Title: Coding sequences: a history of sequence comparison algorithms as a scientific instrument.

Author: Hallam Stevens

Department of History of Science

Harvard University

1 Oxford Street

Science Center 371

Cambridge, MA 02138.

Post-print version: This article appeared in *Perspectives on Science,* volume 19, issue 3, pp. 263-299 and the final version is available at:
http://www.mitpressjournals.org/doi/abs/10.1162/POSC_a_00042?journalCode=posc#.VbYHJhOqqko

Abstract:

Sequence comparison algorithms are sophisticated pieces of software that compare and match identical or similar regions of DNA, RNA, or protein sequence. This paper examines the origins and development of these algorithms from the 1960s to the 1990s. By treating this software as a kind of scientific instrument used to examine sets of biological objects, the paper shows how algorithms have been used as different sorts of tools and appropriated for different sorts of uses according to the disciplinary context in which they were deployed. These particular uses have made sequences themselves into different kinds of objects.

Introduction

Historians of molecular biology have paid significant attention to the role of scientific instruments and their relationship to the production of biological knowledge. For instance, Lily Kay has examined the history of electrophoresis, Boelie Elzen has analyzed the development of the ultracentrifuge as an enabling technology for molecular biology, and Nicolas Rasmussen has examined how molecular biology was transformed by the introduction of the electron microscope (Kay 1998, 1993; Elzen 1986; Kohler 1994; Rasmussen 1997).[1] Collectively, these historians have demonstrated how instruments and other elements of the material culture of the laboratory have played a decisive role in determining the kind and quantity of knowledge that is produced by biologists. During the 1960s, a versatile new kind of instrument began to be deployed in biology: the electronic computer (Ceruzzi 2001; Lenoir 1999). Despite the significant role that computers now play in almost all of biology, they have received comparatively little historical attention.[2] Indeed, perhaps part of the reason for this is that computers have come to play so many roles in biological work that their influence has been difficult to analyze.[3] Computers are best understood not a single instrument, but as many different kinds of instruments. The aim of this paper is to examine just one way in which computers have been used: that is, as an instrument for comparing sequences to one another (protein sequences to protein sequences and nucleic acid sequences to nucleic acid sequences). Rather than taking computer hardware – the machine itself – as the appropriate object for analysis, I here examine the development of a particular form of *software* as an instrument. Unlike computer hardware, which can often be adapted to a number of uses (even simultaneously), software is usually designed for the solution of a well-defined problem. Michael Mahoney has argued that the computer is a 'protean machine' - an object that is "what we make of it

1 One could also point to Robert Kohler's (1994) work on the fruit fly, Jean-Paul Gaudillière (2001) on laboratory mice and Hannah Landecker (2007) on the technologies of tissue culture.

2 Exceptions to this include November 2006, Fujimura 1999, and Hagen 2001.

3 To name but a few, computers are used for simulations, for data analysis, for controlling instruments, for managing databases, and for managing laboratories.

(or have now made of it) through the tasks we set for it and the programs we write for it." What is important in the history of computing is not stories about physical devices, but analyzing "software as model, software as experience, software as medium of thought and action, software as environment within which people work and live..." (Mahoney 2005: 122, 127-128). It is through software that individuals do things with computers, and as such it is towards specific pieces of software that we should look if we want to understand what role computers have played in scientific practice.

Algorithms for comparing DNA and protein sequences have become the most ubiquitous, and most important, software tools for biology. Increasingly sophisticated and efficient methods for comparing sequences have allowed sequences to be assembled, ordered, categorized, and related to one another – that is to say, they have transformed sequences from arbitrary strings of letters into meaningful biological objects. When sequence comparison algorithms were first developed and used in the 1960s they were just one among many ways in which computers were deployed in biology.[4] Singling out this strand in the history of computational biology should not be taken to suggest that all other ways in which computers were used were unimportant. Rather, tracing the history of computational sequence comparison software as the history of an instrument will provide insight into why it became such an important tool in biological research. This paper, then, is the story of how algorithms have come to make sequences meaningful and what consequences these particular forms of meaning have. The paper will show how constellations of disciplinary, technological and epistemological circumstances produced sequence comparison algorithms as particular kinds of instruments. However, it has not only been the instruments themselves that have been transformed, but also biology and its objects of study. The introduction of computational methods, first by biochemists and later by mathematicians, physicists, and computer scientists, meant that the use of sequence comparison algorithms was closely tied to disciplinary debates over what counted as doing biology and

4 Others included simulation, modeling of evolution, prediction and visualization of protein structure, and data recording. See November 2006.

valid biological knowledge. The historical actors discussed in this paper remained, at least until the mid-1980s, remarkably isolated from the mainstream of biological research. In the 1960s and 1970s, using computers to do biology was an extremely unusual activity by most standards; even into the 1980s, the community of computational biologists was sufficiently small that everyone knew everyone else. There was almost no criticism of computational biology by other biologists – rather, they largely ignored it since it was not considered relevant to biology proper. Computational biologists were usually not trained as biologists – often they had picked up computer skills either from a training in physics, or in the context of other non-biological work. In short, the authors of sequence comparison algorithms were a small, isolated, and marginalized group.

Algorithmic innovations provided opportunities to ask and answer different sorts of questions. Behind these developments lay long term disputes over the legitimacy of computational techniques vis-à-vis traditional experimental approaches. These powerful softwares became tools for negotiating which kind of practices counted as proper biological work. Perhaps most importantly, these re-configurations made sequences themselves different kinds of objects with different kinds of meanings.

The paper describes a transition between two distinct uses of sequence comparison algorithms. This is not supposed to suggest that practices associated with the earlier period were completely replaced by the practices of the later decades. The progression was a synchronic process through which later practices overlaid – not completely replaced – earlier ones.[5] The first period traces the origins of sequence comparison algorithms – the argument here is not only that the introduction of computing was facilitated by the information discourse of molecular biology, but also that this discourse picked out a particular set of computational problems as interesting and important. These were the problems of tracing evolutionary history through the comparison of sequences. Since molecular biologists in 1950s

5 For instance, so-called molecular anthropologists have maintained a concerted effort to understand genes as historical origin narratives as can be seen from the literature surrounding the Human Genome Diversity Project, the HapMap project, and National Geographic's Genographic Project. See Reardon 2005 and Sommer 2008.

and 1960s understood DNA as a Book of Life, it followed that its writing must contain an origin story, an account of where we came from. Such an origin story could be reconstructed by examining differences between protein-coding sequences. Computers became a tool through which molecular evolutionists could highlight and demonstrate the objectivity and statistical precision of molecular methods in their battle against traditional morphological approaches to evolution.

The second part examines the role of algorithms in the 1980s as the number of available protein and nucleotide sequences began to rapidly increase. Algorithms such as FASTA and BLAST, although ostensibly solving the same problem as their predecessors, in fact came to have an additional and different set of uses: that is, their main goal was no longer the reconstruction of evolutionary lineages through the comparison of proteins, but rather the management of sequence information. Such management was necessary for the production of knowledge about gene function as molecular geneticists sought to show how genes determined biological function. The textual and linguistic metaphors that dominated molecular biology in the 1950s and 60s, although still present, were overlaid by new notions of sequence-as-data that had to be banked, sorted, searched and organized in order to do biological work.

Part I: Origins

Historians and philosophers of biology have had much to say about the role of textual and cryptographic metaphors in mid twentieth century biology. Lily Kay has argued that "DNA was conceptualized as programmed information, as a Book of Life of the nascent information age" and tracked the ways in which information metaphors were "remarkably seductive and productive both operationally and culturally." (Kay 2000: 327-238). Similarly, Hans-Jörg Rheinberger sees both the past of molecular biology and the future of molecular medicine as driven by the basic fact that

biologists "have come to envision the fundamental processes of life as based on the storage, transmission, change, accumulation, and expression of genetic information." (Rheinberger 2000: 22). This literature has also explored the role of information and coding metaphors in opening biology up to computing.[6] Code-talk was productive for two reasons. First, on a linguistic level, when DNA became information, it became susceptible to information processing machines; the coding of life could be imagined to be like 'coding' software or programming a computer. Second – on a more practical level – for biologists who wished to emphasize the centrality and importance of strings of letters to biology, computers offered ready-made tools for rapid symbol manipulation.

In September 1964, a symposium was held at Rutgers University with the title "Evolving Genes and Proteins." Participants included Linus Pauling, Christian Anfinsen, Salvador Luria, Arthur Kornberg, Alexander Rich, Edward Tatum, and Tracy Sonneborn. Taking place as the genetic code was beginning to be decoded (and as significant numbers of protein sequences were becoming available) the symposium was an attempt to understand the molecular basis of evolution (Bryson and Vogel 1965). No doubt most of the participants would have agreed with Tatum's remarks in his opening address: "'The Age of DNA' is now in full flower of rapid, exponential development [and] is characterized by the gratifyingly successful attempts to delineate in detail the molecular structures, processes, and forces which underly [sic] the specificity, replication, and functioning of the hereditary material." (Bryson and Vogel 1965: 4). Evolution had to be understood not through studying fossils, or morphology, but by paying close attention to the relationships between homologous protein sequences across different species. Tatum predicted that new techniques would soon open up research on "controlled evolution, including genetic engineering, on the evolution of life on other planets and in

6 In fact Kay makes a point of the fact that computational and mathematical approaches to "cracking the genetic code" (for example, George Gamov's efforts) came to nothing because "the genetic code is not a code." Kay's argument is that information was productive at a discursive and metaphorical, rather than a practical, level. See also Strasser 2006 and November 2006.

other solar systems, and even on the origin of life itself." (Bryson and Vogel 1965: 9).[7]

Such a view was controversial. The idea that molecular biology would soon be able to tackle the most profound of questions, was not accepted by everyone. A group of biologists led by Ernst Mayr, George Gaylord Simpson, and Theodosius Dobzhansky argued that the molecular should not and could not be privileged over the morphological in evolutionary studies, defending organismic biology against the colonizing forces of the molcularists (Morgan 1998; Dietrich 1994; Dietrich 1998; Aronson 2002). During the conference, Pauling and Zuckerkandl demonstrated a powerful method whereby molecules could be used to deduce evolutionary relationships. Sequence comparison could tell an origin story. Largely through Pauling and Zuckerkandl's work, the term 'information' began to apply not just to the structure and function of macromolecules, but also to *history*.[8] In 1962, Zuckerkandl had coined the term 'molecular anthropology' to refer to the practice of reconstructing evolutionary relationships from information contained in nucleotides (Sommer 2008). Molecular anthropologists compared sets of protein sequences in order to construct phylogenetic trees; such a tree described a particular historical relationship between protein molecules and (by implication) a genealogy for the organisms which they had been extracted from. This methodology formed the justification for the majority of sequence comparison efforts until the 1980s.

