sex differences, and individual differences, are largely controlled by protein differences. Motivation and mental capacity, goals and satisfactions, as well as diseases may be linked to proteins. We shift over our fingers the first grains of this great outpouring of information and say to ourselves that the world be helped by it. The *Atlas* is one small link in the chain from biochemistry and mathematics to sociology and medicine.

Although the extent to which Dayhoff's social concerns and the entire *Atlas* endeavor should be linked to her gender is debatable; nonetheless, it is evident that she operated in a framework in which the intellectual value of "passive" collecting was perceived, by her and by others, to be gendered in way that made it less authoritative than "active" experimenting.[146]

Conclusions

In this paper I have tried to bring into historical perspective the central epistemic and material practices underlying the current use of molecular sequences to produce biological knowledge. The practices of collecting, comparing, and computing amino acid sequences grew out of distinct research agendas of the 1950s – the investigation on the structural basis of molecular function, the attempt to solve the genetic code, and the application of computers to the management and analysis of biological information. The *Atlas of Protein Sequence and Structure* drew heavily on this heterogeneous set of research agendas and combined them into a powerful instrument for the production of biomedical knowledge. The success of this collection as a research tool in fields ranging from protein biochemistry to molecular evolution attests to the heuristic value of these practices and to the rise of sequences as key object of study in the life sciences. By creating this sequence collection and developing methods to analyze sequences, Dayhoff made sequences matter, in a way that has grown in importance ever since the creation of the *Atlas*.

[146] Keller, 1992.

The historical trajectory of these practices offers an opportunity to reconsider the historiography of molecular biology and, more generally, of the life sciences in the twentieth century. The achievements of molecular biology and the power of the molecular approach to the study of life have been attributed to several factors, including their reliance on a very limited number of model organisms such as the T-even bacteriophages, *Escherichia coli*, *Neurospora crassa*, *Drosophila melanogaster*, or *Caenorhabditis elegans*.[147] Yet, as I have tried to show here, some essential contributions of molecular biology and biochemistry, such as the understanding of the structural basis of protein function and the solution to the genetic code, relied on a much wider range of organisms including salmon, pigs, silkworms and reindeer (among many others). Biochemists, who had been trained in the comparative biochemistry tradition, provided the kind of attention to biological diversity that physicists-turned-molecular-biologist, for example, often lacked. Via comparative biochemistry, the field of molecular biology thus seems to have borrowed, much more than has been recognized previously, from the natural historical "way of knowing," relying on the comparisons of structures from very diverse organisms. A broader historical exploration of how a wide range of organisms entered laboratories and came to be studied at the molecular level would shed additional light on the development of the comparative perspective within the experimental sciences.[148]

Historians of the life science have insisted on the transformative role of scientific instruments such as electron microscopes, ultracentrifuges, and electrophoresis apparatuses in the development of a molecular vision of life.[149] Surprisingly, the computer has been almost completely left out of the picture.[150] Whereas the informational vision of life based on information theory and postwar cybernetics has been extensively studied,[151] we still lack a satisfying historical understanding of the impact of computers and networks, not as metaphors for living systems,[152] but as forces shaping the material and social dynamics of the biological

[147] For a model organism-centered history of biology, see Endersby, 2007.

[148] For an exploration of how wild organisms came to be studied experimentally in serological taxonomy, see Strasser 2010.

[149] See Rasmussen, 1997; Elzen, 1986; Kay, 1988, respectively.

[150] The see however Lenoir, 1999; Hagen, 2001; de Chadarevian, 2002, Chap. 4; Francoeur and Segal, 2004; November, 2004.

[151] Kay, 2000; Segal, 2003.

[152] Keller, 1995.

sciences.[153] When computers have been considered at all in the historiography of the life sciences, it has been mainly as powerful calculators. Yet computers have played a much more diverse role in the development of biology, in particular by making the management and comparison of vastly increased amounts of information possible, as the computerized *Atlas of Protein Sequence and Structure* attests. Dayhoff's sequence collection, which became available on magnetic tapes, and eventually online in 1978, represents just one example of the burgeoning computerization of biological information. A number of other computerized data collections, such as the Protein Data Bank (protein structure coordinates) and GenBank (DNA sequences), were created on the same model as the *Atlas* and similarly became essential tools for biomedical research. The introduction of computers into the management of scientific data offers several promising venues for future research. For example, all these databases have combined scientific data and bibliographic literature and were shaped by techniques developed in library science, a field at the cutting-edge of the management of computerized records, and yet one largely neglected by historians of science. The place of computers in the counterculture movements of the 1970s and the development of shared computer resources also offer promising avenues to understanding the rise of open access practices in science.[154]

