Hist. Sci., xlix (2011)

THE COMPARATIVE AND THE EXEMPLARY: REVISITING THE EARLY HISTORY OF MOLECULAR BIOLOGY

Bruno J. Strasser
Yale University

Soraya de Chadarevian
UCLA

INTRODUCTION

The image and caricature of the biologist roaming the field with a collecting box or poring over pinned up insect and butterfly cases endured well into the twentieth century. In contrast the molecular vision of life, and even more so molecular biology, the quintessential science of the late twentieth century, has most often been connected with experimenting and intervening on a handful of model organisms and systems. Underpinning this widespread opposition is the idea that biology as practised by naturalists was "merely" descriptive, systematic, and comparative. Naturalists would rely on collecting and comparing specimens, providing names and descriptions, and thereby document biological diversity. This way of approaching the natural world was supposedly superseded by the experimental approach in the late nineteenth century, with molecular biologists taking up that banner in the twentieth. Using ever more powerful instruments, imported from the physical and chemical sciences, and focusing their attention on a few well-chosen model systems, they studied processes and their underlying mechanisms. This view of the history has been promoted by latter-day biologists and by historians alike.[1] Historians of molecular biology have insisted, for example, on the crucial role of the Rockefeller Foundation, since the 1930s, in promoting molecular approaches in the life sciences by providing funding for the acquisition of costly new physical and chemical instruments.[2] The molecular vision of life was predicated upon sophisticated means of intervention which produced representations enabling further manipulations of life.[3] Equal attention has been dedicated to the strategic choice of a few simple model organisms on which much of the early work in molecular biology was developed. We can think here of Max Delbrück's introduction of phage for the study of genetics, and the extensive use of bacteria and viruses in studying genetic mechanisms in molecular terms, but also of moulds (*Neurospora crassa*), flies (*Drosophila melanogaster*), and later, worms (*Caenorhabditis elegans*), mice (*Mus musculus*), and weeds (*Arabidopsis thaliana*), which all gained the enviable status (at least from the researcher's perspective) of model organism for the production of knowledge in molecular biology.[4]

One of the historiographic side-effects of writing disciplinary histories, such as those of molecular biology mentioned previously, has been a tendency to emphasize

0073-2753/11/4903-0317/$10.00

cognitive, methodological, and sociological unity within the disciplines and differences among disciplines. By contrast, John Pickstone's call to focus on "ways of knowing" (or "working knowledges"), rather than disciplines, can help us make visible the heterogeneity of cognitive and material practices within disciplines and the similarities among disciplines. Indeed, according to Pickstone, even though ways of knowing have their own historicity and have, for example, enjoyed their greatest successes at different times, they do not replace each other, like Kuhnian paradigms, but add new layers in the makeup of science, technology and medicine.[5] Pickstone's historiographic approach, unlike that of many others proponents of 'styles', is not taxonomic, but analytic; it reveals the different components which make up scientific practices. This perspective proves fruitful for a closer examination of the working practices of modern biologists. As we show in this paper, molecular biology did not take shape exclusively as an experimental science focused on 'exemplary' model organism and systems. Rather, much work in molecular biology can be described more accurately as comparative and relied far more than previously recognized on collections. This is true not only for late twentieth-century genomic scientists, who have been derided as "molecular birdwatchers",[6] but also of early day molecular biologists, those most vehement in condemning natural history.[7] In their studies of the most challenging problems of the new science, namely the structure and function of proteins and nucleic acids and the deciphering of the genetic code, they relied heavily on collections and comparisons of molecular data, the distinctive approach of the comparative biological sciences, usually associated with nineteenth-century natural history, anatomy, and embryology. Distinguishing exemplary and comparative practices not only throws new light on the work practices of molecular biologists, but it also challenges the usefulness of the very distinction between experimentalism and natural history, which has structured so many histories of the life sciences.[8]

In his book and later articles, Pickstone has defined different ways of knowing, of which the central elements are "meanings", "natural history", "analysis", and "synthetic experimentation". For our purposes, "analysis" (the attempt to understand compound elements in terms of their individual parts) is particularly useful in that it subverts the distinction between ("mere") description, attributed to natural history, and experimentation, supposedly the hallmark of any truly modern science. This distinction has been a powerful rhetorical tool in the hand of the advocates of experimentalism in their quest for institutional authority, but has obscured our understanding of some significant continuities in research practices, such as the ones we highlight in this paper. Pickstone's "analytical" sciences have been concerned with "deconstructing" organisms, conceptually or materially, into components, such as organs, tissues, cells, molecules or genes.[9] Yet, the analytical sciences, of which molecular biology is certainly a good example, have operated according to at least two distinct epistemic practices that we propose to call "comparative" and "exemplary". We would like to introduce them as subdivisions of Pickstone's category of "analysis". Before examining in some details three key episodes in the history of

"classical molecular biology" to show how comparative and exemplary practices have played out, we start by briefly discussing how we understand Pickstone's categories.