However, Pauling and Zuckerkandl's bold vision of understanding all evolutionary history through sequence comparison was constrained by two factors. The first was that protein sequence comparison was limited to sequences where the differences between two sequences was visually obvious. Comparing highly conserved sequences (such as cytochromes) was trivial because the differences between them could be counted by eye. But for other sequences – where the best alignment between the two was less obvious – inferring the evolutionary trajectory of proteins involved a large

7 Linus Pauling and his coworkers had conducted experiments in 1949 linking Mendelian hereditary traits to molecular changes in protein structure (Pauling et al. 1949), however it was not yet proved that specific changes in DNA *sequences* were linked to changes in protein structure.

8 For the detailed context surrounding Pauling and Zuckerkandl's work see Suárez Díaz 2008a.

number of repetitive operations. For instance, a comparison of ten amino-acid sequence to a fifty amino-acid sequence requires checking forty potential alignments, each of which involves checking the identity of ten amino acids, generating a total of 400 steps.[9] Comparison of entire protein sequences was often done by writing out the sequences on long strips of paper and sliding them past one another to produce each alignment in turn; once the best alignment was found, the investigator then had to find the 'difference' between the two sequences by counting the number of matched and mis-matched amino-acids.

It is important to distinguish here between to distinct operations: sequence comparison and tree building. For a given set of related sequences, an evolutionary tree or phylogeny relating the sequences to one another can only be built once all the sequences have been compared to each other in pairs. A tree can then be assembled by joining the two most closely related sequences on one branch, and then adding the next most closely related and so on until all the sequences are included. Even in cases where alignments between sequences were obvious, reconstructing parsimonious trees from pairwise comparisons could prove extremely difficult because of the large number of possibilities for connecting the branches.[10] In the 1960s, molecular evolutionists built algorithms for both sequence alignment and tree-building; although this paper focuses on the sequence comparison step, often it was tree-building that was more computationally intensive (because only sequences with obvious alignments were chosen for comparison). Nevertheless, both were necessary for building evolutionary histories, and the methods were most often used in tandem.

Walter Fitch was one of the strongest advocates of Pauling and Zuckerkandl's molecular methods. Trained in comparative biochemistry at Berkeley, in the early 1960s Fitch began to work on

---

9 This is the case if there are assumed to be no gaps in either sequence. If gaps of arbitrary length are allowed, the problem becomes much worse.

10 For an unrooted tree with three nodes (sequences) there is only one possible tree, but for a tree with just ten nodes (sequences) there are over two million possibilities.

applying molecular biology to evolutionary problems.[11] In 1965, Fitch, working at the University of Wisconsin (Madison) Medical School, designed an algorithm for determining "evolutionary homology" between proteins (Fitch 1966). Allowing a computer to perform the amino-acid to amino-acid comparisons allowed Fitch to take into account the possibility of small gaps in the alignment. If entire sequences are compared at once, a small gap in the middle of one sequence would completely throw off the alignment. Instead of comparing entire sequences, then, Fitch used a sliding window to compare sub-sequences for homology. Testing his program on the α- and β-hemoglobins (which contained approximately 150 amino-acids each) required 13 104 sequence comparisons of thirty amino-acids each, or 393 120 letter-to-letter comparisons. Such a task would have been almost inconceivable without a computer. As such, Fitch's algorithmic instrument immediately enabled more sensitive and more realistic determinations of sequence homology and hence evolutionary distance between sequences. The computer, then, allowed advocates of molecular evolution to overcome one of the major hurdles towards implementing their program.

The computerization of molecular methods served to highlight its advantages over its morphological competitors. Edna Suárez has argued that protein sequences – in particular those collected and compared by Fitch and Emaneul Margoliash in 1967 (Fitch and Margoliash 1967) – "provided the material on which to apply explicit statistical criteria that the older taxonomists were not able to provide." (Suárez 2008b). Computers – in particular sequence comparison and tree-building algorithms – were tools with which to highlight and demonstrate the objectivity and precision of molecular methods through applying rigorous statistical methods to the construction of phylogenies. Comparison of sequences eliminated the 'judgment' involved in traditional taxonomic methods, providing instead a quantitative measure of difference. The automation of both the sequence alignment

11 His first paper on this subject appeared in 1964 (Fitch 1964). In the 1980s Fitch went on to become the co-founder on the journal *Molecular Biology and Evolution* and the first president of the Society for Molecular Biology and Evolution.

and tree-building steps reinforced the perceived objectivity of the molecular methods.[12]

The second factor constraining protein sequence comparison was a lack of knowledge of protein sequences themselves. Although the first protein sequence – that of insulin – had been determined in 1953 by Frederick Sanger, the sequencing of proteins remained sporadic into the 1960s; no efforts had been made to systematically collect sequence information. Margaret Dayhoff was the first person to realize the importance of accumulating protein sequences for biology. Dayhoff had studied quantum chemistry at Columbia under George Kimball, gaining her PhD degree in 1949. For her dissertation work she used punch-card fed computers designed for business operations at the Watson Computing Laboratories to calculate the molecular resonance energies of polycyclic organic molecules (Dayhoff and Kimball 1949; Hunt 1984; Hagen 2001; Strasser 2006). Dayhoff continued to apply electronic computers to biological problems and in the early 1960s she and her co-workers at the National Biomedical Research Foundation (NBRF) were working on the use of computer programs for assisting in protein sequence determination. The NBRF was an unusual context for biological work. Its founder, Robert S. Ledley, a qualified dentist, had been exposed to computers doing work in operations research in military contexts. After the war, as a member of the RNA Tie Club, he had pursued various ways of applying computers to biomedical problems. In particular, in the late 1950s Ledley developed a computer for assisting medical diagnosis, founding the NBRF in 1960 "in order to explore the possible uses of electronic computers in biomedical research." (November 2006: 165). It was in this highly interdisciplinary context that Dayhoff and her colleagues began to apply computers to protein problems. Although they were not yet applying computing to sequence comparison, Dayhoff's FORTRAN routines were designed to aid in planning experiments, detecting errors, determining the consistency of experimental data, and assessing the reliability of results (Dayhoff 1964). The aim was

12 Suárez 2008b quotes a paper by Thorne, Kishino, and Felsenstein from 1991: "It is possible, and among some researchers, popular to align sequences by eyeball. The eyeball technique is time-consuming, tedious, and irreproducible... Computer-aided sequence alignment does not possess the disadvantages of the eyeball technique." (Thorne et al. 1991: 114).

to speed up the experimental determination of protein sequence information in order understand the "evolutionary history of life." By 1965, Dayhoff had published the first edition of her "Atlas of Protein Sequence and Structure," a collection of all the known protein sequences in a common format. Although this first edition contained only about seventy sequences, the amount of sequence data grew rapidly: by the following year the Atlas contained over one hundred sequences (in addition to just three nucleotide sequences), and by 1968 the number had quickly climbed to 197 sequences (and six nucleotide sequences).[13]

The organization and curation of Dayhoff's sequence collection was highly computerized: sequences, names, citations, and comments were stored on punched cards and "alignments, the three-letter notation sequences, the amino-acid compositions, the page layouts and numbering, and the author and subject index entries from the data section are produced automatically by computer." (Dayhoff and Eck 1968: viii). Although the work of gathering sequences from various sources, checking consistency of the data, and transforming it into a uniform format were difficult problems in themselves, Dayhoff's real interest was not collection. Rather, as the introduction to her *Atlas* makes clear, the aim was to use the sequences to make a contribution to "the theory of the evolution of proteins and nucleic acid and to the mathematical treatment of the data." (Dayhoff and Eck 1968: viii).[14] She understood her work as a necessary labor for further progress in studies of molecular evolution. As such, Dayhoff wrote computer programs that produced phylogenetic trees, detected chromosomal duplications, simulated protein evolution, and generated multi-species alignments. These last were printed on long strips of paper that folded out from the back of the *Atlas*. Although some three-dimensional crystal structures of proteins were also reproduced in the *Atlas,* the emphasis was on accumulating the linear strings of

13 For a detailed account of Dayhoff's work see Strasser 2006.

14 A detailed explication of what Dayhoff was trying to achieve through her sequence collection efforts can be found in: Dayhoff (1969). For instance: "The comparative study of proteins... provides an approach to critical issues in biology: the exact relation and order of origin of the major groups of organisms, the evolution of the genetic and metabolic complexity of present-day organisms and the nature of of biochemical processes... Because of our interest in the theoretical aspects of protein structure our group at the National Biomedical Research Foundation has long maintained a collection of known sequences" (p. 87).

sequence information in order to learn about evolutionary history through sequence comparison.

