Hist. Phil. Life Sci., 32 (2010), 65-90

Making Room for New Faces: Evolution, Genomics and the Growth of Bioinformatics

Edna Suárez-Díaz

Universidad Nacional Autónoma de México
Facultad de Ciencias, Circuito Exterior, Cd. Universitaria
Copilco, Coyoacán, 04510,
Mexico City, México

ABSTRACT – Genomics poses challenges that are specific to historians of science. Such challenges are not necessarily met by most recent sociologically-oriented approaches. This paper argues that historians of genomics can draw some lessons from the history of molecular biology, in part because some of the actors, concepts, and tools have made a transition between the two fields. More importantly, historians face the marginalization of scientific fields and actors that played a role in the integration of both ultra-disciplines. While biochemistry and genetics played an underrated role in the rising of molecular biology, research on the molecular evolution of informational molecules (molecular phylogenetics) played a neglected but nevertheless central role in the development of conceptual and analytical bioinformatics tools for genomics. Even today genomic tools incorporate underlying assumptions that show their origins in problems of comparative biology. This is particularly true in the case of the algorithms for sequence alignment, first proposed by Needleman and Wunsch (1970). The present essay also makes reference to areas in the history of science that require further investigation for an understanding of the transformations brought about by genomics to biological research, namely, the role of automation – beyond sequencing – and the intersection of biology and mathematics.

KEYWORDS – Presentism, genomics, molecular phylogenetics, sequences, data bases, bioinformatics, automation

Introduction

Writing the history of genomics poses a number of challenges to historians of science. The history of science, as practiced in the last decades, aims to depict the social and material dimensions of scientific practice in its own context. In this broad definition, history touches the goals traditionally associated with sociology. However, the writing of history is supposed to deliver detailed genealogical narratives that show

temporal connections between ideas, fields, events, actors, and objects that mediate between past and present. Some of these connections may not be apparent from the viewpoint of the synchronic analysis characteristic of sociological studies. Genomics is even more obscured by the fact that we are dealing with a case in the recent history of science. On some particular issues the temporal dimension does not have enough depth and thus it may be easy to take sociological accounts for historical ones, not that these perspectives are exclusive. On the contrary, sociological and historical accounts illuminate each other. [1] But in order to be useful to each other, the historical perspective needs to bring into light the actors, objects, and spaces that are not revisited from a more contemporary or synchronic perspective.

The present essay aims to shed light on some of the specific tasks of historians of science interested on the mediations that eventually transformed a complex set of practices in 20th-century biology into genomics. To do so, I have divided this article into three sections. First, I make reference to some historiographical problems faced by historians of molecular biology and their relevance for investigations into genomics. I rely on previous critical work in the historiography of science in general and molecular biology in particular. The second part takes one of the most conspicuous concerns of historians of molecular biology, namely, the contribution of disciplines "left on the margins," to reflect on actors and tools that have not figured prominently in recent accounts of genomics. In particular, I argue for the usefulness of histories of tools as a means to bring into focus the contribution of fields that have been left out from the picture in both sociological-oriented (STS) accounts and those of scientists. This reflection is illustrated in the third section by a brief account of the development of algorithms for sequence alignment and comparison, initiated by Saul B. Needleman and Christian Wunsch (1970). This case study shows the relevance, connections, and even continuities between tools developed for taxonomic and phylogenetic concerns and the tools of sequence analysis that figure so prominently in genomic practices.

[1] As will be clear in the text, the focus on history and its challenges is not intended as an overall critique of the STS perspective, nor of sociologically informed accounts. Moreover, because of the integrative nature of genomics, some of the questions faced by historians cannot be met without an interdisciplinary approach, as Sabina Leonelli has forcefully argued (this issue). The automation of scientific research and the transformation that this process has provoked in the organization of work, is a clear example of an area of study requiring all types of approaches in science studies.

The History of Science as Written by Scientists

In contrast to the first professional accounts of the history of molecular biology (Olby 1974; Judson 1979; Morange 1998), students of genomics have used a vast array of interpretative tools taken from the social sciences and the humanities. Most of these studies describe a dynamic enterprise in which layers of science, technology, industry, and government intersect and interact in profound and novel ways, reflecting not only current trends in science and technology studies, but the diverse transformations that have taken place in late 20th- and early 21st-century life sciences research. Accounts of genomics range from the analysis of the new organization of scientific research (Fujimura 1999; Hilgartner 2004; Kaufman 2004; Ramillon 2007; Leonelli 2007; Strasser 2008) to the latest biotech and state ventures in the context of broader reflections on the globalization of markets, politics and information (Parry 2004; Reardon 2004; Sunder Rajan 2006; Fortun 2008; see also Thacker 2006). All these studies have obvious and, to my view, beneficial impacts on the way historians may approach the vast field of genomics. At the very least, they have opened a number of spaces to look for historical research. For their part, historical accounts have also paid attention to the development of material and social practices of genomics, mainly focusing on model organisms (Ankeny 2001; 2010; De Chadarevian 2004; Leonelli 2007) and the Human Genome Project (Kevles and Hood 1992; Beatty 2000; Cook-Deegan 2004).

Still, there is a lot of work left in this area for professional historians. Despite the obvious difference in the diversity of approaches, the historiography of genomics has several points of contact with the historiography of molecular biology, if only because many scientists, including prominent figures like James D. Watson, Craig Venter, Charles Cantor and Francis Collins, have made their way from one field of practice to the other. Thus, a first historical question arises as to the context where such personal reconfigurations have taken place. This is analogous to the specific and sometimes very local configurations that led many biochemists, physicists, and biophysicists from the 1960s to call themselves "molecular biologists."[2]

[2] This issue has captured a lot of attention from historians of biology. See, for example De Chadarevian (1996) for the case of chemist Fred Sanger, or Santesmases (2006) for biochemist Severo Ochoa and the papers brought together by Tibor Frank on physicist Leo Szilard (2004). More generally, De Chadarevian (2004) makes reference to the transformation of physicists, chemists, and biophysicists into molecular biologists at the Cambridge Laboratory of Molecular Biology at the end of the 1960s. Gaudillière (1996) also makes reference to the changing status of François Jacob and Jacques Monod from microbial genetics and enzymology to molecular biology.

In the case of genomics, the interaction of former molecular biologists with computer scientists (including former mathematicians and physicists), technological innovators, and administrators, gave way to new types of scientists, some of them seldom worried with wet experiments but some deeply engaged in the details of organismic development.

Moreover, some of the central characters in genomics have been instrumental in extending their personal narratives to account for the transformations in their field of research. In this vein, one of the most conspicuous continuities has been the involvement of leading scientists in the writing and legitimating of the history of genomics, much in the same way as leading molecular biologists of a generation ago wrote their personal accounts of their own field (for instance, Stent 1968). In the past decade this trend has comprised the writing of autobiographies (Sulston and Ferry 2002; Venter 2007) and personal accounts of the field (Cook-Deegan 1994), with an obvious focus on the Human Genome Project. These histories have developed a perspective where previous developments in the life sciences are seen as inevitably leading to the HGP.

Pnina Abir-Am's earlier reflections on the historiography of molecular biology and, in particular, the role played by myths of origins enacted in scientists'accounts, are clearly relevant in a reflection on our present difficulties in the writing of genomics (Abir-Am 1985; 1999). Two different but interrelated aspects of her analysis are of particular interest here. The first one refers to the themes and genres brought about by prominent scientists; the second, to the complexities of scientific research in what Abir-Am calls the "ultra-discipline" of molecular biology, "integrating resources from half a dozen traditional disciplines" (Abir-Am 1985, 73).