WAYS OF KNOWING AND DOING

Pickstone's ways of knowing are part of a group of categories, including Ludwik Fleck's "Denkstil", Thomas Kuhn's "paradigms", Michel Foucault's "episteme", Gerald Holton's "themata", and Alistair Crombie's "styles", which have been devised to describe the historical development of science in methodological terms, but with more nuance than a unique and atemporal "scientific method" would allow.[10] Pickstone's "ways of knowing" are closest to Crombie's "styles" which, as Hacking pointed out, are cognitive practices but include more than simple rules of inference. They produce specific scientific objects and kinds of scientific evidence.[11] This explains, according to Hacking, why styles have been so enduring. Even though they designate historical practices, which can cease to exist, they are particularly robust due to their "self-authenticating" character.[12] Claims made by a science operating according to a particular style apply only to objects defined in that style and their validity can only be evaluated according to the evidence produced in that style. This, like the incommensurability of Kuhn's paradigms, gives styles their historical stability.

One may take issue with Pickstone's specific ways of knowing, their naming, and their exact boundaries, but still recognize the value of understanding science, technology, and medicine in terms of (some kinds of) ways of knowing. Furthermore, the different ways of knowing defined by Pickstone are constrained by his attempt to have them match, chronologically and conceptually, to "ways of working" in society at large. This ambitious intellectual agenda, absent in Crombie for example, would deserve a discussion of its own, and again, one does not need to endorse this particular argument to use Pickstone's categories productively.

We view critically, for example, some of the names Pickstone gives to his categories. Calling a way of knowing "natural historical" somewhat defeats the purpose of the category which is to allow for the exploration of different ways of knowing within disciplines, including natural history. In Pickstone's terms, natural history was all "natural historical" in the eighteenth century, but in the nineteenth some of it was also "analytic" — which is somewhat confusing. Similarly, Pickstone's use of the term "experimental" to designate a way of knowing was in some respects unfortunate because, as he makes clear, nineteenth-century "analytic sciences" (say analytic chemistry), or even natural history, were sometimes carried out in ways which could be called experimental, and many sciences which historical actors designated as experimental, such as biochemistry, were mainly "analytic" for Pickstone. In the present volume's Introduction and especially the Afterwords, Pickstone clarifies his initial reasons for using actors' categories to designate ways of knowing. He also proposes to call them now "reading, sorting, analysis and synthesis", a helpful correction we believe. Indeed, if one wants to propose historiographic categories which depart significantly from the historical actors' use of the terms, it is better to name them differently, or confusion will inevitably result.

In our paper, we will focus on two scientific approaches that we propose to designate as "comparative" and "exemplary". Both have played out in Pickstone's analytical sciences, but also in the other ways of knowing he proposes. It comes as no surprise, however, that Pickstone chose the analytical sciences to discuss the comparative approach, since it had been particularly prominent there.[13] The "comparative analysis" practices, based on the collection and comparison of the parts of a wide range of organisms, have enjoyed a wide popularity in zoological museums and botanical gardens since the early nineteenth century in a number of scientific endeavours referred to as natural historical. Yet comparative analysis has also been an essential practice associated with experimental work, for example in comparative physiology, comparative embryology, or, as we show in this paper, in molecular biology. In a trivial sense, all sophisticated experimental work is comparative, because researchers compare the result obtained from a control and those from a manipulated system. Together with the historical actors of the previously mentioned disciplines, we use the term in a much more specific sense, to designate cases where researchers compared variations present in nature (in different species, populations, or individuals), not variations induced by the experimenter. Yet not all biological experimentation has been comparative. Often researchers have focused on particular model organisms to analyse characteristics which seem to be common to many, if not all, organisms. We refer to this approach as "exemplary".[14]