Although the methods of Fitch and Dayhoff were effective for the comparison of short sequences of interest, significant refinements in the apparatus were required if it was to be able to tackle full-length protein sequences. In 1967, Christian Wunsch was pursuing both his MD and a PhD in biochemistry at Northwestern University, studying heterocyclic analogues of serotonin. He was hoping to correlate their kinetic constants with molecular orbital calculations from quantum chemistry. During his graduate work Wunsch had become "enamored with computers and interested in all kinds of problems computers were being used to solve, developed into a good programmer, and even did some contract programming to help pay the bills." (Wunsch, personal correspondence).[15] Northwestern had obtained its first computer – a CDC 3000 – in 1965 and Wunsch had taught himself to program in FORTRAN and begun to solve problems using techniques of successive approximation. At a meeting of the Biochemistry Journal Club, Saul Needleman, a faculty member in the Biochemistry Department, presented a paper by Margoliash[16] that used an algorithm to determine the similarity between two amino-acid sequences.

> After the meeting I told him that I thought a better algorithm would be to use an exhaustive search method over short, overlapping sequence domains, then link the best results together - it was a method that would allow naturally for evolutionary insertions and deletions in the compared sequences - something missing in earlier methods. Needleman, who did not program, offered to give me some money to purchase computer time to try out my idea. With about $200 I opened a computer account and began working on the problem without much success. (Wunsch, personal correspondence).

From his work in quantum chemistry Wunsch would have been familiar with the use of matrices and he began to cast the sequence matching problem in matrix terms. It was soon obvious that an

15 The narrative that follows is based on the same. I rely here exclusively on Wunsch's retrospective account. However, there appear to be no other sources, published or archival, which shed light on the origins of the Needleman Wunsch algorithm.

16 It is likely that this was Fitch and Margoliash 1967. Needleman and Margoliash had worked together previously on determining the sequence of cytochrome c in rabbits (Needleman and Margoliash 1966).

exhaustive search would be impossible for long sequences and Wunsch set about determining the practical limits on such a search. In the course of trying to eliminate redundant comparisons from his counting, Wunsch realized that "by recording the number of possibilities in each cell of the next-to-last row, one did not need to count them again for any path that proceeded from a cell in an earlier row. Indeed, by making the method iterative, row-by-row, one could tally the number of paths that could follow from any given cell..." (Wunsch, personal correspondence). It was an attempt to prove the *impossibility* of computing the full sequence match that provided the solution to the problem. Although Wunsch's account of his own 'eureka' moment is perhaps exaggerated, it is useful in illustrating how the development of sequence comparison required the importing of techniques and methods well beyond the purview and expertise of most biologists. Wunsch – neither a biologists nor a computer expert – was able to make a fundamental contribution in this fluid disciplinary space. Needleman and Wunsch algorithm was published in 1970 (Needleman and Wunsch 1970).[17] The basic idea is that some computations require many small, identical operations to be repeated over-and-over again; a large computational saving can be made by storing the results of such computations and re-using them. When gaps are inserted into a sequence, the same stretches of sequence must be compared multiple times – Needleman and Wunch's algorithm stores the results of these comparisons in a matrix such that such duplication of effort is minimized. Once all the comparisons have been performed, the best overall alignment can be determined by tracing a pathway through the matrix.[18] The innovations that Wunsch introduced depended not only on his background in quantum chemistry, but also on his fascination with

17 In fact, the algorithm was first presented as a paper in 1967 under the title "A method for finding similarities in amino acid sequence of two proteins" at the 154th meeting of the American Chemical Society. According to Wunsch, he then revised and submitted the paper to the Journal of Molecular Biology, placing himself as first author. The paper was at first rejected and Wunsch delayed resubmission because he was completing his medical degree; meanwhile, and without consulting Wunsch, Needleman resubmitted the paper where it was published with Needleman as first author. Wunsch excluded Needleman from his dissertation committee and the pair never spoke again (Christian Wunsch, personal correspondence).

18 The iterative approach applied in Needleman-Wunsch later began to be associated with the techniques of 'dynamic programming,' a set of techniques invented by Richard Bellman in the 1940s for unrelated purposes. Needleman and Wunsch were unaware of this work. On the invention of dynamic programming see Dreyfus 2002 and Bellman 1984.

computers which suggested to him the power of an iterative approach to the problem.

The application of iterative methods meant that large-scale sequence comparison became a viable proposition. Computing power, for a while at least, could keep up with the growth in sequence information. It was proved mathematically that Needleman and Wunch's method was guaranteed to yield the best alignment between two sequences. Because of this is could be used to "detect and define" the homology between sequences, and thus to measure "evolutionary distance" (Needleman and Wunsch 1970: 452). During the 1970s, this notion of 'distance' became the most important way of thinking about sequence comparison. Protein – and later nucleotide – sequences became "living fossils" whose text could narrate a story about the past. All that was needed to discover this story was the right distance metric which would place species, varieties, and individuals in the right order.

By 1972, Stanislaw Ulam – famous for his work on the hydrogen bomb – had turned some of his attention to mathematical problems in molecular biology. In that year he published a paper framing the sequence homology problem as one of defining a distance or metric space for sequences (Ulam 1972).[19] This distance was defined as the minimal mutational path by which one sequence could turn into another sequence, either through insertions or deletions or through point mutations. Work on biology at Los Alamos took place within the Theoretical Biology and Biophysics Division (T-10). In the late 1960s, George I. Bell, a student of Hans Bethe, began to work seriously on biological problems. Bell's work focused on immunology and in 1970 he published a paper providing an explicit quantitative model of the immune system which could be explored computationally. Bell began T-10 in 1974 with the aim of developing theoretical approaches to biology that would complement the mostly experimental approaches pursued elsewhere. Bell was quickly joined by Walter Goad. Goad, another theoretical physicist who had come to Los Alamos in 1950 to work on the hydrogen bomb, had spent the 1970-71 academic year on sabbatical working with Francis Crick at the Medical Research Council

19 Ulam was not the first to use the 'distance' concept for sequence comparison, but he was the first to formalize it in a mathematically precise sense of a metric. On Ulam's contribution see Goad 1987.

Laboratory of Molecular Biology in Cambridge. After this visit he turned his full attention to theoretical biology. For a group working on biological problems, T-10 had a most unusual set of knowledge and skills in mathematics, physics, and computers, drawn from their bomb work. Further work at Los Alamos elaborated Ulam's ideas to give a precise algorithm for reconstructing phylogenetic trees from protein sequence data (Beyer et al. 1974). Soon afterward, Peter Sellers at Rockefeller University proved that a sensible definition of sequence distance could be found which satisfied the triangle inequality – the most important mathematical principle for demonstrating that a measure satisfies the formal mathematical criteria of a distance (Sellers 1974). By formulating the concept of distance between sequences in a formal, mathematical sense, Ulam and his co-workers hoped to produce a precise ordering of sequences that told a story of evolutionary origins, an amino-acid-based history.

All these attempts to subject sequences to computerized matching algorithms derived their plausibility in part from the information discourse of molecular biology. Kay uses Pauling as an example of one individual who, in the 1940s, used pre-informatic metaphors of 'pattern and template' (Kay 2000: 49-51). However, his notion of the molecule as 'semantide' at the Rutgers conference suggests his commitment, by the early 1960s, to the informatic paradigm. Others at the Rutgers conference, including Alexander Rich and Noboru Sueoka spoke easily of DNA, RNA, and protein as "information" and "codes" (Bryson and Vogel 1965: 453-459 (Rich), 479-485 (Sueoka). For Dayhoff, proteins were molecules that not only carried information about structure, but also "they contain information about their own origin and history and about the ancestry and evolution of the organisms in which they are found." (Dayhoff 1969: 87). Not all such computational work drew on informatic metaphors: during the 1960s and 1970s taxonomists also developed algorithms to systematically compare sets of morphological characteristics, yet their work did not adopt the informatic perspective (Hagen 2001). But informatic metaphors applied most readily and powerfully to *molecules* (that is,

RNA, DNA, and protein) – in fact such molecules were defined as sequences or codes[20]; and it was here, in the encoding of molecular sequences into computer codes, that the informatic metaphor persisted. Wunsch's training in quantum chemistry and programming allowed him to develop a procedure that immediately cast the molecule into an informatic form – a matrix – susceptible to computational methods. Working alongside George Gamow, John von Neumann, Martynas Yčas, and Nicholas Metropolis, Ulam was deeply involved in the cybernetic vision of biology; in the mid-1950s he contributed to work on the distribution of nucleotides in RNA (Kay 2000: 156-159). His mathematical contribution to the computational problems of sequence matching in the 1970s were an extension of this program – an attempt to show how one could gain biological insight by rendering sequences into mathematical codes. Indeed, a large part of his paper on 'biomathematics' is devoted to speculations about the kind and quantity of information contained in the genetic code (Ulam 1972).[21] Since together all codes formed a text, a "Book of Life," they could be made to tell a story about evolutionary history. By comparing sequences, putting them in precise order, reconstructing them into trees, this origin story became manifest. Sequence comparison became an important problem not only because it was computationally tractable, but also because it made sense of the code of life in a way that was interwoven with the dominant discourse of molecular biology in the 1960s and 70s. Sequence comparison made information into narrative.

The development of sequence comparison algorithms as instruments of biological practice was tied to both (sub-)disciplinary competition and notions of what a sequence meant. Such instruments were important in transforming evolutionary studies on the molecular level into a set of practical and routine operations. This contributed to the ability of molecular studies to legitimate themselves with

20 In 1957 Crick argued that "any genetic information in the nucleic acid is carried by the base sequence, and only by this sequence." (Crick 1957: 175).