Concerning the first, historians have for a long time debated on the epistemic value of the history of science as written by scientists. Besides the resonance of this debate on claims about disciplinary boundaries and cognitive authority, the arguments of historians have revolved around the perils of presentist agendas and, at worst, of whiggish history.[3] More specifically, many of the questions that the historian of recent science asks are, by definition, shaped by the same contemporary concerns of his/her subjects. The globalization of knowledge, to name one of the most conspicuous concerns, certainly affects genomic practices as much as historical research.

[3] For instance, see the opposing arguments of Forman (1991), banning all types of presentism, arguing for a thorough historicism as a way to attain i*ndependence* from scientists and the position defended by Hull (1979) and Brush (1995), who defend the virtues and unavoidability of a mild presentism. The consensus among historians of science, in the decades after George Sarton, is a common cause against whiggish history. A thorough revision of references on this debate is included in Brush (1995).

Thus, today it is quite accepted that historians should not rely on scientists' accounts of their fields in order to attain independence (Forman 1991), but it also seems clear that scientists' accounts can be useful and also historically meaningful. Meaningful contributions by scientists, however, are not always easy to discern. As early as the mid- to late-1960s, the first personal accounts on the origins and development of molecular biology were published by some of the most visible scientists in the field (Cairns, Stent and Watson 1966; Watson 1968; Stent 1968). These ranged from autobiographies (Watson 1968; Jacob 1988; Crick 1988; Stent 1998, etc.) to any subject related to their science: from the nature of discovery (Jacob 1988), the aims of science, the *nature* of nature and human nature (Monod 1972), and even the end of progress (as Gunther Stent did in 1969), to "the less glorified aspects of science and society," as noted by Pnina Abir-Am on a review of molecular biologists' autobiographies (1991, 327).

These actors' accounts revealed the growth of a less mythical and more candid relation between science and society, touching on personal, national, and institutional rivalries, as well as "micro- and macro-politics of dubious morality." Perhaps, says Abir-Am, such "antiheroic narratives" resulted from scientists' growing awareness of public interest in a less idealized image of science. Or, as Michael Bishop, co-discoverer of the oncogenes and Nobel laureate says in his own autobiographical account, the point was to show that scientists are "supremely human," in an effort to reconcile science and society. At the end of the 1960s and through the 1970s and 1980s, such efforts clearly reflected the changing position of science within society at large, a question of the greatest interest to professional historians (Bishop 2004).

Abir-Am argues, however, that in this context and despite the very different types of characters, most of molecular biologists' autobiographies shared the plot of a "scientific Cinderella," depicting a life that grows from marginality to stardom in the context of post-War biology and the development of biotechnology (Abir-Am 1991, 327). In its recent obituary of Gunther Stent (June 16, 2008), for instance, *The New York Times* quotes Stent's colleague stating that he belonged to the "small band of pioneers," that included a few scientists in the circle surrounding Watson, Crick, and Delbrück as well as the French Pasteurians. Not surprisingly, the "small band of pioneers" includes the most vocal representatives of the Cinderella plot and, according to Abir-Am, the main builders of a legitimate order for the new biology.

Despite the mass of work devoted to deconstruct scientists' accounts and agendas, a fair historical evaluation must recognize the value of the autobiographical genre. What is more significant about an autobiographical account is not that it evokes the past, but that it offers

"different appreciations of what it is desirable to recall" (Abir-Am 1985, 326). Such considerations are more than relevant when evaluating the recent publications of many of the most prominent first-generation genomicists. A small band of genomicists obviously includes Craig Venter (2007), but also figures like John Sulston (Sulston and Ferry 2002) and Francis Collins (2007).[4] The Cinderella of the market and the heroes of public interest have greatly influenced the subjects and themes chosen by sociologically-informed accounts of genomics. Notoriously, with a few exceptions (Cantor 1992; Cook-Deegan 1994), these genomicists' narratives have not paid attention to the actual builders of the material and conceptual tools, methods, and models of this interdisciplinary multi-layered field.

In this context, and notwithstanding his opinions on other aspects of the debate, Paul Forman's advice seems useful as a regulatory ideal: "only by *thoroughly* historicizing scientific knowledge – explaining possession of specific pieces or structures of it, not by appealing to a trascendent reality . . . , but by reference to mundane factors and human actors – can historians of science move from whiggery and toward intellectual independence" (Forman 1991, 78). As I will demonstrate, the recent focus on the historicity of such "mundane factors" as instruments and tools has shown a good exit from this endless debate.

Even in those areas where attention to technology and innovators has been granted, such as the case of sequencing technologies, we still know very little about the development of competing alternatives, such as the innovations that took place in Japan (see Fujimura 1999). More importantly, we lack historical studies that cover the broad range of tools used in genomics. The focus on automated sequencers, given the prominent role of Leroy Hood and Applied Biosystems in the history of the HGP, has had the unintended consequence of creating an historical vacuum. We know nothing, for instance, about the development of robotics for the construction of genetic and molecular libraries, even though these seemingly "humble" and more trivial technologies are as important to genomics as the sequencing machines.[5] On this issue, it is also relevant not to forget that the drive to automation has not been limited to sequencing procedures and extends to the automation of inference procedures.

[4] Francis Collins's book, however, stands alone in his personal struggle to bring science and religion together. In his more recent book Collins (to be delivered in 2010) seeks to present genomic medicine in a vivid and accessible way, and "sets out hope without hype, and will enrich the mind and uplift the heart" (from Amazon's Editorial Reviews).

[5] Vincent Ramillon made me realize this point (personal communication).

One of the few references to the many other unacknowledged technologies of genomics, dating back to electrophoresis and nucleic acid hybridization, is the paper by Charles Cantor included in Daniel Kevles's and Leroy Hood's volume on the Human Genome Project (1992). This case illustrates both the meaningful historicity of the actors' accounts and the lack of attention by professional historians of science to the more mundane objects of scientific research. In this case the marginalization is all the more inexcusable given Cantor's prominent role in the history of the HGP. Cantor, former director of the Department of Energy Human Genome Project and co-author with Berkeley colleague Cassandra Smith of one of the first textbooks on genomics (Cantor and Smith 1999), has been granted 54 US patents. Some of them, like the one on the automatic production of molecular libraries and pulse field gel electrophoresis, are included in his textbook, a valuable source for historians of science interested in the range of genomic techniques.

Moreover, to date, even the triumphant technologies of automated sequencers and bioinformatics appear in most of the available accounts as subservient to the main plots we have inherited from the scientists' accounts: either the maverick scientist-entrepreneur who has fought for his individual right to advance scientific research (Venter 2007) or the intersection of industry, government, and university research (Cook-Deegan 1994; Sulston and Ferry 2002; and sociologically-informed accounts of Hilgartner 2004; Kaufman 2004).