The crucial difference between the comparative and the exemplary approaches resides in how their local claims are made universal.[15] In the comparative approach, it is the systematic comparison of a wide diversity of cases (or species in biology) that reveals regularities which are turned into universal claims. In the exemplary approach, the results obtained from a single case, deemed to be exemplary, are taken to have universal validity, or as the molecular biologists Jacques Monod and François Jacob famously put it, "what is true of E. coli is true of the elephant".[16] Researchers sharing this point of view have often criticized the comparative approach as simply "natural historical", i.e. aimed at showing the range of different biological phenomena present in different species. Yet comparative practices have also been used to uncover regularities and produce universal (or at least general) knowledge. Here, the exemplary and the comparative have had the same aim (universal knowledge), have been carried out in the same place (the laboratory), with the same kind of data (results of experimental analysis), but have proceeded through a different epistemic route.

The history of molecular biology has often been told as a story of experimentalism triumphant, illustrated by the elucidation of the structure and function of proteins and the deciphering of the genetic code. Pickstone has already pointed out that much of molecular biology can better be viewed as analytical rather than experimental. Building on Pickstone's work, we will show here that the history of molecular biology can usefully be understood as mobilizing both comparative and exemplary practices. Some of molecular biologists' greatest intellectual achievements rested not only on experimental ingenuity, but also on the constitution of collections, from a broad range of organisms or mutants, and on the systematic comparison of their

elements.[17] Since exemplary practices, involving unique model systems and organisms such as viruses, worms, and mice, have attracted much more attention, we will focus here on comparative practices which have been largely neglected and their relation to exemplary practices. The following historical examples on the study of protein structure, protein function, the genetic code, and molecular development should provide ample evidence for the usefulness of this perspective.

A COMPARATIVE ANATOMY OF PROTEINS

The study of the structure and function of proteins occupied a central place in the early history of molecular biology. While biochemists tended to study the function of proteins by determining the flow of energy and reaction products in enzyme-catalyzed reactions, molecular biologists embraced the structural study of proteins, including enzymes, as key to their function. In doing so, they did not only explore "exemplary" proteins, such as insulin, haemoglobin, and ribonuclease, extracted from a single species, but determined a wide range of protein structures from many species and compared their results systematically. In so doing, they were not adopting outdated natural historical approaches. Rather their work was widely viewed as cutting edge biology. Natural history and the modern experimental sciences have thus both resorted to comparative practices.

Research on protein sequences as a means to understand protein functions perfectly illustrates the combination of exemplary and comparative approaches. The idea that protein sequences determine protein conformation, and thus protein functions, became established in the 1950s in a series of experiments that were taken as exemplary for all proteins and all organisms.[18] In 1949, the physical-chemist Linus Pauling demonstrated that sickle cell anemia resulted from a physical change in the structure of the haemoglobin molecule, most likely a difference in amino acid composition. By 1957, the biochemist Vernon M. Ingram showed that a single amino acid difference caused the abnormal function of the haemoglobin molecule. This result was interpreted, in the exemplary mode, as showing that in all proteins, sequences determined functions.[19] The link between sequences and functions was made more precise that same year, when the biochemist Christian B. Anfinsen found that the protein ribonuclease, after having been denaturated in vitro, would fold spontaneously in its original state. Again, this result was interpreted as having general relevance and from that time it became a common — but therefore not less bold — assumption among molecular biologists that the one-dimensional protein sequences held the key to the three-dimensional structure and function of proteins.[20]

Yet the knowledge of a protein sequence alone, although a major experimental achievement in and of itself, would not indicate which part of the sequence was essential for the function of the molecule or how it performed its task. To answer this question researchers resorted to the collection and the comparison of sequences from many species or from variants within species. They might have been comparing the latest results of experimental science, but their approach belonged to a long standing comparative tradition in biology. They compared structures to learn about

function, as Félix Vicq d'Azyr, George Cuvier and other comparative anatomists had done before them.[21] In this and the next section we will discuss two examples of how researchers collected and compared sequences from various species and from various pathological conditions in humans. While drawing parallels to earlier comparative approaches we will also point to the differences between nineteenth- and late twentieth-century comparative practices.