Finally, the importance of biological collections for the rise of the experimental life sciences constitutes an invitation to revisit the relations between the natural historical and the experimental "ways of knowing." Elsewhere, I have compared the *Atlas* to natural history collections including botanical, zoological, and anatomical collections, and discussed how collecting constituted perhaps the most defining practice of natural history.[155] I have examined the later challenges faced by Dayhoff in collecting sequences from experimentalists for her *Atlas* and tried to show how the moral economy on which she based her collecting efforts – in particular, its ideas about property, privacy, and priority – resonated with that of the natural history tradition, but conflicted with that of the experimental sciences. The comparison of features among a wide range of species is also a characteristic epistemic practice in natural history, and has played a similarly key role in the rise of sequence collections, where protein and, later, nucleic acid sequences were compared to produce knowledge about molecular structure, function, and evolution. This is not to say that those engaged in the collection and

[153] For such a project on computers and American political culture, see Edwards, 1996.

[154] For preliminary exploration of this theme, see Strasser, 2008.

[155] Strasser, 2006a, 2008, 2010.

comparison of molecular sequences were naturalists – they were not. Their professional identities were those of experimentalists and theoreticians, but a number of their research practices, like those of the naturalists, rested on a "way of knowing" which valued collecting and comparing features from various species in order to produce knowledge about the natural world. This way of doing science put them in an uneasy professional position, where the legitimacy of their approaches was challenged from all sides – by experimentalists, naturalists, and theoreticians. The dominant narrative of the powerful experimental (molecular) sciences taking over all fields in the biological and medical sciences in the twentieth century thus needs to be qualified. Molecular evolution, for one, can be understood not only as the introduction of molecular concepts and practices into "traditional" evolutionary studies but also as the importation of comparative and other natural historical practices into experimental approaches to evolution. By examining more broadly the history of sequence databases and other biological collections, one might be able to write a different history of the life sciences in the twentieth century than the one that has focused almost exclusively on the triumphs of experimentation over natural history.

Acknowledgments

I would like to thank Robert S. Ledley for giving me access to the archives of the NBRF, Winona C. Barker, Ruth E. Dayhoff, and Judith E. Dayhoff for sharing material and memories, the participants at the workshop and conferences at the University of Pennsylvania, the Max Planck Institute for the History of Science, and Yale University, as well as Helen Curry, Marianne Sommer, Angela Creager, Joel Hagen, Sage Ross, Michael Dietrich, Brendan Matz, and Joe November, as well as an anonymous referee for helpful comments.

References

Abir-Am, Pnina. 1992. "The Politics of Macromolecules Molecular Biologists, Biochemists, and Rethoric." *Osiris* 7: 164–191.

Allen, Garland E. 1978. *Life Science in the Twentieth Century*. Cambridge/London: Cambridge University Press.

Anfinsen, Christian B. 1959. *The Molecular Basis of Evolution*. New York: Wiley.

Anfinsen, Christian B, Åqvist, Stig E.V., Cooke, Juanita P. and Jönsson, Börje. 1959. "A Comparative Study of the Structures of Bovine and Ovine Pancreatic Ribonucleases." *Journal of Biological Chemistry* 234(5): 1118–1123.