One generation ago, as an unintended consequence of the leading actors' prolific writing, some historians of molecular biology built what Abir-Am (1985) calls a "second-order" legitimating regime, one that institutionalized the actor's narratives within history itself.[6] On the other hand, many historians of science as well as scientists whose disciplinary practices had been left on the margin, struggled to offer alternative narratives of molecular biology. In particular, historians of molecular biology soon learned that the image of physics as the main contributor to the origins of molecular biology and the corresponding dismissal of the role played by biochemistry and classical genetics, was only a socio-professional myth and a legitimating resource, enacted by physicists turned molecular biologists. The myth, however, left its mark. It took the work of many historians of science to deconstruct it and replace it by a more equilibrated and sophisticated account of the disciplinary

[6] Following the themes and genres of this early group of "chroniclers," historians were trapped in the actor's plots. *The Eighth Day of Creation* (1978), written by journalist-historian of science Horace F. Judson and by far one of the most popular books on the history of molecular biology, including dozens of useful interviews, shares many of the tropes, actors, and episodes of the *small band*.

sources that gave rise to molecular biology (see Abir-Am 1992; papers included in De Chadarevian and Gaudillière 1996).[7] This last point takes us directly to my second historiographical concern in writing the history of genomics, the difficulties of writing about a multi-layered, inter- or even trans-disciplinary field that integrates the practices and socio-professional features of at least a handful of disciplines.

Molecular Evolution and Bioinformatics

When the practices of what became to be known as "molecular biology" started to form at the end of the 1930s and up to the late 1950s, interdisciplinarity was not the exchange commodity it is today. Molecular biology integrated physicists, organic chemists, biochemists, geneticists, and all kinds of life science practitioners escaping from medical research (on the latter, Gaudillière 1996). Issues of prestige, epistemic authority, and power were at the core of the redefinition of disciplinary boundaries in post-war biology. Some historians of molecular biology have also documented, not at the level of the labels, but at the level of scientific practices, the many contingencies affecting the professional careers and choices of individual scientists that made their ways from one field of research to the new interdisciplinary field. Thus, for instance, Soraya de Chadarevian has accounted for the career of chemist Fred Sanger in the context of "negotiations between biochemists and molecular biologists (mostly crystallographers) accompanying the construction of the new field of molecular biology" (1996, 362). Sanger's work on the sequencing of insulin confirmed that proteins consisted of specific sequences of chains of amino acids, thus providing a direct link with the idea of genetic information and the early speculations about the genetic code. De Chadarevian also documents the fact that the crystallographers feared their field would someday become obsolete and the idea that predicting function from sequences (structure) alone might some day play a central role in the understanding of biological systems. Thus, when deciding to move to the new MRC Laboratory of Molecular Biology at Cambridge in the mid 1960s after winning the Nobel Prize in 1959 (for

[7] The literature on this issue is enormous and continues to grow. For instance, to get a feeling of the centrality of biochemistry and genetics in the everyday activities and research programs of molecular biologists, one could just go through the pages of Larry Holmes's recent and last book on Seymour Benzer (2006, ch. 2) for a detailed reevaluation of classical genetics to the early research of Max Delbrück or Maria Jesus Santesmases's also recent biography of Severo Ochoa (2006) and his reconfiguration from biochemistry to molecular biology.

the first time), Sanger was not in a weak position as a chemist-biochemist, a well-established discipline to which he belonged by training. However, collaborative networks that had taken place at the bench level "were used for and transformed by institutional and disciplinary developments." Thus, Sanger's own contribution to the "success story" of the Molecular Biology Laboratory at Cambridge helped to change the power balance between disciplines. Becoming a molecular biologist, thus, was not only a label. "Names matter. They are not only labels or reference terms for historical accounts, but strategic tools" (De Chadarevian 2002, 206; Powell et al. 2007).

The transformation of knowledge production in the last decades of the 20th century and the impact it has had on the disciplinary structure of the life sciences has delivered a new research landscape for scientists and historians at the beginning of the 21st century (De Chadarevian and Rheinberger 2006; 2009; Morange 1998; Suárez 2009). Inter-disciplinarity does not need to be construed from the ground anymore (Gibbons et al. 1994; Nowotny et al. 2001). It does not need to be justified in a world where scientific research is often socially distributed. However, this does not mean that issues of cognitive and political power do not arise where professional boundaries are reconfigured by scientific and technological developments. On this issue, no historical or sociological research has been done for the case of genomics. Nevertheless, it would be desirable to understand the alliances and struggles between mathematicians and physicists in their interaction with life scientists, given that the main components of genomics are computer science (bioinformatics and mathematics) and molecular biology.

Indeed, protein and nucleic acid sequence databases, tools for sequence analysis, and computers are three elements of this new configuration, whose development took place in very close relation to each other and with overlapping goals, in particular during the last three decades. This configuration, however, came together for the first time in the mid-1960s, with the earlier attempts to construct molecular evolutionary trees by comparing protein sequences. I will not attempt the impossible task of showing that the origins of bioinformatics and comparative or structural genomics lie in the rise of molecular phylogenetic analysis. Sequence analysis is a much broader endeavor, driven by the relatively autonomous interests of mathematicians and computer scientists. But attention to the development of concrete analytical tools shows the historical connection between problems and methodological goals in bioinformatics and evolutionary biology.

This connection is frequently acknowledged by practitioners of bioinformatics. Current textbooks, ranging from *Bioinformatics for*

Dummies (Claverie and Notredame 2007) to Jin Xiong's *Essential Bioinformatics* (2006), include a "historical" section on this issue in their introductory chapters. For instance, in his introduction, Xiong claims: "the earliest bioinformatics efforts can be traced back to the 1960s, although the word bioinformatics did not exist there. Probably, the first major bioinformatics project was undertaken by Margaret Dayhoff in 1965, who developed a first protein sequence data base called *Atlas of Protein Sequence and Structure*" (2006, 3). Most textbooks, also include chapters on the uses of bioinformatics for phylogenetic analysis. One could say that the application of bioinformatics tools in the construction of evolutionary trees is one of its most obvious uses. But in view of what the historiography of molecular biology has taught us, we should be suspicious of the scientists' claims to past disciplinary ancestors and myths of origins. By redirecting our attention to the relation between evolutionary biology and bioinformatics we find meaningful historical relations that are more subtle and complex than the scientists' historical accounts. In particular, several of the tools and practices of bioinformatics remain deeply tied or *entrenched* to problems of evolutionary biology.

For example, evolutionary trees are frequently the underlying model in the construction of algorithms, optimization criteria, and software packages for multiple sequence alignment (MSA). A tree model consists of a tree topology and a model of accepted mutations along the branches. Basically, this is a classical Darwinian representation for ancestral relations. But the entrenchment of the tree model in MSA algorithms may also be a source of problems and a door to the development of methods and new perspectives in genomics. Given that multiple sequence alignments are used for several purposes in molecular biology and genomics, the tree model sometimes fails. For instance, in cases of lateral gene transfer, where an entire ancestry line cannot be modeled by a unique tree or, in cases of convergent evolution, a dialogue between evolutionary biologists and computer scientists is of help to develop different alignment methods and software packages, not only in order to solve problems in phylogenetics, but to address specific analytic questions in functional and structural genomics (Durand 1997).

What is relevant about this example is that it points to the overlapping of models and resources between molecular phylogenetics or, broadly speaking, evolution, bioinformatics, and genomics. This overlap cannot be accounted for by the simplistic idea that the construction of phylogenies is one of the many applications of bioinformatics and genomics. The problems faced by molecular evolutionists in the 1960s helped to shape the first attempts to develop methods for sequence analysis and many of the underlying considerations in bioinformatics tools still carry assumptions and commitments associated with phylogenetic analysis.