In 1953, after almost a decade of research, the biochemist Frederick Sanger, working at the University of Cambridge, determined the first complete amino acid sequence of a protein, the small hormone insulin.[22] This achievement was made possible by the ingenious sequencing techniques Sanger had devised, and provided new insights into the primary structure of proteins. It was rewarded by the Nobel Prize in Chemistry just five years later. What is often forgotten in accounts of this feat is that Sanger, as well as a number of protein researchers coming after him, also extensively studied the sequences from various species. The hypothesis behind this practice was that similarities and differences in sequences would provide clues about which part of the protein was responsible for its function. Identical regions, which had been preserved through evolution, might indicate the presence of an essential part of the molecule, such as the "active centre", responsible for its catalytic activity, whereas variable regions might indicate parts of the molecule which had not been under the pressure of natural selection, and thus were probably of lesser functional importance.

As early as the 1950s, this mode of reasoning, and the underlying comparative practice, became widespread among protein researchers, as the following example should make clear. Sanger, who had sequenced insulin isolated from bovine material, did not rely only on this "exemplary" species, chosen mainly as a matter of availability. He also examined insulin from other species, eventually sequencing pig, sheep, horse and whale insulin. This was only a small collection, but it already offered promising opportunities for comparison.[23] Indeed, in 1956, by aligning these five insulin sequences, Sanger and his co-workers discovered that the differences were confined to a small portion of the molecule, namely the disulfide bridge. This came as a surprise because the variability to one disulfide bridge had seemed to indicate that it was not essential for the function of the molecule as was initially thought.[24] The fact that they did not question the rationale behind the sequence comparison approach but called for more studies of species differences showed how much trust they placed on this comparative practice.[25]

The promoters of molecular biology in Cambridge, Francis Crick, Sydney Brenner and Max Perutz, considered Sanger's work on the structure and function of proteins of such fundamental importance for their intellectual and disciplinary projects that they enrolled him in the plan for the new Laboratory of Molecular Biology that opened in 1962.[26] Already in 1957, Crick had drawn on the results of Sanger's comparative studies in his famous lecture "On protein synthesis" in which he outlined the "central dogma" of molecular biology. According to Crick, Sanger's work showed that "these sequences are the most delicate expression possible of the phenotype of an organism",[27] a conclusion which could not have been reached without adopting

comparative analysis. Crick went as far as to predict that "before long we shall have a subject which might be called 'protein taxonomy' — the study of the amino acid sequences of the proteins of an organism and the comparison of them between species".[28] The study of evolutionary history based on protein sequences did indeed become a lively area of research.

Throughout Europe and the United States, protein researchers followed in Sanger's footsteps, not only in determining the sequence of new proteins, but also in performing systematic comparisons of protein sequences from many species. In Vienna, the protein chemist Hans Tuppy, a student of Sanger, sequenced parts of the cytochrome c protein in horse, ox, pig, salmon, and chicken. In 1954, he found that the sequences in the vicinity to what he suspected to be the active site of the molecule were identical in horse, ox, and pig, a result that was all the more remarkable given the many physical differences between these proteins.[29] Tuppy, like Sanger, also took advantage of the first known sequence to infer the others from data on amino acid composition alone. The comparative and the exemplary approaches were thus closely intertwined. For Tuppy, Sanger, and others however, sequence comparisons were far more important than simply a shortcut to the determination of sequences. It was one of the few available methods to assess the relations between structure and function in proteins, a major intellectual agenda among biochemists and early molecular biologists.[30] Tuppy systematized this approach by examining large sets of homologous protein sequences, including insulin, haemoglobin, and trypsin, which had been determined by various other researchers.[31]

Sanger and Tuppy worked mainly on proteins taken from a few domestic species such as ox, horse, rabbit, and pig, whose blood and meat could easily be purchased from slaughterhouses. Other protein researchers such as the chemists Margareta and Birger Blombäck at the Karolinska Institute in Stockholm extended this approach to a much broader range of species. In the early 1950s, they embarked on what would turn out to be a lifelong study of the clotting factor fibrinopeptide. Using the new sequencing method developed by Pehr Edman, which made protein sequencing easier than Sanger's method, they determined the sequences of fibrinopeptides from various mammalian species, including a number of wild species such as badger, bison, fox, green and rhesus monkey, guinea pig, llama, mink, red deer, and reindeer. A comparison of the sequence of fibrinopeptide from 22 species revealed portions which had "been stationary during mammalian evolution" and which were thus likely to be "of importance for directing thrombin action".[32]