21 For instance, Ulam becomes concerned with the problem of how much information is contained in the genetic code, comparing it to a mathematical encoding of prime numbers to demonstrate how information might be compressed through "inductive or recursive rules." pp. 289-290.

respect to traditional morphological and organismic approaches to evolutionary questions. In doing so, however, sequence comparison algorithms tied sequences more tightly to one particular use: that is, to the construction of phylogenies or evolutionary histories. That is, sequence comparison algorithms transformed information molecules into stories about evolutionary history. Here disciplinary, instrumental, and epistemological transformations occurred together. Molecular evolution, sequence comparison algorithms, and sequences-as-stories arose through a kind of co-production: particular objects (sequences) are produced both by instruments (algorithms) and by disciplinary imperatives (molecular evolution). Likewise, the development of sequence comparison techniques and molecular evolutionary approaches were reinforced by each other and also by the particular forms of objects they produced.[22]

Part II: Searching

In 1976, Allan Maxam and Walter Gilbert developed a method of reliably determining the sequence of short DNA fragments using radioactive labels (Maxam and Gilbert 1977). Less than a year later, Frederick Sanger developed an even more efficient (and less dangerous) sequencing method: the dideoxy (or chain-termination) was quickly adopted by most people interested in DNA sequences (Sanger et al. 1977).[23] From the late 1970s onward, DNA sequences gradually became preferred over protein sequences in evolutionary studies. Because of the degeneracy of the genetic code many mutations are 'invisible' at the protein level; as such, DNA was considered a more direct representation of mutational events than protein. The new sequencing techniques provided access to a more fundamental representation of evolutionary events. The rise of fast, reliable, and scalable DNA

22 On co-production see Jasanoff 2006, introduction. Jasanoff uses the term to show how objects are produced both culturally and naturally in a way that is inextricable at all levels. Here co-production indicates how instruments, disciplines, and their objects of study might be similarly produced together.

23 It was less dangerous because it could be performed using lesser amounts of radiation and toxic chemicals.

sequencing methods marked a change not only from protein to DNA, but also an increase in the number of sequences available. But it also caused a qualitative shift in the kind of sequence comparison that would be required. Since both protein and nucleic acid sequences were just strings of letters (twenty for proteins, four for DNA or RNA), from a purely formal point of view the analysis of each would be exactly the same. In practice, however, there were several differences which required the development of new methods of sequence comparison. The most important of these was that a given stretch of DNA did not necessarily code for protein – it could be an untranslated region, a promoter, or a chunk of uncharacterized DNA between gene regions. Indeed, the region of interest – the protein coding section – could be buried in the middle of a long DNA sequence.

In addition to this technological change, however, at the end of the 1970s the disciplinary agenda of molecular biology had begun to shift. Inspired by the new molecular techniques that allowed the copying and editing of DNA, some evolutionary biologists turned their attention to the problem of showing how the evolution (on the molecular-genetic level) could account for diseases and physical traits. Those interested in studying sequence no longer had to confine themselves to problems of molecular evolution – what can we learn about evolution from protein and DNA sequence? - but now shifted to problems of molecular genetics – showing that DNA sequence was the basis of complex physical and behavioral traits in humans.[24] This tendency was reinforced by the opportunities that genetic engineering seemed to provide for a molecule-based medicine: discovering the gene for cystic fibrosis or breast cancer was, many believed, just a small step from finding a cure.[25] The ability to gain

---

24 On this shift see Keller (1992): "After 1970, both the development of techniques permitting direct intervention in the structure of DNA sequence and the use of these techniques in the study of human genetics took off exponentially." (p. 291). In addition, the molecular evolutionists became more and more established. Although organismic and morphological approaches did not completely disappear, by 1980 the molecular evolutionists had largely convinced their competitors of the strength of their approach: "Although some taxonomist can still ignore molecular evidence, in many cases both classification and phylogenetic reconstruction have been significantly influenced by molecular biology." Hagen 1999, p. 340.

25 Robert Cook-Deegan, in his history of the HGP argues that: "During the 1970s and 1980s, genetics was drawn out of the backwater and entered the mainstream of biomedical research, emerging as the dominant strategy to understand mysterious diseases" (p. 10). Cook-Deegan connects this rise to the development of RFLP (restriction fragment length

access to the DNA sequence directly and on a large scale caused a qualitative shift in the kinds of questions that could be asked: in particular, molecular biologists could begin to attack the problem of how DNA sequence determined biological function. This genetic determinist program dominated biology in the 1980s and 1990s as the epistemic agenda shifted toward an accumulation of knowledge of gene function.[26]

These technological and disciplinary shifts meant that new uses could be envisioned for the instruments of sequence comparison. In particular, biologists needed to find where genes resided within long sequences of DNA. The solution to this problem was to make a distinction between global and local similarity. Algorithms such as Needleman-Wunch were able only to give the best global alignment between two sequences – that is, they always took the matching or mismatching of every base into account in calculating the similarity score. Around 1980, Sellers and Walter Goad realized that one could instead define a 'local' similarity. This was to ask a very different question: not 'How similar are these two sequences?' but rather 'Which parts of these long sequences look most similar?' Global comparisons continued to be used and biologists continued to be interested in using these algorithms to answer evolutionary questions, but local alignments opened up new problems for computational sequence analysis.

In the summer of 1974, Temple Smith, a young nuclear physicist, and Mike Waterman, a young mathematician had joined Ulam and Bill Beyer at Los Alamos to work on problems of molecular biology and evolution. Smith and Waterman were both teaching at small and intellectually isolating universities in the midwest (Smith in Michigan and Waterman in Idaho) and relished the opportunity to

---

polymorphism) maps in the late 1970s which “not only made gene-hunting easier but also opened entirely new possibilities for for tracing the inheritance of multiple genes...” (p. 46). In the early 1980s RFLP maps located genes for Huntington's disease and Duchenne muscular dystrophy. These, and the discovery of the gene for cystic fibrosis (mapped in 1985 and identified in 1989), demonstrated the power of genetic approaches to medicine (Cook-Deegan 1994, pp. 44-45). It was also around this time that the first DNA tests were being developed for use in forensics (see Jeffreys et al 1985). Also see Yoxen (1982), who argues that the concept of 'genetic disease' was constructed during the 1970s.

26 On the role of genetic determinism in biology see Keller 1992. On the extent to which this vision also dominated popular culture see Nelkin and Lindee 1996.

spend their summers working on novel research at a world-famous laboratory (Waterman 1999). Smith's background in physics had included computational work in the analysis of cross-section data from nuclear physics experiments (Temple Smith interview, 12/2/2007). The Los Alamos collaboration, which lasted until the end of the decade, added mathematical and computational rigor to sequence alignment algorithms.[27] In 1980, however, Smith and Waterman realized that a change in mismatch scoring would result in a remarkably different result.[28] They showed that Needleman and Wunch's matrix algorithm could be modified to determine local rather than global similarity (Smith and Waterman 1981). Their formulation guaranteed the return of "a pair of segments, one from each of two long sequences, such that there is no other pair of segments with greater similarity (homology)." (Smith and Waterman 1981: 195).

The Smith-Waterman algorithm marked a break with the first phase of sequence comparison. It removed the necessity for the alignment to be centered on whole genes or proteins – any fragment of sequence could be compared to any other fragment. Since now it was possible to search for fragments of similarity one did not need to assume any structured relationship between the sequences. Sequence comparison could now be used not only for reconstructing evolutionary relationships, but for totally new sorts of analyses. The most well known example of this is the work of Russell Doolittle, often recited as a sort of folkloric tale in bioinformatics. Indeed Doolittle's story is important not because it necessarily typified the way in which sequences came to be used, but rather because its reception and the importance that was attributed to it by other biologists suggests how sequence comparison algorithms were being re-imagined as new kinds of powerful tools.

Doolittle was trained as a biochemist at Harvard and had been working on problems of

27 For instance Beyer et al. 1974, Waterman et al. 1976, Waterman et al. 1977, Smith et al. 1981.

28 When an alignment results in the superposition of two non-identical nucleotides a negative mismatch or penalty score is applied. The innovation of Smith-Waterman was to not let the overall match score drop below zero. Others, at Los Alamos and elsewhere, had previously used similar techniques to find locally strong matches: Dayhoff, for instance, had a program called 'search' that used fixed overlapping protein sequence regions to search a database. Peter Sellers (1979) was the first to clearly define local similarity and Walter Goad and Minoru Kanehisa (1982) were the first to implement a useful program for finding local alignments (David Lipman, personal communication, November 11th 2008).

molecular evolution and protein sequence alignment at University of California San Diego since the 1960s. In 1981, Doolittle published an article in *Science* with the title "Similar amino-acid sequences: chance or common ancestry?" (Doolittle 1981). His aim was to use sequence comparison not to reconstruct the hierarchy of species, but to learn more about protein function: since evolution tends to preserve function, much might be learned about how proteins work through sequence comparison. To achieve this Doolittle compiled his own database of proteins, which he called *Newat* (for "new atlas"), that built on Dayhoff's *Atlas* but was more up-to-date, more representative, and less redundant.[29] It was searching *Newat* on his VAX computer that Doolittle discovered an unusually high degree of similarity between a gene in Simian Sarcoma Virus (a cancer-causing virus) and a human growth factor gene. As the subsequent publication noted, "This relationship raises intriguing possibilities about the mechanism of action at the molecular level of both the *onc* gene product and the growth factor." (Doolittle et al. 1983: 276).