Indeed, the origins of bioinformatics can be traced to three different but intersecting lines of research and technological development already present in the mid-1960s: the accumulation, first, of protein sequences, and later of RNA and DNA sequences, due to previous research in protein and nucleic acid biochemistry and automation (see García-Sancho 2009), followed by the development of data bases by statisticians and computer scientists (Smith 1990; Strasser 2008); the implementation of fruitful research programs based on the idea that biological macromolecules carry information; and the introduction of computers in biological research (Lenoir 1999; Hagen 2000; 2001; November 2006).

Regardless of the problematic connection between the "information discourse" in molecular genetics and information theory (Kay 2000; Fox Keller 2002; Brandt 2005; Suárez 2007), the idea that proteins and nucleic acids carry evolutionary or phylogenetic information was advanced as a fruitful research program since the early 1960s, in large part thanks to the work of Emile Zuckerkandl and Linus Pauling (Zuckerkandl and Pauling 1965a). Using passionate rhetoric, they reduced the problem of reconstructing the history of species using a quantitative comparative analysis of the information contained in its semantides; that is, molecules with meaning (Zuckerkandl 2005, personal communication), which they treated as historical documents (Dietrich 1998; Hagen 1999; Suárez 2007; Sommer 2008).[8]

During the first half of the 1960s and well into the 1970s, the comparative analysis of proteins made use of qualitative and semi-quantitative methods, given the scarcity of sequence data (Needleman and Margoliash 1963; Margoliash 1963; Zuckerkandl and Pauling 1962, 1965b). As more amino acid sequences became available, the need for computational tools to search in the nascent databases grew. By the end of the 1960s, however, Pehr Edman had already developed a fully automated protein sequencing machine (the *sequenator*) that used the degradation reaction he had developed some years before (Edman 1967; 1970). Using a somehow different approach, during the early 1970s Stanford Moore and William Stein at the Rockefeller Institute, using semi-automated techniques, were able to sequence the 174 amino acids of ribonuclease in half the time that Fred Sanger had used to sequence the insulin molecule (Moore and Stein 1973).[9]

[8] Curiously enough, Georges Cuvier – founder of paleontology – seems to have been the first to use the historical metaphor in the biological context. In his *Preliminary Discourse* of 1812, he described fossil bones as "historical documents" (Rudwick 1997, 205).

[9] Miguel García-Sancho (2009) has reconstructed Fred Sanger's work on the development of sequencing methods for proteins and nucleic acids, adopting the view of sequencing as a field of practice (instead of a mere technique). This approach seems to open the space for a more dynamical histori-

The first generation of molecular phylogeneticists already had realized that each single residue of a long biological macromolecule and not the entire protein, constituted a character. Thus, each amino acid or, eventually, each nucleotide composing a macromolecule could be present in different character states. The implication was that the quantitative comparison of several molecules, composed by tens or hundreds of these residues, required enormous computation capabilities which were impossible to perform by the individual scientist. Thus, the quantitative analysis of large data sets intersected in two obvious ways with the introduction of computers in the life sciences: by incorporating the new machines in universities and later at the laboratory bench (Lenoir 1999; Hagen 2001; November 2006)[10] and in the development of algorithms or programs (software packages) to perform the comparative analysis (Hagen 2000; Suárez and Anaya 2008).

The first quantitative methods for estimating the minimum distance between two molecules with the help of computers became possible once the number of available protein sequences increased (Fitch and Margoliash 1967; Dayhoff et al. 1965; Dayhoff 1969). One of the first computer algorithms for constructing phylogenetic trees was conceived by Walter Fitch, then at the University of Wisconsin.[11]

Fitch used the 20 sequences of cytochrome c obtained by biochemist Emanuel Margoliash of Abbot Laboratories in Chicago, to develop a "distance matrix method." Such a method seeks for a tree that best predicts a set of mutation distances between two molecules. Mutation distance was "defined here as the minimal number of nucleotides that would be needed to be altered in order for the gene for one cytochrome to code for the other." He added: "This distance is determined by a computer making a pair-wise comparison of homologous amino acids" (Fitch and Margoliah 1967, 280).

cal analysis on the development of these procedures and its relative transformation into automated procedures.

[10] The use of computers in biology as a project promoted by the NIH in the 1960s (Lenoir 1999; November 2006), or the impulse towards automation and development of computational tools related to the Human Genome Project are examples of this (see Kevles and Hood 1992). Nevertheless a wide variety of scientific research programs in traditional or new disciplinary domains between the 1960s and 1980s became targets of the computerization of biology. Conspicuously, one of these was the nascent field of studies on molecular evolution (see Hagen 2000; 2001; Suárez and Anaya 2008).

[11] Fitch's work became the basis upon which most sequence-analysis tools took place during the 1980s. But similar efforts in the direction of constructing phylogenetic trees had taken place almost simultaneously by population geneticists Anthony Edwards and Luigi Luca Cavalli-Sforza in the mid-1960s (Edwards and Cavalli-Sforza 1964; Cavalli-Sforza and Edwards 1967). However, instead of protein sequences they used data on blood group variation in human populations and methods of maximum likelihood.

Indeed, the development of Fitch's algorithm was made possible by the availability of small computers in American universities in the 1960s and their increased applicability for biological research. Most probably, Fitch used an IBM machine of the 360 Series bought by the University of Wisconsin (Fitch, personal communication 2007). And, as Hagen (2000) has pointed out, Fitch, like Dayhoff, used the FORTRAN programming language devised by IBM. A thorough account of how computers were introduced into systematic research in the 1960s has been given by Joel Hagen (2000; 2001). The proliferation of molecular data facilitated the intersection of mathematics and systematic research, allowing the development of a "statistical frame of mind" in the profession (Hagen 2000; Suárez and Anaya 2008). This is all the more important, given the strong rivalries between the cladistic, evolutionary, and pheneticist schools of taxonomic thought in the 1960s (Hull 1988). Regardless of the methodological commitments of each school, a pragmatic approach developed among the first practitioners of molecular phylogenetics, who enthusiastically undertook the development of tools for sequence comparison and analysis.

The Tools of Comparison

Although one of the first computer trees was made after protein sequence data, Fitch's algorithm incorporated a definition to measure the minimum distance between molecules in terms of nucleotide substitution. Previously, he had done research on the genetic code (Fitch 1966) and like many of his contemporaries, he held the idea that genes carried more phylogenetic information than proteins (Fitch 1995, personal communication). Hence, even when the first computer-based analysis and molecular databases depended on the available protein sequences and DNA sequencing looked well into the future, the most popular algorithms already addressed the task as an analysis of nucleotide differences.[12]

Very soon, scientists realized that the alignment of sequences for comparative purposes faced serious conceptual and practical problems, affecting later developments in bioinformatics. Such problems were characteristic of evolutionary biology. One was the definition of homology in the context of the comparison of molecules. I will not go into details here, since a thorough account on the way in which the concept of

[12] Margaret Dayhoff and her colleagues, however, developed a very useful matrix for amino acid substitutions or PAMs (Dayhoff et al. 1965). PAM matrixes assign values for probabilities of amino-acid substitutions and thus provide a way to understand the functional or biological (adaptive) value of protein evolution. PAM matrixes have been recently revived by the work of Martin Vingron and his colleagues at the Max Planck Institute for Molecular Genetics (see Suárez and Anaya 2008).

homology has been present in molecular phylogenetics would require a different paper.[13] Suffice it to say that in order to assign a hypothetical function to a sequence, the scientist must presuppose an ancestry relation with sequences whose function has been already recognized (by traditional genetic and biochemical methods).[14] Thus, regardless of their awareness of evolutionary concepts and their definition of them, the annotation of genomes requires the assumption of homology.