Anonymous. 1962. "Computing in the University." *Datamation* 8: 27–30.
Aronson, Jay D. 2002. "'Molecules and Monkeys': George Gaylord Simpson and the Challenge of Molecular Evolution." *History Philosophy of Life Sciences* 24(3–4): 441–465.
Aspray, William and Williams, Bernard O. 1994. "Arming American Scientists – NSF and the Provision of Scientific Computing Facilities for Universities, 1950–1973." *IEEE Annals of the History of Computing* 16(4): 60–74.
Baldwin, Ernest. 1937. *An Introduction to Comparative Biochemistry*. Cambridge: The University Press.
Baldwin, Ernest. 1966. *An Introduction to Comparative Biochemistry*. Cambridge: The University Press.
—— 1962. "Synthetic Polynucleotides and the Amino Acid Code. V." *Proceedings National Academy Science* 48: 613–616.
Benson, Keith R. 1988. "From Museum Research to Laboratory Research: The Transformation of Natural History into Academic Biology." Ronald Rainger, Keith R. Benson and Jane Maienschein (eds.), *The American Development of Biology*. Philadelphia: University of Pennsylvania Press, pp. 49–83.
Bernhard, S.A., Bradley, D.F. and Duda, W.L. 1963. "Automatic Determination of Amino Acid Sequences." *IBM Journal of Research and Development* 7(3): 246–251.
Blombäck, Birger, Blombäck, Margareta and Grondahl, Nils Jakob. 1965. "Studies on Fibrinopeptides from Mammals." *Acta Chemica Scandinavica* 19: 1789–1791.
Bowler, Peter J. and Morus, Iwan Rhys. 2005. *Making Modern Science: A Historical Survey*. Chicago: University of Chicago Press.
Brandt, Christina. 2004. *Metapher und Experiment: von der Virusforschung zum genetischen Code*. Göttingen: Wallstein.
Brenner, Sydney. 1957. "On the Impossibility of All Overlapping Triplet Codes in Information Transfer from Nucleic Acid to Proteins." *Proceedings National Academy Science* 43(8): 687–694.
Brown, H, Sanger, Frederick and Kitai, Ruth. 1955. "The Structure of Pig and Sheep Insulins." *Biochemical Journal* 60(1–4): 556–565.
Chargaff, Erwin. 1955. "Isolation and Composition of the Desoxypentose Nucleic Acids and of the Corresponding Nucleoproteins." Erwin Chargaff and J.N. Davidson (eds.), *The Nucleic Acids: Chemistry and Biology*. New York: Academic Press.
Coleman, William. 1971. *Biology in the Nineteenth Century: Problems of Form, Function and Transformation*. Cambridge: Cambridge University Press.
Creager, Angela N.H. 2002. *The Life of a Virus: Tobacco Mosaic Virus as an Experimental Model, 1930–1965*. Chicago: University of Chicago Press.
Dayhoff, Margaret O. 1964. "Computer Search for Active Site Configurations." *Journal of the American Chemical Society* 86(11): 2295–2297.
—— 1969. "Computer Analysis of Protein Evolution." *Scientific American* 221: 86–95.
Dayhoff, Margaret O. and Ledley, Robert S. 1962. "Comprotein: A Computer Program to Aid Primary Protein Structure Determination." *Proceedings of the Fall Joint Computer Conference*. Santa Monica: American Federation of Information Processing Societies.
Dayhoff, Margaret O., Lippincott, E.R. and Eck, Richard V. 1964. "Thermodynamic Equilibria in Prebiological Atmospheres." *Science* 146(1461): 1461–1464.
Dayhoff, Margaret O, Eck, Richard V, Chang, Marie A. and Sochard, Minnie R. 1965. *Atlas of Protein Sequence and Structure*. Silver Spring: National Biomedical Research Foundation.

Dayhoff, Margaret O., Eck, Richard V., Lippincott, E.R. and Sagan, Carl. 1967. "Venus: Atmospheric Evolution." *Science* 155(3762): 556–558.

de Chadarevian, Soraya. 1996. "Sequences, Conformation, Information: Biochemists and Molecular Biologists in the 1950s." *Journal of the History of Biology* 29(3): 361–386.

—— 1998. "Following Molecules: Haemoglobin Between the Clinic and the Laboratory." Soraya de Chadarevian and Kamminga Harmke (eds.), *Molecularizing Biology and Medicine: New Practices and Alliances, 1910s–1970s*. Amsterdam: Harwood Academic Publishers, pp. 171–201.

—— 1999. "Protein Sequencing and the Making of Molecular Genetics." *Trends in Biochemical Sciences* 24: 203–206.

—— 2002. *Designs for Life. Molecular Biology after World War II*. Cambridge: Cambridge University Press.

de Solla Price, Derek J. 1963. *Little Science, Big Science*. New York: Columbia University Press.

Dietrich, Michael R. 1994. "The Origins of the Neutral Theory of Molecular Evolution." *Journal of the History of Biology* 27: 21–59.

—— 1998. "Paradox and Persuasion: Negotiating the Place of Molecular Evolution Within Evolutionary Biology." *Journal of the History of Biology* 31: 85–111.