Although Doolittle was not the first to investigate protein function by sequence comparison, this story is often narrated by biologists and in bioinformatics textbooks as a 'eureka' moment for demonstrating how computers in general and sequence comparison in particular could be valuable to biology.[30] Such computational approaches lay decidedly outside the mainstream of biological practice. Many practitioners came from other fields such as physics (Temple Smith) or statistics (Mike Waterman) and were regarded as performing a kind of 'theoretical' work that was not highly valued amongst biologists.[31] Doolittle's finding, reported in the *New York Times*, was a boon for those who wished to promote such extra-experimental practices – it demonstrated a route through which

29 As the story goes, Doolittle had his secretary and his 11-year-old son manually entering sequence information into his computer. Doolittle 2000.

30 For example, one textbook narrates the story thus: "In 1983 [Doolittle] stunned cancer biologists when he reported that a newly reported sequence for platelet-derived growth factor (PDGF) was virtually identical to a previously reported sequence for the oncogene known as v-sis." This was big news and the finding served as a wake-up call for molecular biologists: "searching all new sequences against an up to date database is your first order of business." (Jones and Pevzner 2004: 79). Doolittle 1997 and 2000 describe similar work that had been carried out before his own.

31 This was particular true of Dayhoff's sequence collection work, which was often written off as mere theory (Strasser 2006: 117).

computers could produce meaningful biological results (Schmeck 1983). In particular, it showed a way out of one major dilemma posed by DNA sequencing. Since proteins were often isolated from particular parts of an organism or cell, their function was often known (or could be inferred from its physical or chemical properties); DNA sequence, on the other hand, was just sequence and as such its function (if it had one at all) was invariably unknown (Strasser 2006). Explaining how DNA controlled the biology of organisms required methods for determining what particular strings of DNA did. Doolittle's work showed how sequence comparison could be used to achieve this by demonstrating similarities between sequences of known and unknown function.[32]

Doolittle's work demonstrated that sequence comparison could be useful not just for reconstructing evolutionary relationships and telling stories about the past, but for understanding biological function in a way that was abstracted from genes, species, and evolutionary hierarchies. Work using sequence comparison to construct phylogenies did not stop – Doolittle himself spent much time on the problems of multiple sequence alignment and tree reconstruction (Doolittle 2000: 31). However, Doolittle showed how the instruments of sequence comparison could be used to further the aims of the genetic determinist-reductionist research program. As such, sequence comparison began to flourish in different roles.

Partly in response to the new possibilities for sequence comparison, from the early 1980s renewed efforts were made at collecting nucleic acid sequence in a centralized repository. Apart from Doolittle, the leaders in this field were Margaret Dayhoff at the NBRF and Walter Goad at Los Alamos National Laboratories. The early history of GenBank, as the preeminent repository came to be called, has been detailed elsewhere (Strasser 2006). However, the relationship between sequence comparison algorithms and the development of databases has not been examined.

One of the key figures was David Lipman. After finishing medical school, Lipman began to get

32 This was based on the assumptions that the one-dimensional sequence largely or fully determined the three dimensional structure of the associated protein and that similar structure implied similar function.

interested in mathematical problems in biology, and specifically in the problem of how the influenza virus evades the immune response. This prompted him to apply to the National Institutes of Health to become a medical staff fellow, where he hoped to pursue work in computational biology. Although Lipman had a hard time finding someone who would encourage the kind of work he wanted to do, he ended up at the Mathematical Research Branch of the National Institute of Arthritis, Diabetes, Digestive, and Kidney Diseases. Here he met John Wilbur, who was working on modeling in neuroscience. Lipman convinced Wilbur to spend some time working on problems in molecular evolution. In doing this Wilbur and Lipman were using both Dayhoff's database at the NBRF and Goad's database at Los Alamos over dial-up modem connections (David Lipman interview, 4/9/2008). They soon saw an opportunity to deploy their mathematical and computational skills: as databases grew larger, searching at speed was necessary in order to make searching an entire database practical. It is worth quoting extensively from the abstract of their paper which outlined a solution to this problem:

> With the development of large data banks of protein and nucleic acid sequences, the need for efficient methods of searching such banks for sequences similar to a given sequence has become evident... Currently, using the DEC KL-10 system we can compare all sequences in the entire Protein Data Bank of the National Biomedical Research Foundation with the 350-residue query sequence in less than 3 min and carry out a 500-base query sequence against all eukaryotic sequences in the Los Alamos Nucleic Acid Data Base in less than 2 min. (Wilbur and Lipman 1983: 726).

What is significant about this is not so much the specific times reported, but the fact that times were reported at all – other popular methods up to this time did not report search speeds, since they did not anticipate searches against entire databases.[33] Although the extent to which Wilbur and Lipman's algorithm represented a speed-up over Needleman-Wunsch and Smith-Waterman is not clear, it did put forward a new criteria by which sequence comparison algorithms should be assessed, namely their

33 For example, the publication describing the popular Korn-Queen program made no comment on the speed of the algorithm (Korn et al. 1977).

speed at searching large databases.[34]

Lipman had assumed that since most of the new sequence data was DNA, that a successful search algorithm should focus on DNA sequence matching. As such, the Wilbur-Lipman algorithm used a particular matrix form (a unitary matrix) that only allowed for simple matches, mismatches or gaps. After reading a paper by Dayhoff, Lipman realized that searching protein similarities in a way that took into account the relatedness of specific amino-acid pairs would provide a much more powerful way to detect more distant relationships between sequences. A protein sequence could change substantially – while retaining its essential function – by swapping one amino-acid for a similar one (for example, a mutation that caused one hydrophobic amino-acid to be replaced by another); over time, proteins might accumulate many such like-for-like mutations, making it appear superficially very different from its ancestor. Lipman's method – by 'matching' similar as well as identical amino-acids – could align such distant cousins (Lipman, personal communication).[35]

Lipman collaborated with Bill Pearson to write a new algorithm that could detect these more distant relationships while maintaining speed. This eventually became known as FASTA (pronounced *fast-ay*) (Lipman and Pearson 1985). It is based on the notion that matching sequence regions are likely to contain matching small 'words' or *k*-tuples. The algorithm first uses a hash or lookup table to identify occurrences of such words and then identifies the regions which contain the highest density of such matching words. A high-scoring alignment can be found in a time proportional to the length of the sum of the sequences (that is, linearly). Lipman and Pearson showed that under certain parameter choices their algorithm was fully equivalent to Needleman-Wunsch. However, their approach was to sacrifice accuracy in favor of time. FASTA is what is known as a heuristic algorithm – although it is likely to

34 Needleman-Wunsch was also a significant speed-up over older algorithms such as those developed by Fitch and Dayhoff. In that case, however the algorithm was designed essentially to solve the same problems of protein sequence alignment; Wilbur and Lipman's algorithm, on the other hand, was a necessary response to a new set of problems associated with large sets of sequences.

35 The Dayhoff paper was Barker and Dayhoff 1982.

produce the best match between any two sequences, unlike Needleman-Wunsch, it is not guaranteed to do so. No longer was the emphasis on finding an exact 'distance' or hierarchy of sequences; rather, FASTA was designed to perform a search across many sequences, rapidly returning any matches of interest, which could then be subjected to further analysis.

Before Wilbur and Lipman had even published their work, Michael Waterfield's lab used it to link a viral oncogene to a human platelet-derived growth factor, publishing the results just days before Doolittle's identical finding (Harding 2005; Waterfield et al. 1983).[36] Michael Fortun has argued that the Human Genome Project was characterized by a culture of speed – a kind of biology that was distinctive not because it was essentially different but because it was *faster* (Fortun 1999). Although it is extremely unlikely that it was the speed of Wilbur and Lipman's algorithm that was decisive in the 'race' for this result, this cancer gene finding suggested that speed would be newly important in sequence comparison. This exciting result, reported in the *New York Times* and the *Washington Post*, hinted at a gold-mine of discoveries waiting to be made if only it was possible to dig through the sequence databases fast enough (Schmeck 1983; Hilts 1983). As these databases grew in size, speed became even more important – to ask questions about gene function, such searches had to be able to be performed rapidly. Because FASTA was fast, it allowed particular problems (those of inferring gene function by sequence comparison) to remain practical. Moreover, it meant that accumulation of sequences in data banks remained useful practice – as long as the speed of algorithms could keep up with the amount of data, growth could continue. This was not speed for speeds sake, but rather these algorithms allowed biologists (helped by computer scientists and mathematicians) to answer questions that would not otherwise have been conceivable.

Sequence comparison is here a tool for dealing with the proliferation of sequences in the expanding data banks – information management was the primary goal. Indeed, Wilbur and Lipman

36 Waterfield's publication in Nature on July 7 came eight days ahead of that by Doolittle and Mike Hunkapiller in Science on July 15 (Doolittle et al. 1983).

were aware that the question which sequence comparison was trying to answer had shifted: "it may be fairly asked whether the more optimal alignment of a few relatively isolated sequence elements (not parts of *k*-tuple matches) that can be obtained by the full Needleman-Wunsch alignment over our method really gives a more accurate picture of biological truth. To this question, we do not know the answer." (Wilbur and Lipman 1983: 730). There is no mention here of evolution, trees, or the hierarchical relationship of sequences: the kind of biological truths that were being sought through sequence comparison had changed.