A different practical problem was faced when scientists attempted the alignment of sequences of different length. In order to give a quantitative measure of similarities (or differences) between two homologous molecules their sequences need to be compared "side by side." However, given their different lengths, the comparison requires the introduction of criteria for maximum match that, in turn, require the introduction of what they call *gaps* or *indels*. The phylogenetic tree constructed by Fitch and Margoliash in 1967 did not address the problem of alignment in an explicit way. During the 1970s and well into the 1980s many of these studies still made no use of computer algorithms, being restricted to a semi-quantitative comparison of sequences by aligning sequences "by the eye." Fitch's approach was based on the search for nonrandom alignments by comparing all possible combinations of sequences of a given length. Well until the 1970s, and even in the late 1980s, a common comparative approach was to align sequences "by the eye." This practice has been considered to introduce subjective criteria to sequence comparisons, however it is still common to argue that biological criteria may involve fixing the computer alignments "by the eye" (Suárez and Anaya 2008).

The first computer algorithm for sequence alignment was developed by Saul B. Needleman and Christian Wunsch (Needleman and Wunsch 1970). In the early 1960s Needleman had collaborated with Emmanuel Margoliash on the semi-quantitative comparative analysis of rabbit cytochrome c, at the Biochemical Research Department of Abbott

[13] For more references and a brief account of the positions held by molecular phylogeneticists see Suárez and Anaya (2008).

[14] The first quantitative analysis of hemoglobins and cytochromes c in the 1960s assumed that the molecules compared were homologous; that is, that they shared a common ancestor. But homology was inferred on the basis of similarity. This position raised the critiques and concerns of biologically-minded phylogeneticists (for instance, Fitch 1970; 2000), who do not equate similarity (which can be attained by adaptive *convergence*) with homology. However, more often than not, homology continues in practice to be inferred on the basis of sequence similarity. In the context of molecular evolution, Zuckerkandl and Pauling (1965b) suggested the importance of gene duplication (first proposed by Haldane) to account for sequence similarities. Later on, Walter Fitch's distinction between *orthologous* and *paralogous* molecules became part of the genetic analysis of homology and became a crucial conceptual tool in functional genomics. Homologous sequences are orthologous if they were separated by a speciation event; while paralogous molecules were separated by a gene duplication event and thus occupy two different positions in the genome (Fitch 1970; 2000).

Laboratories in North Chicago (Margoliash, Needleman and Stewart 1963). Also, in 1969, Needleman and Blair had published an analysis on the evolution of cytochrome c, in which an analysis of sequences and early attempts to develop alignment methods were applied to the origins of eukaryotic cells. Needleman's evolutionary interests are reflected in his and Wunsch's algorithm, which incorporates biological considerations that most recent programs aim to eliminate.[15]

Using dynamic programming, Needleman and Wunsch relied on Fitch's approach, which calculates the minimum number of mutations needed for a given substitution. As mentioned above, a similar approach had been developed by Dayhoff and Ecke for the analysis of data on protein sequences at the National Biomedical Research Foundation. However, Dayhoff and Ecke's mutation matrices (known as PAM) calculated the probability of one amino acid to be substituted by another amino acid in a protein chain. Such matrices have remained useful to detect homology, in the alignment of sequences and other problems (Hagen 2000), but their utility has remained linked to evolutionary problems. Instead, Needleman and Wunsch's tool, as they defined it in 1969, was "a computer adaptable method for finding similarities in the amino acid sequences of two proteins." Clearly, their program had been developed in the context of evolutionary problems and with the purpose of solving phylogenetic questions, but it could be more generally extended for analytical purposes. Shortly afterwards, Sellers (1974) developed a similar algorithm that instead of maximizing similarities, as Needleman's and Wunsch's proposal was designed to minimize the differences between two molecules (in the context of what is called global alignments, namely, tools that compare complete molecules). A few years later, in the first of many future collaborations, phyisicist Temple Smith and mathematician Michael Waterman from Los Alamos National Laboratory at New Mexico (see below), together with Walter Fitch, proved that under certain circumstances both algorithms were equivalent (Smith Waterman and Fitch 1981). Thus, the basic tool for global alignment of sequences is often known as the Needleman-Wunsch-Sellers algorithm.

[15] Saul Needleman's professional career looks a little bit erratic. Trained as a biochemist, he contributed to protein sequence techniques while at Abbot Laboratories and then at Northwestern University at Chicago. Then in the late 1980s, working at the Navy Drug Screening Laboratory at the Great Lakes, he focused on research on drug addictions (Needleman 1990). Needleman was considered a "medical legal chemist" and he used his biochemistry and mathematical skills to develop blood tests for several drugs and to analyze the distribution of drug addictions among different army corps.

As gene and protein databases grew and diversified during the decade of the 1980s as a result of improvements in sequencing techniques and automation, more efficient methods of comparative analysis were required. In 1980, within the context of building a national nucleic acid sequence database at Los Alamos National Laboratory, such sequence analysis tools seemed urgent, if one did not want to get swamped amidst the ocean of data. Waterman proposed a metric of similarity between molecular sequences and he, together with Smith, developed their algorithm for searching similarities in databases (Smith and Waterman 1981). Both scientists were part of the Theoretical Biology group under the lead of physicist Walter Goad and had a very specific purpose: to search sequences by finding sequence similarities at large databases such as GenBank, created in 1982 under Goad's leadership (Smith 1990; Strasser 2008). This bioinformatics tool is a variation of the Needleman-Wunsch algorithm, but instead of looking for global alignments, it searches for local alignments. That is, the Smith-Waterman tool compares segments of all possible lengths, optimizing also the similarity measure.

It is not my intent to detail the way in which the Waterman-Smith algorithm performs. What is interesting is that both biological (evolutionary) and statistical considerations are part of the algorithm: it searches for correct alignments in regions of low similarity between distantly related biological sequences. Like its precursor, the Needleman-Wunsch algorithm, the Waterman-Smith uses dynamic programming, which has the advantage that it seeks to find the optimal local alignment with respect to the scoring system that is used, for instance the substitution matrix (PAM) and the gap scoring system (see Waterman 1984; Suárez and Anaya 2008).

The Smith-Waterman algorithm is the basis for many sequence comparison programs. It does not necessarily meet the needs and criteria of evolutionary biologists, since it looks for patterns, instead of overall similarities between molecules.

Thus, by the mid 1980s, in a review of developments in dynamic programming and comparison of macromolecules written by Waterman, the focus was on the usefulness of these tools for rapid database searches, instead of searching for evolutionary relations (Waterman 1984, 474). Waterman, by then holding a joint appointment in the Department of Molecular Biology and the Department of Mathematics at the University of Southern California, noticed the importance of these analytical tools in the wake of the several international databases that were being created around the world (in Japan and Europe) to organize the information available (1984, 474).