Doolittle, Russell F. and Blombäck, Birger. 1964. "Amino-Acid Sequence Investigations of Fibrinopeptides from Various Mammals: Evolutionary Implications." *Nature* 202: 147–152.

Doolittle, Russell F., Singer, Seymour J. and Metzger, Henry. 1966. "Evolution of Immunoglobulin Polypeptide Chains: Carboxy-Terminal of an IgM Heavy Chain." *Science* 154(756): 1561–1562.

Eck, Richard V. 1961. "Non-Randomness in Amino-Acid 'Alleles'." *Nature* 191: 1284–1285.

—— 1962a. "A Simplified Strategy for Sequence Analysis of Large Proteins." *Nature* 193: 241–243.

—— 1962b. "I. The Protein Cryptogram: I Non-Random Occurrence of Amino Acid "Alleles"." *Journal of Theoretical Biology* 2: 139–151.

—— 1964 "Cryptogrammic Detection of a Pattern in Amino Acid "Alleles": Its Use in Tracing the Evolution of Proteins." *Proceedings on the 17th Annual Conference on Engineering in Medicine and Biology*, Vol. 6, p. 115.

Eck, Richard V. and Dayhoff, Margaret O. 1966a. "Evolution of the Structure of Ferredoxin Based on Living Relics of Primitive Amino Acid Sequences." *Science* 152(3720): 363–366.

—— 1966b. *Atlas of Protein Sequence and Structure*. Silver Spring, MD: National Biomedical Research Foundation.

Edwards, Paul N. 1996. *The Closed World: Computers and the Politics of Discourse in Cold War America*. Cambridge, MA: MIT Press.

Elzen, Boelie. 1986. "Two Ultracentrifuges: A Comparative Study of the Social Construction of Artefacts." *Social Studies of Science* 16: 621–662.

Endersby, Jim. 2007. *A Guinea Pig's History of Biology*. Cambridge: Harvard University Press.

Farber, Paul Lawrence. 2000. *Finding Order in Nature: The Naturalist Tradition from Linnaeus to E. O. Wilson*. Baltimore/London: The Johns Hopkins University Press.

Felsenstein, Joseph. 2004. *Inferring Phylogenies*. Sunderland, MA; Sinauer Associates.

Fitch, Walter M. 1964. ''The Probable Sequence of Nucleotides in Some Codons.'' *Proceedings National Academy Science* 52: 298–305.

Fitch, Walter M. 1966a. ''Evidence Suggesting a Partial, Internal Duplication in the Ancestral Gene for Heme-Containing Globins.'' *Journal of Molecular Biology* 16(1): 17–27.

—— 1966b. ''The Relation Between Frequencies of Amino Acids and Ordered Trinucleotides.'' *Journal of Molecular Biology* 16(1): 1–8.

Fitch, Walter M. and Margoliash, Emanuel. 1967. ''Construction of Phylogenetic Trees.'' *Science* 155(760): 279–284.

Florkin, Marcel. 1944. *L'évolution Biochimique*. Paris: Masson & cie.

—— 1949. *Biochemical Evolution*. New York: Academic Press.

Francoeur, Eric and Segal, Jérôme. 2004. ''From Model Kits to Interactive Graphics.'' S. de Chadarevian and N. Hopwood (eds.), *Models: The Third Dimension of Science*. Stanford, CA: Stanford University Press, pp. 402–429.

Gamow, George. 1954. ''Possible Relation Between Desoxyribonucleic Acid and Protein Structure.'' *Nature* 173: 318.

Gamow, George and Metropolis, Nicolas. 1954. ''Numerology of Polypeptide Chains.'' *Science* 120(3124): 779–780.

Gamow, George, Rich, Alexander and Yčas, Martynas. 1956. ''The Problem of Information Transfer from the Nucleic Acids to Proteins.'' *Advances in Biological and Medical Physics* 4: 23–68.

Gamow, George and Yčas, Martynas. 1955. ''Statistical Correlation of Protein and Ribonucleic Acid Composition.'' *Proceedings National Academy Science* 41(12): 1011–1019.

Garcia-Sancho, Miguel. 2010, in press. ''A New Insight into Sanger's Development of Sequencing: From Proteins to DNA, 1943–1977.'' *Journal of the History of Biology*.

Gaudillière, Jean-Paul. 2002. *Inventer La Biomédecine: La France, l'Amérique et la Production des Savoirs du Vivant: 1945–1965*. Paris: La Découverte.