Finally, in the United States, the biochemist Christian B. Anfinsen was pursuing a similar project using ribonuclease, and also argued that "variations from species to species may yield valuable information on the location of the site of enzymatic activity".[33] His 1959 book, *The molecular basis of evolution*, drawing on the work of Sanger, Tuppy, the Blombäcks, and others, presented numerous sequence alignments from a wide range of proteins. It did much to popularize the comparative approach among protein researchers and the idea that similarities in sequence would indicate "the minimum structure which is essential for biological function".[34]

For these protein researchers the comparative approach was essential for linking knowledge about protein structure to an understanding of their function, as Cuvier had already argued for anatomy a century earlier. Yet, they were not inspired by reading Cuvier and other nineteenth-century comparative anatomists. They drew on the much more familiar tradition of comparative biochemistry (and comparative physiology) which sought to shed additional light on the function and the generality of biochemical systems by comparing them among various organisms. The biochemist Ernest Baldwin, who had been one of Frederick Sanger's mentors at Cambridge,[35] wrote a popular *Introduction to comparative biochemistry* that was first published in 1937 and went into new editions through the late 1960s.[36] In line with Frederick Gowland Hopkins's programmatic vision, Baldwin's main interest was to produce generalizations about the biochemical basis of life.[37] The study of various species was a way to reach that goal, and for Baldwin "a starfish, or an earthworm, neither of which has any clinical or economic importance per se, is as important as any other living organism and fully entitled to the same consideration".[38] In Cambridge, Francis Crick, while transiting from physics to biology, read Baldwin's comparative biochemistry work and was much impressed by it.[39] The Belgian biochemist Marcel Florkin also published an influential little book in 1944, *L'evolution biochimique*, translated five years later into English.[40] Florkin, too, reviewed the biochemistry of numerous organisms in order to stress "the unity of the biochemical plan of animal organization".[41] The comparative biochemists of the first half of the twentieth century, represented by Baldwin and Florkin, were a major source of inspiration for the protein researchers who made molecular biology half a century later. However, molecular biologists being eager, for disciplinary reasons, to distance themselves from their colleagues in biochemistry, downplayed its considerable influence on the development of their field.[42]

The rise of comparative practices, however, did not simply result from the borrowing of the comparative biochemists' intellectual agenda. It also resulted from major improvements in the technologies which provided the data to be compared and from the establishment of data collections. Comparative studies came to play an increasingly important role in the development of molecular biology as data about protein sequences began to accumulate, thanks to the development of automated protein sequencers in the late 1960s.[43] But the systematic comparison of sequences on a larger scale required something more than large amounts of data. Data collections, performing the same function as museums for natural history, became indispensable tools for the comparative approach.[44] In 1965, the physical-chemist Margaret O. Dayhoff began publishing her *Atlas of protein sequence and structure*, which listed all the known protein sequences and aligned them to make comparisons easier for the reader.[45] The *Atlas* became a common fixture in molecular biology laboratories. Sanger, Perutz, and Kendrew, to name but a few, had their own copies. Each time researchers would determine a sequence, either from a new protein or a new organism, they would immediately compare it to those presented in the *Atlas* to find clues about the presence of the active site or the evolutionary history of the protein. At

Stanford University the molecular biologist Joshua Lederberg praised the *Atlas* as "an important contribution to the next stage of molecular biological architecture".[46] The *Atlas* and its computerized edition represented a key tool for all those who engaged in comparative practices of protein sequences.

The comparative biochemistry tradition was not the only one which inspired molecular biologists trying to understand how structures determined functions. As the following example shows, inspiration also came from a more clinical tradition where different pathologies were collected and compared.

COMPARISONS AND THE PATHOLOGY OF MOLECULES

Max Perutz's study of the three-dimensional structure of haemoglobin by X-ray analysis is often seen as the paradigmatic case of protein structure determination even if it was not the first protein structure that was solved. Perutz started the X-ray analysis of haemoglobin in the late 1930s and the project would occupy him for large part of his long career. It became the flagship project of the emerging science of molecular biology in Cambridge and won Perutz a Nobel prize.[47]

At the time Perutz embarked on the structural study of haemoglobin, proteins were considered to hold the key to all life processes, including inheritance. Perutz was attracted to haemoglobin because of the interesting shift in crystal structure between the oxygenated and de-oxygenated forms of the molecule that was observable under the microscope. Haemoglobin was known to show a cooperative binding effect for oxygen, the so-called "Bohr effect". Perutz hoped that studying the structure of the molecule would provide clues to its oxygen binding mechanism.