Today, the most popular tools for finding similarities in a database are BLAST (with several modifications and adaptations) and FASTA. The paper presenting BLAST, published in 1990 by David Atschul and collaborators at the NCBI (National Center for Biological Information, a branch of the NIH), was the most quoted paper in that decade. Like the Smith-Waterman algorithm, BLAST is restricted to find regions of similarity between biological sequences not global alignments between pairs of complete molecular sequences. However, instead of using the exhaustive approach of the Smith-Waterman tool, BLAST uses a heuristic approach that approximates the Smith-Waterman algorithm. BLAST and FASTA emphasize speed over sensitivity, reflecting the new realities of large database searches related to the Human Genome Project (started in 1988). BLAST, for instance, is over 50 times faster than the previous methods of dynamic programming. Also, BLAST is faster than FASTA, since it only seeks for the most significant patterns in similarity. In fact, BLAST is a family of programs. In searching the NCBI webpage, there are different alternatives: BLAST for search in a nucleotide database using a nucleotide query, search in a protein database using a protein query, search of a protein database using a translated nucleotide query, etc. However, BLAST cannot guarantee the optimality of sequence alignments of basic dynamic programming, which does incorporate biological (evolutionary) as well as statistical criteria. The heuristics used in these searching tools gives solutions which are good enough, but the alignments do not necessarily provide homologous sequences. The criteria used in these programs, however, may include the weighting of gaps or indels required to perform an alignment of sequences of different length. Thus, for certain specific tasks within genomics, scientists still need the sequence analysis tools that incorporate evolutionary considerations.

Concluding Remarks

Bioinformatics today comprises an enormous and growing field of applications and practices. Some of its most conspicuous uses are displayed on the "ultra-discipline" of genomics, but also fields like biogeography and population biology. These range from the creation and development of databases, to the advancement of computational and statistical techniques in order to address practical and theoretical problems in the analysis of biological information. This means that a growing number of bioinformatics applications have nothing or almost nothing to do with evolution. Other areas of bioinformatics and

genomics, by contrast, still seem intimately linked to evolutionary concerns: gene location and sequence alignment, and many problems of functional and comparative genomics (including the annotation of genomes) illustrate this point. Such areas of practice incorporate assumptions about the evolutionary process and sometimes they even incorporate evolutionary models, as in the case of calculations for multiple-hit events in nucleotide substitution.

As historians, we need to make room for those fields, actors and tools that have been systematically marginalized in sociologically oriented and in actors' accounts of genomics, but not just for a "sense of justice." In paying attention to these marginalized themes, we are guided toward problems that demand the attention of a new generation of historians; that is, to the role of automation in biology, including the automation of data collection (for instance, the robotization of molecular or gene libraries), and the automation of statistical inferences, entrenched in software packages. Also, the intersection of biology and mathematics, and in particular statistics, opens demanding questions concerning the nature of biological knowledge and the transformation of scientific attitudes and practices in biological research (Hagen 2001; Suárez and Anaya 2008). These two broad questions constitute important areas of research that are specific to genomics and bring us back to the material and conceptual components of scientific practice.

However, a big challenge arises for historians of science wanting to understand the origins and development of genomic tools. Maybe it is not a new type of challenge, but its dimension seems insurmountable at this point. I am referring to the large number of software programs and families of programs that overwhelm the student of science who pretends to get a glance of this field.[16]

The esoteric character of this field, due to the black-boxing of programming practices into software packages, makes it unwise to attempt a reconstruction of its scientific development.

Bioinformatics poses specific difficulties for historians-of-molecular-biology-turned-historians-of-genomics and it seems to require the cooperation of a very rare species, the historian of computer sciences interested in biology. Moreover, in the case of bioinformatics and genomics we are dealing with a collective or network-situated knowledge, which reflects the transformations of biological research at the

[16] For instance, Joe Felsenstein's personal open-source webpage – probably the most comprehensive in its kind – includes 385 software packages and 52 free servers, to provide analytical tools for molecular phylogeneticists alone. See: http://evolution.genetics.washington.edu/phylip/software.html (August 31, 2009).

beginning of the 21st century. This fact presents historians with new challenges as to the sources of her/his research.

Finally, a word on genres in the writing of history. The selective nature of the human memory is always surprising. It reminds us of how individuals construe their own personae everyday, rebuilding the image they have of themselves and the image they want others to have of them. Some months ago, Vincent Ramillon came to my office at the MPIWG with exciting news. In the 1980s Venter was sequencing neurotransmitters at his lab at the National Institute of Neurological and Communicative Disorders and Stroke at the NIH. But he was doing more than that. He was well involved in the phylogenetics of neurotransmitters and, moreover, he addressed his research with an evolutionary perspective that is far removed from the reductionist rhetoric and the Cinderella/maverick plot of his later accounts on the HGP (Venter 2007). In a review he wrote of his field as late as 1988 he claims:

> The findings (of phylogenetic analysis) so far indicate that the evolution and development of the nervous system was not dependant upon the formation of new or better transmitter substances, receptor proteins, transducers and effector proteins but involved better utilization of these highly developed elements in creating advanced and refined circuitry. This is not a new concept . . . . In a 1953 article discussing chemical aspects of evolution (Danielli 1953) Danielli quotes Medawar, "endocrine evolution is not an evolution of hormones but an evolution of the uses to which they are put; an evolution not, to put it crudely, of chemical formulae but of reactivities, reaction patterns and tissue competences . . . . Evolution is the history of changing uses of molecules, and not of changing synthetic abilities (Danielli 1953). (Venter et al. 1988, 151-152)

Venter published on the phylogenetics of neurotransmitters and the evolution of the nervous system for almost a decade in the 1980s, while his research focused on the adrenalin receptors. Not a single word about his evolutionary interests is ever mentioned in his memoirs (Venter 2007), though he makes reference to evolutionary processes at least three times. The first was to account for "what (makes) us uniquely human" (Venter 2007, 323), despite the large sets of genes we share with the fly. The other two refer to his recent work in synthetic biology: "I want to take us far from shore into unknown waters, to a new phase of evolution, to the day when one DNA-based species can sit down at a computer to design another. I plan to show that we understand the software of life by creating true artificial life" (Venter 2007, 357).

Revisiting his days at the State University of New York at Buffalo, and at the National Institute of Neurological Disorders and Stroke, in Bethesda, Maryland, he does not mention his acquaintance with the tools of evolutionary analysis and comparison at the time. This, despite the fact that such techniques were instrumental in his early research and might have allowed him the familiarity to devise computer strategies for assembling whole genomes. All of which reminds us of the perils of relying on the actors' accounts but, also, of the meaningful areas of historical research opened up by focusing on the more mundane aspects of science.

Acknowledgements

I would like to thank Conacyt and UNAM for generously supporting the research for this paper through projects #J50219-H and PAPIIT IN308208. Also, I would like to thank the Proalmex grant "Heredity and Evolution: Genetic and Epigenetics" (funded by Conacyt and the DAAD), which made possible a research stay at the Max Planck Institute for the History of Science, in Berlin, during November-December, 2009.

References

Abir-Am P.G., 1985, "Themes, Genres and Orders of Legitimation in the Consolidation of New Scientific Disciplines: Deconstructing the Historiography of Molecular Biology", *History of Science*, 23(1): 73-117.

Abir-Am P.G., 1991, "Noblesse Oblige: Lives of Molecular Biologists", *Isis*, 82(2): 326-343.