Hagen, Joel B. 1999. ''Naturalist, Molecular Biology, and the Challenge of Molecular Evolution.'' *Journal of the History of Biology* 32: 321–341.

—— 2000. ''The Origins of Bioinformatics.'' *Nature Reviews* 1: 231–236.

—— 2001. ''The Introduction of Computers into Systematic Research in the United States During the 1960s.'' *Studies in the History and Philosophy of Biological and Biomedical Sciences* 32(2): 291–314.

Harris, J. Ieuan, Naughton, Michael A. and Sanger, Frederick. 1956. ''Species Differences in Insulin.'' *Archives of Biochemistry and Biophysics* 65(1): 427–438.

Hunt, John A. and Ingram, Vernon M. 1958. ''The Chemical Effects of Gene Mutations in Some Abnormal Human Haemoglobins.'' Albert Neuberger (ed.), *Symposium on Protein Structure*. New York: Wiley.

Jardine, Nicholas, Secord, James A. and Spary, Emma C. (eds.). 1996. *Cultures of Natural History*. London/New York: Cambridge University Press.

Jukes, Thomas H. 1962a. ''Beta Lactoglobulins and Amino Acid Code.'' *Biochemical and Biophysical Research Communications* 7(4): 281–283.

—— 1962b. ''Possible Base Sequences in Amino Acid Code.'' *Biochemical and Biophysical Research Communications* 7(6): 497–502.

—— 1962c. ''Relations Between Mutations and Base Sequences in Amino Acid Code.'' *Proceedings of the National Academy of Sciences of the United States of America* 48(10): 1809–1815.

—— 1963. "Some Recent Advances in Studies of the Transcription of the Genetic Message." *Advances in Biological and Medical Physics* 9: 1–41.

Kay, Lily E. 1993. *The Molecular Vision of Life. Caltech, the Rockefeller Foundation and the Rise of the New Biology*. New York: Oxford University Press.

—— 1988. "Laboratory Technology and Biological Knowledge: The Tiselius Electrophoresis Apparatus, 1930–1945." *History and Philosophy of the Life Science* 10:51–72.

—— 2000. *Who Wrote the Book of Life. A History of the Genetic Code*. Stanford: Sanford University Press.

Keller, Evelyn Fox. 1992. *Secrets of Life, Secrets of Death: Essays on Language, Gender, and Science*. New York: Routledge.

—— 1995. *Refiguring Life, Metaphors of Twentieth-Century Biology*. New York: Columbia University Press.

—— 2000. *The Century of the Gene*. Cambridge: Harvard University Press.

Kohler, Robert E. 1982. *From Medical Chemistry to Biochemistry. The Making of a Biomedical Discipline*. Cambridge: Cambridge University Press.

—— 2002. *Landscapes and Labscapes: Exploring the Lab-Field Border in Biology*. Chicago: Chicago University Press.

Ledley, Robert S. 1955. "Digital Computational Methods in Symbolic Logic, with Examples in Biochemistry." *Proceedings of the National Academy of Sciences* 41(7): 498–511.

—— 1959a. "Reasoning Foundations of Medical Diagnosis; Symbolic Logic, Probability, and Value Theory Aid Our Understanding of How Physicians Reason." *Science* 130(3366): 9–21.

—— 1959b. "Digital Electronic Computers in Biomedical Sciences." *Science* 130(3384): 1225–1234.

—— 1960. "Letters to the Editor." *Science* 131(3399): 474–564.

—— 1965. *Use of Computers in Biology and Medicine*. New York/Saint Louis: McGraw-Hill.

Ledley, Robert S. and Lusted, L.B. 1959. "Probability, Logic and Medical Diagnosis." *Science* 130(3380): 892–930.

Lengyel, Peter, Speyer, Joseph F. and Ochoa, Severo. 1961. "Synthetic Polynucleotides and the Amino Acid Code." *Proceedings of the National Academy of Sciences* 47: 1936–1942.

Lengyel, Peter, Speyer, Joseph F, Basilio, Carlos and Ochoa, Severo. 1962. "Synthetic Polynucleotides and the Amino Acid Code. III." *Proceedings of the National Academy of Sciences* 48: 282–284.

Lenoir, Timothy. 1999. "Shaping Biomedicine as an Information Science." M.E. Bowden, T.B. Hahn and R.V. Williams (eds.), *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems*. Medford: Information Today, pp. 27–45.