Perutz himself has often re-told the tortuous story that led him to an understanding of the cooperative mechanism of haemoglobin function.[48] Here we want to focus on a single aspect of his work: how practices of collecting and comparing were combined with thinking in exemplary terms about the structure and function of the molecule. When Perutz set out to study the structure of haemoglobin, whose size and complexity far exceeded any other molecule that had been studied with the new technique of X-ray crystallography, researchers expected that proteins had a regular structure. The regularity of the molecule would help in deducing the structure from the diffraction images. Protein researchers at the time also supported the view, typical of the exemplary approach, that all proteins shared a common structure. Thus knowing the structure of one protein would give decisive clues for the structure of all proteins.

One approach to solve the structure of hemoglobin consisted in comparing the diffraction patterns of adult with foetal haemoglobin, a project taken up by John Kendrew when he joined Perutz's laboratory in the late 1940s. Later Perutz expanded his comparative approach to include sickle cell haemoglobin. The hope to gain structural clues was not fulfilled, but the comparative approach was nonetheless to prove useful in later stages of the analysis.

When Perutz and his collaborators, after labouring for years, finally obtained full scale atomic models of both the structure of oxygenated and de-oxygenated haemoglobin using the heavy atom replacement method, the function of the molecule

still remained elusive. At this juncture a collection of haemoglobin variants proved crucial. It had been gathered by the clinical biochemist Hermann Lehmann, both in sampling trips around the world and through his broad network of clinical collaborators.[49]

Lehmann, a German émigré working in Britain, contributed to the description of the first haemoglobin variants in humans. Finding additional variants and studying their distribution in different world populations became his all consuming interest. His collecting activity was eventually put on a more secure footing when he became founding head of the Medical Research Council Unit for Abnormal Haemoglobins, established at St Bartholomew's Hospital, London. This Unit, which combined research with clinical work, later moved to Cambridge following his appointment there. Further recognition for Lehmann's collecting work came in the mid-1960s when the World Health Organization designated his Unit as an International Reference Centre for Abnormal Haemoglobins. By that time the collection included more than 100 haemoglobin variants, growing to triple that number ten years later.

Variants were identified first by electrophoretical analysis, later by protein fingerprinting techniques that allowed researchers to pinpoint the amino acid substitutions in each of the haemoglobin molecules.[50] It was this detailed chemical knowledge, combined with Lehmann's extensive clinical knowledge of the various pathologies, that helped Perutz out of his impasse in the haemoglobin work. Built into the model, the amino acid substitutions provided decisive clues on the reaction mechanism of haemoglobin. Lehmann knew the symptoms of the haemoglobinopaties and could link the effects to changes in the sequence and thus with structural changes. In this way he and Perutz could figure out which part of the molecule produced which physiological effect. The researchers showed that the haemoglobin molecule was insensitive to replacements of most of the amino-acid residues on its surface, but very sensitive to any small alteration of internal non-polar contacts. Replacements near the heme groups and at the contacts between the different subunits which composed the haemoglobin molecule affected respiratory functions. The two authors displayed their findings in long tables relating clinical symptoms and abnormal haemoglobin properties to structural effects of amino acid replacement, as well as in detailed stereochemical drawings.[51] Understanding the structural basis of protein function thus relied heavily on comparative work. What has often been celebrated as a triumph of molecular biology's powerful experimental approach is thus better described as resulting from analysis, and more specifically from a combination of comparative and exemplary approaches.

COMPARATIVE APPROACHES TO THE GENETIC CODE

The study of the relationships between protein structure and function, a defining aim of the early molecular biologists, was not the only endeavour that borrowed from the use of collections. Another landmark in the history of molecular biology, the elucidation of the genetic code, relied far more than has been previously recognized on comparative analysis. The standard story highlights the failed attempts to crack the

code theoretically between 1954 and 1961 and the final victory of the experimental approach, following Marshall Nirenberg and Heinrich Mattaei's "exemplary" determination, by biochemical methods, of the first codon.[52] Yet, as we show here, the solution to the code resulted from both exemplary and comparative studies.