Abir-Am P.G. 1992, "The Politics of Macromolecules: Molecular Biologists, Biochemists and Rhetoric", Osiris, 2nd Series, 7: 164-191.

Ankeny R., 2001, "The Natural History of *Caenorhabditis elegans* Research", *Nature Review Genetics*, 2: 474-476.

Ankeny R., 2010, "Historiographic Reflections on Model Organisms: Or, How the Mureaucracy May Be Limiting Our Understanding of Contemporary Genetics and Genomics", *History and Philosophy of the Life Sciences*, 32(1): 91-104.

Beatty J., 2000, "Origins of the U. S. Human Genome Project: Changing Relationships between Genetics and National Security". In: Sloan P. (ed.), *Controlling our Destinies: The Human Genome Project from Historical, Philosophical, Social and Ethical Perspectives*, Notre Dame: University of Notre Dame Press, 131-153.

Bishop J.M., 2004, *How to Win the Nobel Prize. An Unexpected Life in Science*, Cambridge: Harvard University Press.

Brandt C., 2005, "Genetic Code Text and Scripture: Metaphors and Narration in German Molecular Biology", *Science in Context*, 18(4): 629-648.
Bronwyn P., 2004, *Trading the Genome: Investigating the Commodification of Bio-information*, New York: Columbia University Press.
Brush S.G., 1995. "Scientists as Historians", *Osiris*, *2nd Series*, Vol. 10, *Constructing Knowledge in the History of Science*, 215-231.
Buckingham W., 1961, *Automation. Its Impact on Business and People*, New York: The New American Library.
Cairns J., Stent G.S. and Watson J.D. (eds), 1966, *Phage and the Origins of Molecular Biology*, Cold Spring Harbor: Cold Spring Harbor Laboratory Press.
Cantor C.R., 1992, "The Challenges to Technology and Informatics". In: Kevles D. and Hood L. (eds), *The Code of Codes*, Cambridge: Harvard University Press, 83-97.
Cantor C.R. and Smith C.L., 1999, *Genomics: The Science and Technology Behind the Human Genome Project*, New York: Wiley-Interscience.
Cavalli-Sforza L.L. and Edwards A.W.F., 1967, "Phylogenetic Analysis: Models and Estimation Procedures", *Evolution*, 21(3): 550-570.
Claverie J.M. and Notredame C., 2007, *Bioinformatics for Dummies.* 2nd edition. Indianapolis: Wiley Publishers.
Collins F., 2007, *The Language of God: A Scientist Presents evidence for Belief*, New York: Free Press.
Collins F., 2010, *The Language of Life: DNA and the Revolution in Personalized Medicine*, New York: Harper.
Crick F., 1988, *What Mad Pursuit. A Personal View of Scientific Discovery*, New York: Basic Books.
Cook-Deegan M.R., 1994, *The Gene Wars: Science, Politics and the Human Genome*, New York: Norton & Company.
Dayhoff M.O., 1969, Computer Analysis of Protein Evolution, *Scientific American*, 221: 86-95.
Dayhoff M.O., Eck R.V., Chang M.A. and Sochard M.R., 1965, *Atlas of Protein Sequence and Structure*, Silver Spring: National Biomedical Research Foundation.
De Chadarevian S., 1996, "Sequences, Conformation, Information: Biochemists and Molecular Biologists in the 1950s", *Journal of the History of Biology*, 29: 361-386.
De Chadarevian S., 2002, *Designs for Life: Molecular Biology after World War II*, Cambridge: Cambridge University Press.
De Chadarevian S., 2004, "Mapping the Worm's Genome: Tools, Networks, Patronage". In: Gaudillière J.P. and Rheinberger H.J. (eds), *From Molecular Genetics to Genomics: The Mapping Cultures of Twentieth Century Genetics*, New York: Routledge, 95-110.
De Chadarevian S. and Gaudillière J.P., 1996, "The tools of the discipline: biochemists and molecular biologists," *Journal of the History of Biology, 29:* 327-330.
De Chadarevian S. and Rheinberger H.J. (eds), 2006, *History and Epistemology of Molecular Biology and Beyond: Problems and Perspectives,* Berlin: Max Planck Institute for the History of Science. Preprint Number 317.

De Chadarevian S. and Rheinberger H. J. (eds), 2009, "Disciplinary Histories and the History of Disciplines: the Challenge of Molecular Biology," *Studies in History and Philosophy of Biological and Biomedical Sciences,* 40(1), Special Issue.

Dietrich M.R., 1998, "Paradox and Persuasion: Negotiating the Place of Molecular Evolution within Evolutionary Biology", *Journal of the History of Biology*, 31: 85-111.

Durand D., 1997, "A New Look at Three Models for Multiple Sequence Alignment". In: Farach M., Roberts F., Vingron M. and Waterman M. (eds), *DIMACS Special Year Volume, Mathematical Support for Molecular Biology*, 65-84.

Edman P. and Begg G., 1967, "A Protein Sequenator", *European Journal of Biochemistry*, 1: 80-91.

Edman P., 1970, "Sequence Determination", *Molecular Biology, Biochemistry and Biophysics,* 8: 211-255.

Fitch W.M., 1970, "Distinguishing Homologous from Analogous Proteins", *Systematic Zoology*, 19: 99-113.

Fitch W.M., 1971, "Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology", *Systematic Zoology*, 20(4): 406-416.

Fitch W.M., 1988, "This Week's Citation Classic", *Current Contents*, 27: 16.

Fitch W.M., 2000, "Homology, a Personal View on Some of the Problems", *Trends in Genetics*, 16(5): 227-231.

Fitch W.M. and Smith T.F., 1983, "Optimal Sequence Alignments", *Proceedings of the National Academy of Sciences, USA*, 80: 1382-1386.

Fitch W.M. and Margoliash E., 1967, "Construction of Phylogenetic Trees", *Science* 155(3760): 279-284.

Fitch W.M. and Smith T.F., 1983, "Optimal Sequence Alignments", *Proceedings of the National Academy of Science*, 80: 1382-1386.

Forman P., 1991, "Independence, Not Transcendence, for the Historian of Science", *Isis*, 82(1): 71-86.

Fortun M., 2008, *Promising Genomics: Iceland and deCODE. Genetics in a World of Speculation*, Berkeley: University of California Press.

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

Fujimura J.H., 1999, "The Practices of Producing Meaning in Bioinformatics". In: Fortun M. and Mendelsohn E. (eds), *The Practices of Human Genetics*, Dordrecht: Kluwer Academic Publishers, 49-87.

García-Sancho M., 2009, "A New Insight into Sanger's Development of Sequencing: from Proteins to DNA, 1943-1977", *Journal of the History of Biology, Published online, 23 aprile 2009.*

Gaudillière J. P., 1996, "Molecular biologists, biochemists and messenger RNA: The birth of a scientific network," *Journal of the History of Biology,* 29: 417-445.

Gibbons M. et al., 1994. *The New Production of Knowledge: The Dynamic of Science and Research in Contemporary Societies*, London: Sage.

Hagen J.B., 1999, "Naturalists, Molecular Biologists and the Challenges of Molecular Evolution", *Journal of History of Biology*, 32: 321-334.

Hagen J.B., 2000, "The Origins of Bioinformatics", *Nature Reviews Genetics*, 1: 231-237.