The first steps towards improving algorithms for sequence comparison in large databases were taken largely by a group of individuals in and around Los Alamos (Goad, Temple Smith, Waterman, Minoru Kanehisa). From 1982, Los Alamos was also the place at which the main sequence repository, GenBank, was being developed and managed. The building of sequence data banks and the building of tools for sequence analysis were at first considered to be separate activities: in 1981 the NIH contemplated two separate Requests for Proposals for the two tasks. The contract for the data bank was awarded to Bolt Beranek and Newman (partnered with Los Alamos) in 1982, while the second was never offered. However, the following year the NIH did award a contract to a company founded a group of Stanford computer scientists called IntelliGenetics to build and maintain a set of computer resources for biology called BIONET. At first, then, banking and tool-building remained separate. By the second half of the 1980s, however the two activities were drawing closer together, conceptually and practically. By 1987, when it came time for the five-year GenBank contract to be renewed, the NIH awarded the new contract to IntelliGenetics (again partnered with Los Alamos), bringing tool development and banking under one roof.[37]

The National Center for Biotechnology Information (NCBI) was created by Congress in 1987 to serve the growing informational and information technology needs of the biomedical research

37 The more recent history of GenBank has not been narrated in detail, but for an overview see Strasser 2008.

community. David Lipman, appointed director of the new Center, envisioned an institution which worked on both building improving databases and the tools for using them. Lipman was dissatisfied with the way in which GenBank was being run – in particular, along with many others, he thought that the structure of the database required fundamental revisions (David Lipman interview, 4/9/08). By building up NCBI as a key center of research on both algorithms and information management, Lipman could make a compelling argument for the relocation of GenBank to the Center itself.

As sequence databases continued to grow in size, a sequence comparison search would return many matches. The problem was that for a very large database, one would expect to find some medium-sized sequence strings purely by chance.[38] Therefore, the important question was determining the "unlikelihood" of a particular hit – the more unlikely it was to occur by chance, the more weight could be attached to it as a biological finding. Lipman wanted an algorithm that ignored less-significant matches while high-scoring (that is, very unlikely) matches could be found very fast (Stephen Altschul interview, 4/11/08). The result, published in 1990, was BLAST (the Basic Local Alignment Search Tool). The new algorithm was specifically oriented towards the searching of large databases. "The discovery of sequence homology to a known protein or family of proteins often provides the first clues about the function of a newly sequenced gene," the authors began. "As the DNA and amino acid sequence databases continue to grow in size they become increasingly useful in the analysis of newly sequenced genes and proteins because of the greater chance of finding such homologies." (Altschul et al. 1990: 403). Like FASTA, BLAST begins by finding all the instances of 'words' of fixed length within the query sequence. A deterministic finite automaton is then used to search for these words in every position in each database sequence.[39] When a 'seed' match is found, the algorithm attempts to

---

38 For instance, if the human genome was a random distribution of 3 billion nucleotides, we would expect any and every 15-letter combination of nucleotides to occur at least once by chance since 4^{15} is less than 3 billion. 4^{20}, however is significantly more than 3 billion, so if we found a match to a 20-letter sequence we might attach some significance to this finding.

39 A deterministic finite automaton is a concept that comes from the theory of computation and is used to model the behavior of a simple computing machine. Here is it used to efficiently and systematically search for short words in long

extend the alignment outwards from both ends of the word in an attempt to find an alignment of sufficiently high score to be surprising. By limiting these extensions to relatively rare seeds, BLAST achieved as great as a ten-fold reduction in search time over FASTA and other algorithms available at the time of its introduction.

The BLAST algorithm was fully oriented towards a new biology which was driven by rapid sequencing and relied on database searching for biological insight. The authors note the utility of their method for comparing cDNAs with partially sequenced genes, and for identifying similar regions of distantly related proteins (Altschul et al. 1990: 404). Within this biology, "the biological significance of high scoring matches may be inferred almost solely on the basis of the similarity score, and while the biological context of the borderline sequences may be helpful in distinguishing biologically interesting relationships." (Altschul et al. 1990: 404). The "biological significance" of these matches was not in reconstructing evolutionary relationships, but rather in being able to efficiently use sequence databases to make inferences about the function of a sequence. Lipman describes BLAST as a "gambling game" for finding sequence similarity:

> [T]he key thing with BLAST was to make finding the similarity a gambling game and this was especially good if one had a fairly accurate way of setting the odds... I managed to convince Sam Karlin [a mathematician] from Stanford to work on this and he solved it pretty quick. So rather than just having an heuristic method that probably would find significant matches, we now had the basis to do accurate gambling - we could play off speed & sensitivity quite accurately and determine what were our chances for missing a match of a given significance (Lipman, personal communication).

In other words, for a given BLAST search, one knew the "odds" that one was missing something potentially important. The notion of "significance" (statistical and biological) was the original motivation for BLAST and as such was built into the algorithm itself – it only looked for matches that were certain to be improbable enough that they could be used to say something definite about

---

stretches of sequence.

biological similarity. The success of this statistical approach to sequence comparison put the NCBI on the map as a important locus of bioinformatics research. Combined with NCBI's work on database standards, BLAST demonstrated that the Center was developing computationally and biologically sophisticated solutions to problems of data management. Responsibility for GenBank was transferred to the NCBI in 1992.

In this second phase, sequence comparison algorithms became tools of information management. Although sequence comparison still relied on the fact that sequences might be similar because they were related to one another through evolution, these algorithms were now not only used to construct a tree of life, but also for determining the function of DNA sequence. Beginning in the 1960s, Dayhoff and her coworkers had used sequence comparison for the organization of her *Atlas* of proteins into super-families (Strasser 2006: 112). During the 1980s, however, the ability to sequence DNA on an increasingly large scale led to a disciplinary imperative toward demonstrating the molecular (sequence) basis of all biological function. This resulted in the accumulation of vast amounts of sequence data (mostly at GenBank and its sister databases in the UK (EMBL-Bank) and Japan (DDBJ)) that could only be useful if it could be efficiently searched and compared. Sequence comparison tools such as BLAST made this accumulation of bioinformation possible by offering the possibility that the data could be used to make inference about biological function simply through sequence comparison.

The genetic determinist program, sequence databases, and sequence comparison algorithms mutually justified one another's existence. The program required both the accumulation of DNA sequence and the ability to determine its function; without the databases fast methods of searching would have no purpose, and without the algorithms the repositories would be unsearchable and hence useless. Again we have a three-way inter-dependency between disciplinary goals, instruments, and the knowledge that was created by them. As sequence comparison algorithms became instruments for

information management, sequences themselves became data to be organized, categorized, searched, accessed, deposited, and retrieved. Treating sequences in this way offered a way to discover gene function and support the genetic determinist program. A particular understanding of how biology works led to a particular conception of what kind of an object a sequence could be (and what sort of knowledge it held) and a re-making of the instruments used to interrogate them.

Conclusions:

As sequence comparison algorithms increasingly come to define the way we think (and know) about biology, it is important to reflect on their epistemological status. Describing these algorithms as a set of scientific instruments helps to shed light on the kinds of roles they have played – and are playing – in relation to genes, sequences, and genomes. These instruments have made these biological objects visible and knowable. In the same way that successfully viewing a distant galaxy through a telescope requires not only on the laws of optics, but also techniques of lens-grinding and so on, comparison of sequences requires not only on the logic of the computing machine, but also techniques of programming and using it reliably. And just as the laws of optics and the shape of the lens make a difference to what will be seen through the telescope, the hardware and the software used for sequence comparison makes a difference to how we 'see' sequences. Like telescopes and other instruments, sequence comparison algorithms are often become invisible or taken-for-granted tools. However – as specially designed instruments – disciplinary commitments, technical possibilities, and epistemic categories get built into their design. It is these commitments, possibilities, and categories that have, through the instruments, shaped biologists' understanding of sequences themselves.

This paper has narrated a history of sequence comparison algorithms in two phases. It has shown how the central objects of biology (genes, sequences, genomes) have been transformed by a

transition between these two phases. This periodization should be treated synchronically: biologists (as well as computer scientists and mathematicians) did not stop using sequence comparison algorithms for working on evolutionary questions in the 1980s and 1990s; and examples of using sequence comparison to organize data can be found prior to 1980. Rather, what the paper describes is how technical developments and disciplinary concerns altered ideas about what sort of tool sequence comparison could be. Earlier and later, sequence comparison relied on the same basic fact: that sequences are altered by mutations and acted on by evolution in such a way that measuring differences or similarities in sequence can reveal much how life works. All the algorithms discussed here are measuring "homology," but the meaning and importance of this imputed similarity is different for the the periods described.[40]

In each phase, the particular technological, disciplinary, and epistemological circumstances acted to reinforce one another and to reinforce particular notions of what a sequence was. In the first phase – lasting roughly from the early 1960s to the late 1970s – sequence comparison algorithms were directed toward understanding evolution. In the context of the conflict between morphological and molecular studies of evolution, computational sequence comparison became a tool for augmenting the perceived objectivity of molecular methods. Sequences, embedded in the informational discourse of molecular biology, were a text, a Book of Life. As such, they could be made to tell an origin story – sequence comparison was an important problem because it vested the letters of the genetic code with history. The Needleman-Wunsch algorithm provided a canonical solution to the problems of sequence comparison that could be applied to problems beyond phylogeny reconstruction. In the second phase – lasting roughly from 1980 to the formal beginning of the HGP in 1990 – sequence comparison algorithms began to be used for both functional studies of genes and for organizing and managing the growing body of sequence information. The growing ubiquity of DNA sequencing was coupled to a

40 For an excellent discussion of how the concept of homology produces meaning in bioinformatics see Fujimura 1999.

genetic determinist imperative to demonstrate that the majority of variation amongst organisms could be linked to DNA sequence variation. Organizing sequences according to homology – that is, on the basis of similarity to other sequences (as determined by sequence comparison algorithms) – allowed geneticists to impute the function of many unknown stretches of DNA. Moves towards local rather than global similarity searches, and to heuristic rather than guaranteed-optimal alignments demonstrate how capturing functional similarity became more important than evolutionary hierarchy.