Light, Junnifer S. 1999. "When Computers Were Women." *Technology and Culture* 40(3): 455–483.

Margoliash, Emanuel. 1963. "Primary Structure and Evolution of Cytochrome *C*." *Proceedings of the National Academy of Sciences* 50: 672–679.

Medical Research Council. 1965. *Mathematics and Computer Science in Biology and Medicine*. London: H. M. Stationery Office.

Miles, Wyndham D. 1982. *A History of the National Library of Medicine: The Nation's Treasury of Medical Knowledge*. Washington, DC: U.S. Department of Health and Human Services.

Miller, Stanley L. and Urey, Harold C. 1959. "Organic Compound Synthesis on the Primitive Earth." *Science* 130(3370): 245–251.

Moore, Stanford, Spackman, Darrel H. and Stein, William H. 1958. "Automatic Recording Apparatus for Use in the Chromatography of Amino Acids." *Federation Proceedings* 17(4): 1107–1115.

Morange, Michel. 2000. *A History of Molecular Biology*. Cambridge: Harvard University Press.

Morgan, Gregory J. 1998. "Emile Zuckerkandl, Linus Pauling, and the Molecular Evolutionary Clock, 1959–1965." *Journal of the History of Biology* 31: 155–178.

November, Joseph A. 2004. "LINC: Biology's Revolutionary Little Computer." *Endeavour* 28(3): 125–131.

—— 2006. *Digitalizing Life: The Introduction of Computers to Biology and Medicine*. Doctoral Thesis, Princeton University.

Nyhart, Lynn K. 1996. "Natural History and the 'New' Biology." Nicholas Jardine, James A. Secord and C. Spary Emma (eds.), *Cultures of Natural History*. London: Cambridge University Press.

Oakley, Margaret B. and Kimball, George E. 1949. "Punched Card Calculation of Resonance Energies." *Journal of Chemical Physics* 17(8): 706–717.

Paléus, Sven and Tuppy, Hans. 1959. "A Hemopeptide from a Tryptic Hydrolysate of Rhodospirillum-Rubrum Cytochrome-*C*." *Acta Chemica Scandinavica* 13(4): 641–646.

Pickstone, John V. 1993. "Ways of Knowing: Towards a Historical Sociology of Science, Technology and Medicine." *British Journal for the History of Science* 26: 433–458.

—— 2007. "Working Knowledges Before and After Circa 1800. Practices and Disciplines in the History of Science, Technology and Medicine." *Isis* 98: 489–516.

Rasmussen, Nicolas. 1997. *Picture Control the Electron Microscope and the Transformation of Biology in America, 1940–1960*. Stanford: Stanford University Press.

Rogers, Frank B. 1964. "The Development of MEDLARS." *Bulletin of the Medical Library Association* 52: 150–151.

Sanger, Frederick. 1949. "Species Differences in Insulins." *Nature* 164(4169): 529.

—— 1988. "Sequences, Sequences, and Sequences." *Annual Review of Biochemistry* 57: 1–28.

Segal, Jérôme. 2003. *Le Zéro et le Un. Histoire de la Notion Scientifique d'Information*. Paris: Syllepse.

Shapiro, Marvin B., Merril, Carl R., Bradley, Dan F. and Mosimann, James E. 1965. "Reconstruction of Protein and Nucleic Acid Sequences: Alamine Transfer Ribonucleic Acid"." *Science* 150(698): 918–921.

Smith, Emil L. 1962a. "Nucleotide Base Coding and Amino Acid Replacements in Proteins." *Proceedings of the National Academy of Sciences* 48: 677–684.

—— 1962b. "Nucleotide Base Coding and Amino Acid Replacements in Proteins. II." *Proceedings of the National Academy of Sciences* 48: 859–864.

Sommer, Marianne. 2008. "History in the Gene: Negotiations Between Molecular and Organismal Anthropology." *Journal of the History of Biology* 43: 473–528.

Spath, Susan B. 1999. *C. B. van Niel and the Culture of Microbiology, 1920–1965*. Doctoral Thesis, Berkeley University.

Speyer, Joseph F, Lengyel, Peter, Basilio, Carlos and Ochoa, Severo. 1962a. "Synthetic Polynucleotides and the Amino Acid Code. II." *Proceedings of the National Academy of Sciences* 48: 63–68.