Hagen J.B., 2001, "The Introduction of Computers into Systematic Research in the United States During the 1960s", *Studies in the History and Philosophy of Biological and Biomedical Sciences*, 32(2): 291-314.
Hilgartner S., 2004, "Making Maps and Making Social Order: Governing American Genome Centers, 1988-93". In: Gaudillière J.P. and Rheinberger H.J. (eds), *From Molecular Genetics to Genomics: The Mapping Cultures of Twentieth Century Genetics*, New York: Routledge, 113-127.
Holmes F.L., 2006, *Reconceiving the Gene: Seymour Benzer's Adventures in Phage Genetics*, edited by Summers W.C., New York: Yale University Press.
Hull D.L., 1988, *Science as a Process*, Chicago: The University of Chicago Press.
Judson H.F., 1979, *The Eighth Day of Creation*, New York: Simon and Schuster.
Jacob F., 1988, *The Statue Within: An Autobiography*, New York: Basic Books.
Kaufmann A., 2004, "Mapping the Human Genome at Généthon Laboratory: The French Muscular Dystrophy Association and the Politics of the Gene". In: Gaudillière J.P. and Rheinberger H.J. (eds), *From Molecular Genetics to Genomics: The Mapping Cultures of Twentieth Century Genetics*, New York: Routledge, 129-157.
Kay L.E., 2000, *Who Wrote the Boook of Life? A History of the Genetic Code*, Stanford: Stanford University Press.
Kevles D. and Hood L. (eds), 1992, *The Code of Codes*, Cambridge: Harvard University Press.
Lenoir T., 1999, "Shaping Biomedicine As an Information Science", *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems*, 27-45.
Margoliash E., 1963, "Primary Structure and Evolution of Cytochrome c", *Proceedings of the National Academy of Sciences, USA*, 50(4): 672-679.
Margoliash E., Needleman S.B. and Stewart J.W., 1963, "A Comparison of the Amino Acid Sequences of the Cytochrome c Several Vertebrates", *Acta Chemica Scandinavia*, 17: S250.
Monod J., 1972, *Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology*, New York: Vintage Books.
Moore S. and Stein W., 1972, "The Chemical Structure of Pancreatic Ribonuclease and Deoxyribonuclease", Nobel Lecture. Wwwnobel.se/chemistry/laureates/1972/moore-lecture.html
Morange M., 1998, *A History of Molecular Biology*, Cambridge: Harvard University Press.
Needleman S., and Wunsch C., 1970, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of two Proteins", *Journal of Molecular Biology*, 48(3): 443-53.
November J.A., 2006, *Digitalizing Life: The Introduction of Computers to Biology and Medicine,* Unpublished PhD Dissertation, Princeton University.
Nowotny H., Scott P. and Gibbons M., 2001, *Re-Thinking Science: Knowledge and the Public in an Age of Uncertainty*, Cambridge: Polity Press.
Olby R.C., 1974, *The Path to the Double Helix*, Seattle: University of Washington Press.

Ramillon V., 2007, *Les deux génomiques. Mobiliser, Organiser, Produire: du séquençage à l'expression de la mésure des gènes*, Unpublished PhD dissertation, École des Hautes Études, Paris, France.
Reardon J., 2004, *Race to the Finish. Identity and Governance in an Age of Genomics*, New Jersey: Princeton University Press.
Rudwick M.J.S., 1997, *Georges Cuvier, Fossil Bones and Geological Catastrophes. New Translations and Interpretations of the Primary Texts*, Chicago: The University of Chicago Press.
Sellers P.H., 1974, "On the Theory and Computation of Evolutionary Distances"**,** *Journal of Applied Mathematics (Siam)*, 26: 787-793.
Smith T.F., 1990, "The History of the Genetic Sequence Databases", *Genomics*, 6: 701-707.
Smith T.F. and Waterman M.S., 1981, "Identification of Common Molecular Subsequences", *Journal of Molecular Biology*, 147: 195-197.
Smith T.F., Waterman M.S., Fitch W.M., 1981, "Comparative Biosequence Metrics"**,** *Journal of Molecular Evolution*, 18: 38-46.
Sommer M., 2008, "History in the Gene: Negotiations between Molecular and Organismal Anthropology", *Journal of the History of Biology*, 41(3): 473-528.
Stent G.S., 1968, "That Was the Molecular Biology That Was", *Science*, 160: 390-395.
Stent G.S., 1969, *The Coming of the Golden Age. A View of the End of Progress*, New York: The Natural History Press.
Stent G.S., 1998, *Nazis, Women and Molecular Biology. Memoir of a Lucky Self-Hater*, New Jersey: Transaction Publishers.
Strasser B., 2008, "GenBank – Natural History in the 21st Century?", *Science*, 322: 537-538.
Suárez E., 2001, "Satellite-DNA: A Case Study for the Evolution of Experimental Techniques", *Studies in History and Philosophy o,f Biological and Biomedical Sciences*, 32(1): 31-57.
Suárez E., 2007, "The Rhetoric of Informational Molecules. Authority and Promises in the Early Days of Molecular Evolution", *Science in Context*, 20(4): 649-677.
Suàrez-Díaz E., 2009, "Molecular evolution: concepts and the origins of disciplines," *Studies in history and Philosophy of Biological and Biomedical Sciences,* 40(1): 43-53.
Suárez E. and Anaya V., 2008, "History, Objectivity and the Construction of Molecular Phylogenies", *Studies in the History and Philosophy of the Biological and Biomedical Sciences*, 38: 451-468.
Sulston J. and Ferry G., 2002, *The Common Thread: A Story of Science, Politics, Ethics and the Human Genome*, Washington D.C.: Joseph Henry Press.
Sunder-Rajan K., 2006, *Biocapital. The Constitution of Postgenomic Life*, Durham: Duke University Press.
Thacker E., 2005, *The Global Genome. Biotechnology, Politics and Culture*, Cambridge: MIT Press.
Venter C., 2007, *A Life Decoded: My Genome: My Life*, New York: Penguin Books.
Venter J. et al., 1988. "Evolution of Neurotransmitter Receptor Systems", *Progress in Neurobiology*, 30: 105-169.

Waterman M.S., 1984, "General Methods of Sequence Comparison", *Bulletin of Mathematical Biology*, 46(4): 43-50.
Watson J.D., 1968, *The Double Helix*, New York: Penguin Books.
Wimsatt W.C. and Griesemer J., 1989, "Picturing Weissmanism: a Case Study of Concept Evolution". In: Ruse M. (ed.), *What the Philosophy of Biology Is. Essays for David Hull*, Dordrecht: Kluwer, 75-137.
Xiong J., 2006, *Essential Bioinformatics*, New York: Cambridge University Press.
Zuckerkandl E., and Pauling L., 1962, "Molecular Disease, Evolution and Genic Heterogeneity". In: Kasha M., Pullman B. (eds), *Horizons in Biochemistry: Albert Szent-Györgyi Dedicatory Volume*, New York: Academic Press, 189-225.
Zuckerkandl E. and Pauling L., 1965a, "Molecules as Documents of Evolutionary History", *Journal of Theoretical Biology*, 8: 356-366.
Zuckerkandl E. and Pauling L., 1965b, "Evolutionary Divergence and Convergence in Proteins". In: Bryson V. and Vogel H. (eds), *Evolving Genes and Proteins*, New York: Academic Press, 97-166.