With the completion of the various genome projects it is now perhaps just possible to discern the beginnings of a third phase. The rise of the Human Genome Project and the concomitant availability of supercomputing resources for biology has once again transformed the meaning of sequence comparison. The decision to sequence the *entire* genome (not just the genes) marked a disciplinary turn towards a genomic, rather than genetic, biology. Most recently, I suggest, sequence comparison algorithms became crucial to whole-genome shotgun sequencing methods, and ultimately became to define what it means to do genomics – sequence comparison algorithms are deployed to understand the structure and meaning of whole genomes. In the genome projects a vast amount of protein coding and non-protein coding sequence data had to be ordered, categorized, and made sense of. Newly available supercomputers allowed the development of even more powerful sequence comparison algorithms that made 'sequencing the genome' an interesting, or even thinkable, project. Sequence comparison algorithms were the most important way in which biologists (as well as mathematicians and computer scientists) have attempted to make sense of the genome as a whole. Sequence comparison algorithms were used to compare (and thus conceptually link) short sequences from different parts of the genome. Through such connections the genome came to be understood as operating through a dense set of connections and interactions – the genome sequence became a network.

The shifting use of sequence comparison algorithms has been linked to changes in the

disciplinary configuration of biology. The influence of biochemists, mathematicians, physicists, and computer scientists has radically altered the sorts of questions posed and the kinds of answers given in biology. In particular, sequence comparison algorithms have played a significant role in the legitimation of molecular methods and in the use of statistical methods in understanding living things. Partly due to the influence of sequence comparison, studies of evolution became increasingly dominated by molecules; later, these algorithms made it increasingly plausible to ask questions about the molecular basis of biological function. Both these sorts of uses relied crucially on the ability of computers to perform statistical analysis on large volumes of sequence information. It was partly through the success of sequence comparison in parsing and organizing large sets of biological data that numerical methods borrowed from physics, mathematics, and computer science came to have increasing plausibility in biological work.

The changing status of sequence comparison algorithms as an instrument has been intertwined with the changing ontological and epistemological status of its object of study – the sequence itself. The locus 'information' in sequences has been transformed. Early on, information was associated with a text or code, with a kind of sacred or secret writing that framed the most important biological questions around histories and origin stories. No doubt, this strand has persisted: mitochondrial Eve, the Human Genome Diversity Project, the HapMap project, and the Genographic Project are all attempts to use sequence comparison to investigate (human) history. In this 1980s, however, this notion of information was overlaid with another that desacralized information and made it into data. The information within sequence became stuff to be managed, stored, organized, and searched, just like many other kinds of data.

Following my suggestion above, perhaps now information is increasingly understood in a third sense: as a network. Sequences, as the structure of of the entire organism, are a kind of scaffold on which the system of life is built. Information in genomes is not a static database, but a dynamic web.

These three senses – information as story, information as data, information as network – have rendered sequences different kinds of objects. Examining the development of sequence comparison algorithms has allowed us to see how sequences have been subjected to different sorts of questions, how they have been remade by software and hardware.

References:

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. 'Basic Local Alignment Search Tool.' *Journal of Molecular Biology* 215: 403-410.

Aronson, Jay. 2002. 'History of bioinformatics and molecular biology on the Net: The Dibner-Sloan history of recent science and technology project.' *Mendel Newsletter* 11: 5-8.

Barker, Winona C. and Margaret O. Dayhoff. 1982. 'Viral src gene products are related to the catalytic chain of mammalian cAMP-dependent protein kinase.' *Proceedings of the National Academy of Sciences USA* 79, no 9. (May): 2836-2839.

Barbieri, Marcello. 2003. 'Biology with information and meaning' *History and Philosophy of the Life Sciences* 25: 243-254.

Bellman, Richard E. 1984. *The eye of the hurricane: an autobiography*. Singapore: World Scientific Publishing Company.

Beyer, William A., Myron L. Stein, Temple F. Smith, Stanislaw Ulam. 1974. 'A molecular sequence metric and evolutionary trees.' *Mathematical biosciences* 19: 9-25.

Boniolo, Giovanni. 2003. 'Biology without information.' *History and Philosophy of the Life Sciences* 25: 255-273.

Brandt, Christina. 2005. 'Genetic code, text, and scripture: metaphors and narration in German molecular biology.' *Science in Context* 18, no. 4: 629-648.

Bryson, Vernon and Henry J. Vogel (eds). 1965. *Evolving genes and proteins.* A symposium held at the Institute of Microbiology of Rutgers, September 17-18, 1964. San Diego, CA: Academic Press.

Ceruzzi, Paul. 1999. *A history of modern computing.* Cambridge, MA: MIT Press, 2nd edition.

Cook-Deegan, Robert. 1994. *The gene wars: science, politics, and the human genome.* New York: W.W. Norton.

Crick, Francis. 1957. 'The structure of the nucleic acids and related substances.' pp. 173-179 in *Special Publication of the New York Academy of Sciences,* Part II, Vol. 5. New York: New York Academy of Sciences.

Dayhoff, Margaret O. and George E. Kimball. 1949. 'Punched card calculation of resonance energies.' *Journal of Chemical Physics* 17: 706-717.

Dayhoff, Margaret O. 1964. 'Computer aids to protein sequence determination.' *Journal of Theoretical Biology* 8: 97-112.

Dayhoff, Margaret O. and Richard V. Eck. 1968. *Atlas of Protein Sequence and Structure, 1967-68.* Silver Spring, MD: National Biomedical Research Foundation.

Dayhoff, Margaret O. 1969. 'Computer Analysis of Protein Evolution.' *Scientific American* 221: 87-95.

Department of Energy. 1986. *Sequencing the Human Genome: Summary report of the Santa Fe Workshop.* Santa Fe, NM, 3-4 March 1986. Washington, DC: US Department of Energy, Office of Health and Environmental Research.

Dietrich, Michael R. 1994. 'The origins of the neutral theory of molecular evolution.' *Journal of the History of Biology* 20: 21-59.

Dietrich, Michael R. 1998. 'Paradox and persuasions: negotiating the place of molecular evolution within evolutionary biology.' *Journal of the History of Biology* 31, no 1: 85-111.

Dietrich, Michael R. 2000. The problem of the gene. *Comptes Rendus de l'Academie des Sciences, Paris, Sciences de la Vie / Life Sciences* 323: 1139-1146.

Doolittle, Russell F. 1981. 'Similar amino acid sequences: chance or common ancestry?' *Science* 214, no. 4517: 149-159.

Doolittle, Russell F., Michael W. Hunkapiller, Leroy E. Hood, Sushilkumar G. Devare, Keith C. Robbins, Stuart A. Aaronson Harry N. Antoniades. 1983. 'Simian Sarcoma Virus onc gene, v-sis, is

derived from the Gene (or Genes) encoding a platelet derived growth factor.' *Science* 221, no. 4607: 275-277.

Doolittle, Russell F. 1997. 'Some reflections on the early days of sequence searching.' *Journal of Molecular Medicine* 75: 239-241.

Doolittle, Russell F. 2000. 'On the trail of protein sequences.' *Bioinformatics* 16, no. 1: 24-33.

Dreyfus, Stuart. 2002. 'Richard Bellman on the birth of dynamic programming.' *Operations Research* 50, no. 1 (Jan-Feb): 48-51.

Elzen, Boelie. 1986. 'Two ultracentrifuges: a comparative history of the social construction of artifacts.' *Social Studies of Science* 16: 621-662.

Fitch, Walter M. 1964. 'The probable sequence of nucleotides in some codons.' *Proceedings of the National Academy of Sciences USA* 52 (August): 298-305.

Fitch, Walter M. 1966. 'An improved method of testing for evolutionary homology.' *Journal of Molecular Biology* 16: 9-16.

Fitch, Walter M. and E. Margoliash. 1967. 'Construction of phylogenetic trees.' *Science* 155, no. 760: 279-284.

Fleischmann, Robert D., et al. 1995. 'Whole-Genome random sequencing and assembly of *Haemophilus influenzae* Rd.' *Science* 269, no. 5223: 496-512.

Fogle, Thomas. 1995. 'Information metaphors and the Human Genome Project.' *Perspectives in Biology and Medicine* 38: 535-547.

Fortun, Michael. 1999. 'Projecting speed genomics.' pp. 25-48 in *The Practices of Human Genetics.* Edited by Michael Fortun and Everett Mendelsohn. Dordrecht: Kluwer.

Fujimura, Joan H. 1999. 'The practices of producing meaning in bioinformatics.' pp. 49-87 in *The Practices of Human Genetics.* Edited by Michael Fortun and Everett Mendelsohn. Dordrecht: Kluwer.

Gaudillière, Jean-Paul. 2001. 'Making mice and other devices: the dynamics of instrumentation in American biomedical research (1930-1960).' pp. 175-198 in *Instrumentation between science, state, and industry.* Edited by Bernward Joerges and Terry Shinn. Dordrecht: Kluwer.

Gingeras, T.R., J.P. Milazzo, D. Sciaky, R.J. Roberts. 1979. 'Computer programs for the assembly of DNA sequences.' *Nucleic Acids Research* 7, no. 2(September 25): 529–545.

Goad, Walter B. and Minoru I. Kanehisa. 1982. 'Pattern recognition in nucleic sequences. I. A general method for finding local homologies and symmetries.' *Nucleic Acids Research* 10, no. 1 (January 11): 247-263.

Goad, Walter B. 1987. 'Sequence analysis: contributions by Ulam to molecular genetics.' *Los Alamos Science* Special Issue: 288-291.

Hagen, Joel B. 1999. 'Naturalists, Molecular Biologists, and the Challenges of Molecular Evolution.' *Journal of the History of Biology* 32: 321-341.

Hagen, Joel B. 2001. 'The introduction of computers into systematic research in the United States during the 1960s.' *Studies in the history and philosophy of science part C: Studies in the history and philosophy of biological and biomedical sciences* 32C: 291-314.