—— 1962b. "Synthetic Polynucleotides and the Amino Acid Code. IV." *Proceedings of the National Academy of Sciences* 48: 441–448.

Stacy, Ralph W. and Waxman, Bruce D. 1965. *Computers in Biomedical Research*. New York: Academic Press.

Sterling, Theodor D. and Pollack, Seymour V. 1965. *Computers and the Life Sciences*. New York: Columbia University Press.

Strasser, Bruno J. 2006a. "Collecting and Experimenting: The Moral Economies of Biological Research, 1960s–1980s." *Preprints of the Max-Planck Institute for the History of Science* 310: 105–123.

—— 2006b. "A World in One Dimension: Linus Pauling, Francis Crick and the Central Dogma of Molecular Biology." *History and Philosophy of the Life Science* 28: 491–512.

—— 2006c. *La fabrique d'une nouvelle science: La biologie moléculaire à l'âge atomique (1945–1964)*. Florence: Olschki.

—— 2008. "Genbank: Natural History in the 21st Century?." *Science* 322: 537–538.

—— 2010, in press "Laboratories, Museums, and the Comparative Perspective: Alan A. Boyden's Quest for Objectivity in Serological Taxonomy, 1925–1962." *Historical Studies in the Natural Sciences*.

Strick, James E. 2004. "Creating a Cosmic Discipline: The Crystallization and Consolidation of Exobiology, 1957–1973." *Journal of the History of Biology* 37(1): 131–180.

Suárez-Diaz, Edna. 2007. "The Rhetoric of Informational Molecules: Authority and Promises in the Early Study of Molecular Evolution." *Science in Context* 20(4): 649–677.

—— 2009. "Molecular Evolution: Concepts and the Origin of Disciplines." *Studies in the History and Philosophy of Biological and Biomedical Sciences* 40(1): 43–53.

Suárez-Díaz, Edna and Anaya-Muñoz, Victor H. 2008. "History, Objectivity, and the Construction of Molecular Phylogenies." *Studies in the History and Philosophy of Biological and Biomedical Sciences* 39(4): 451–468.

Tsugita, Akira and Fraenkel-Conrat, Heinz. 1960. "The Amino Acid Composition and C-Terminal Sequence of a Chemically Evoked Mutant of TMV." *Proceedings of the National Academy of Sciences* 46(5): 636–642.

Tuppy, Hans. 1958. "Über die Artspezificität der Proteinstruktur." Albert Neuberger (ed.), *Symposium on Protein Structure*. New York: Wiley, pp. 66–76.

—— 1959. "Aminosaure-Sequenzen in Proteinen." *Naturwissenschaften* 46(2): 35–43.

Tuppy, Hans and Bodo, Gerhard. 1954. "Cytochrom c. III. Zur Frage der Artspezifität von Säugetier-Cytochrom c." *Monatshefte für Chemie* 85(5): 1182–1186.

Tuppy, Hans and Dus, K. 1958. "Eine Untersuchung über Cytochrom-c aus Hefe." *Monatshefte für Chemie* 89(3): 407–417.

Tuppy, Hans and Paléus, Sven. 1955. "Study of a Peptic Degradation Product of Cytochrome-*C*.1. Purification and Chemical Composition." *Acta Chemica Scandinavica* 9(3): 353–364.

Watson, James D. 2001. *Genes, Girls, and Gamow*. Oxford: Oxford University Press.

Wittmann, Heinz-Günter. 1960. "Comparison of the Tryptic Peptides of Chemically Induced and Spontaneous Mutants or Tobacco Mosaic Virus." *Virology* 12: 609–612.

Wolfe, Audra J. 2002. "Germs in Space. Joshua Lederberg, Exobiology, and the Public Imagination, 1958–1964." *Isis* 93: 183–205.

Yčas, Martinas. 1958. ''The Protein Text.'' Hubert P. Yockey (ed.), *Symposium on Information Theory in Biology*. New York: Pergamon Press.

—— 1961. ''Replacement of Amino Acids in Proteins.'' *Journal of Theoretical Biology* 1(2): 244.

Zuckerkandl, Emile and Pauling, Linus. 1962. ''Molecular Disease, Evolution, and Genic Heterogeneity.'' M. Kasha and B. Pullman (eds.), *Horizons in Biochemistry*. New York: Academic Press.

—— 1965. ''Molecules as Documents of Evolutionary History.'' *Journal of Theoretical Biology* 8: 357–366.