Harding, Anne. 2005. 'BLAST: How 90000 lines of code helped spark the bioinformatics explosion.' *The Scientist* 19(16): 21-26.

Hilts, Philip J. 1983. 'Scientists may have found one missing link in cause of cancer.' *Washington Post*, 30th June: A4.

Hunt, Lois. 1984. 'Margaret Oakley Dayhoff, 1925-1983.' *Bulletin of Mathematical Biology* 46, no. 4: 467-472.

Jasanoff, Sheila. 2006. *States of knowledge: the co-production of science and the social order.* New York: Routledge.

Jeffreys, A.J., V. Wilson, S.W. Thein. 1985. 'Hypervariable "minisatellite" regions in human DNA.' *Nature* 314: 67-73.

Jones, Niel C. and Pavel A. Pevzner. 2004. *An introduction to bioinformatics algorithms.* Cambridge, MA: MIT Press.

Kay, Lily E. 1988. 'Laboratory technology and biological knowledge: the Tiselius electrophoresis apparatus, 1930-1945.' *History and Philosophy of the Life Sciences* 10, no. 1: 51-72.

Kay, Lily E. 1993. 'Life as technology: representing, intervening, and molecularizing.' *Rivista di storia della scienza* ser. 2, vol. 1, no. 1: 85-103.

Kay, Lily E. 2000. *Who wrote the book of life? A history of the genetic code.* Palo Alto, CA: Stanford University Press.

Keller, Evelyn Fox. 1992. 'Nature, nurture, and the human genome project.' pp. 281-299 in *The code of codes: scientific and social issues in the human genome project.* Edited by D. Kevles and L. Hood. Cambridge, Massachusetts: Harvard University Press.

Keller, Evelyn Fox. 1994. 'Master molecules.' pp. 89-98 in *Are Genes Us? The social consequences of the new genetics*. Edited by C.F. Cranor. Piscataway, NJ: Rutgers University Press.

Keller, Evelyn Fox. 2002. *The Century of the Gene.* Cambridge, MA: Harvard University Press.

Kohler, Robert. 1994. *Lords of the fly: Drosophila genetics and the experimental life.* Chicago, IL: University of Chicago Press.

Korn, Lawrence J., Cary L. Queen, Mark N. Wegman. 1977. 'Computer analysis of nucleic acid

regulatory sequence.' *Proceedings of the National Academy of Sciences USA* 74(10): 4401-4405.

Landecker, Hannah. 2007. *Culturing life: how cells became technologies.* Cambridge, MA: Harvard University Press.

Lenoir, Timothy. 1999. 'Virtual reality comes of age.' pp. 226-249 in *Funding a revolution: government support for computing research.* Washington, DC: National Research Council.

Lipman, David J. and William R. Pearson 1985. 'Rapid and sensitive protein similarity searches.' *Science* 227, no. 4693: 1435-1441.

Mahoney, Michael. 2005. 'Histories of computing(s).' *Interdisciplinary Science Reviews* 30: 119-135.

McAdams, Harley H. and Lucy Shapiro. 1995. 'Circuit simulation of genetic networks.' *Science* 269(5224): 650-656.

Maxam, Allan M., Walter Gilbert. 1977. 'A new method for sequencing DNA.' *Proceedings of the National Academy of Sciences USA* 74, no. 2 (February): 560-564.

Morgan, Gregory J. 1998. 'Emile Zuckerkandl, Linus Pauling and the molecular evolutionary clock, 1959-1965.' *Journal of the History of Biology* 31, no. 2: 155-178.

National Academy of Sciences. 1988. *Report of the Committee on Mapping and Sequencing the Human Genome.* Washington, DC: National Academy Press.

Needleman, Saul B. and E. Margoliash. 1966. 'Rabbit heart cytochrome c.' *Journal of Biological Chemistry* 241, no. 4 (February 25): 853-863.

Needleman, Saul B. and Christian D. Wunsch. 1970. 'A general method applicable to the search for similarities in the amino acid sequence of two proteins.' *Journal of Molecular Biology* 48: 443-453.

Nelkin, Dorothy and M. Susan Lindee. 1996. *The DNA mystique: the gene as cultural icon.* New York: W.H. Freeman.

November, Joseph A. 2006. 'Digitizing life: the introduction of computers into biology and medicine.' PhD diss., Department of History, Princeton University.

Pauling, Linus C., Harvey Itano, S.J. Singer, Ibert Wells. 1949. 'Sickle cell anemia, a molecular disease.' *Science* 110 (November): 543-548.

Pauling, Linus and Emile Zuckerkandl. 1965. 'Divergence and convergence in proteins.' pp. 97-166 in *Evolving genes and proteins.* Edited by Vernon Bryson and Henry J. Vogel. San Diego, CA: Academic Press.

Rasmussen, Nicolas. 1997. *Picture control: the electron microscope and the transformation of biology in America, 1940-1960.* Palo Alto, CA: Stanford University Press.

Reardon, Jenny. 2005. *Race to the finish: identity and governance in an age of genomics.* Princeton,

NJ: Princeton University Press.

Rheinberger, Hans-Jörg. 2000. 'Beyond nature and culture: modes of reasoning in the age of molecular biology and medicine.' pp. 19-30 in *Living and working with the new medical technologies.* Edited by Margaret Lock, Alan Young, Alberto Cambrosio. Cambridge, UK: Cambridge University Press.

Sanger, Frederick, S. Nicklen, A.R. Coulson. 1977. 'DNA sequencing with chain-terminating inhibitors.' *Proceedings of the National Academy of Sciences USA* 74, no. 12 (December): 5463-5467.

Sarkar, Sohotra. 1996. 'Biological information: a skeptical look at some central dogmas of molecular biology.' pp. 187-231 in *The Philosophy and History of Molecular Biology: New Perspectives.* Edited by Sohotra Sarkar. Dordrecht: Kluwer.

Sarkar, Sohotra. 1997. "Decoding "coding" – information and DNA.' *European Journal for Semiotic Studies* 9, no. 2: 277-298.

Schmeck, Harold M. Jr. 1983. 'Cancer gene linked to natural human substance.' *New York Times*, June 30: B11.

Segal, Jerome. 2003. 'The use of information theory in biology: a historical perspective.' *History and Philosophy of the Life Sciences* 25: 275-281.

Sellers, Peter H. 1974. 'On the theory and computation of evolutionary distances.' *SIAM Journal of Applied Mathematics* 26, no. 4: 787-793.

Sellers, Peter H. 1979. 'Pattern recognition in genetic sequences.' *Proceedings of the National Academy of Sciences USA* 76, no. 7 (July): 3041.

Smith, E.L. and E. Margoliash. 1964. 'Evolution of Cytochrome C.' *Federation Proceedings* 23 (November-December): 1243-1247.

Smith, Temple F., Michael S. Waterman, W.M. Fitch. 1981. 'Comparative biosequence metrics.' *Journal of Molecular Evolution* 18: 38-46.

Smith, Temple and Michael S. Waterman. 1981. 'Identification of common molecular subsequences.' *Journal of Molecular Biology* 147: 195-197.

Sommer, Marianne. 2008. 'History in the gene: negotiations between molecular and organismal anthropology.' *Journal for the History of Biology* 41, no. 3: 473-528.

Strasser, Bruno J. 2006. 'Collecting and experimenting: the moral economies of biological research, 1960s-1980s.' pp. 105-123 in *History and epistemology of molecular biology and beyond: Problems and perspectives. W*orkshop at Max Planck Institute für Wissenschaftsgeschihte, 13-15 October 2005.

Strasser, Bruno J. 2008. 'GenBank – Natural history in the 21st century?' *Science* 332, no. 5901(October 24): 537-538.

Suárez, Edna. 2008a. 'The rhetoric of informational macromolecules: authority and promises in the

early study of molecular evolution.' *Science in Context* 20, no. 4: 1-29.

Suárez, Edna. 2008b. 'Sequences, quantification, and objectivity in the construction of phylogenies.' Paper presented to 'Making Sequences Matter' Conference, Yale University, 19-21 June 2008.

Thorne, J.L., H. Kishino, J. Felsenstein (1991). 'A evolutionary model for maximum liklihood alignment of DNA sequences.' *Journal of Molecular Biology* 33: 114-124.

Ulam, Stanislaw. 1972. 'Some ideas and prospects is biomathematics.' *Annual Reviews of Biophysics and Bioengineering* 1: 277-291.

Venter, J. Craig. 2007. *A life decoded: my genome: my life.* London: Viking Penguin.

Waterfield, Michael D., G.T. Scrace, N. Whittle, P. Stroobant, A. Johnsson, A. Wasteson, B. Westermark, C.H. Heldin, J.S. Huang, T.S. Deuel. 1983. 'Platelet-derived growth factor is structurally related to the putative transforming protein p28sis of the simian sarcoma virus.' *Nature* 304: 35-39.

Waterman, Michael S., Temple F. Smith, W.A. Beyer. 1976. 'Some biological sequence metrics.' *Advances in Mathematics* 20: 367-387.

Waterman, Michael S., Temple F. Smith, M. Singh, W.A. Beyer. 1977. 'Additive evolutionary trees.' *Journal of Theoretical Biology* 64: 199-213.

Waterman, Michael. 1999. *Skiing the sun: New Mexico essays.* http://www-hto.usc.edu/people/msw/newmex.pdf (Accessed 1 November 2008).

Wilbur, W.J. and David J. Lipman. 1983. 'Rapid similarity searches of nucleic acid and protein data banks.' *Proceedings of the National Academy of Sciences USA* 80: 726-730.

Yoxen, Edward J. 1982. 'Constructing genetic diseases.' pp. 144-161 in *The problem of medical knowledge*. Edited by P. Wright and A. Treacher. Edinburgh: Edinburgh University Press.