In 1954, the big-bang theorist George Gamow suggested that the genetic code could be solved as a cryptogram and made a proposal for an overlapping code. However, by examining the few protein sequences that were then available, Francis Crick realized that Gamow's code was flawed.[53] Gamow was not discouraged. He invited a number of molecular biologists and physicists, including Crick, Martynas Yčas, and Sydney Brenner, to join the RNA Tie Club, which he founded to organize the efforts to decipher the code through cryptoanalytical and computational methods. Lily Kay has described in great detail how these theoretical approaches borrowed, sometimes liberally, concepts from cybernetics, cryptography, and information theory.[54] Yet these attempts were not just theoretical speculations of bright minds; they were also constrained by empirical data, in particular by collections of protein sequences.[55]

The "coding problem", as frequently formulated in the 1950s, was how to relate a text written with four letters (made of nucleotides) to a text written with twenty letters (made of amino acids). This task was particularly difficult because researchers had been unable to determine nucleic acid sequences which could be correlated with the corresponding protein sequences.[56] Thus scientists aiming to decipher the genetic code were stuck with examining protein sequences. One of their main strategies was to align homologous protein sequences and look for amino acid differences. They assumed that each amino acid difference was caused by a change of a single nucleotide in the underlying nucleic acid sequence. The two other nucleotides of the codon (supposing codons had three) were believed to be identical for this particular pair of amino acids. This approach made it possible to drastically narrow down the number of possible codes and brought some welcome constraints to the theoreticians' imagination.[57]

But this comparative approach was not only important for the "failed" attempts of the theoreticians at cracking the code. It became essential after the first codon had been determined experimentally in 1961.[58] Indeed, assuming that a mutation from one amino acid to another involved a single nucleotide change, once a few codons were known, a collection of amino acid changes much simplified the determination of the remaining codons.[59] Researchers could rely on the comparison of the diverse sequences produced by evolution to infer other codons, rather than following the delicate (and in some cases technically impossible) strategy devised by Nirenberg and Mattaei of using synthetic polynucleotides specific for each of the sixty-four codons. Nirenberg and Mattaei's success was in part made possible by the fact that they synthesized simple polynucleotides composed only of uracil. But only four, out of sixty-four, codons were that simple. All the others were composed of two or three different nucleotides. Synthesizing polynucleotides with a specific sequence promised to be immensely more difficult.

Researchers, such as the biochemist Severo Ochoa at New York University, relied

extensively on the comparison of sequences from Tobacco Mosaic Virus (TMV) mutants to confirm theoretical codon assignments and infer new ones.[60] For these comparative studies, they relied on existing collections of TMV protein sequences. These had been determined in the previous years by Akira Tsugita and Heinz Fraenkel-Conrat in Wendell Stanley's laboratory at the University of Berkeley and by Heinz G. Wittmann in Georg Melchers's Max Planck Institut für Biologie in Tübingen. They hoped the sequences would become useful in the future to solve the genetic code, as indeed happened.[61]

Historians have generally recognized the importance of TMV mutants for the resolution of the genetic code. Indeed, this approach fitted well within the historiography about model organisms, such as TMV, and the importance of the "exemplary" approach in molecular biology. Yet historians have overlooked that the comparative approach was extended to a much broader set of organisms, including pigs, sheep and horses (not your usual laboratory animals) and proteins, including cytochromes c, insulin and haemoglobin. Within a year after the first codon was determined experimentally, several researchers, including the biochemists Emil L. Smith, at the University of Utah, Thomas H. Jukes, at American Cyanamid Company in New Jersey, Carl R. Woese, at General Electric in New York, and Walter M. Fitch, at the University of Wisconsin, published a number of new codon assignments solely based on the comparison of protein sequences, and later confirmed codon assignment determined by the use of synthetic polynuclotides.[62] The data obtained from the comparison of "natural" sequence variations was particularly important for researchers like Francis Crick since it showed that the results of in vitro experiments were not artifacts.

A comparative approach was also used to show that the genetic code, established first in bacteria, was also valid for humans. On the basis of fifty-nine known substitutions in haemoglobin variants, it could be shown, using the genetic code, that all amino acid substitutions were the outcome of a single mutation in the coding gene, exactly what one would expect if the genetic code was the same in humans as in the organisms from which it had been derived.[63]

By 1966, the code had been entirely solved. Its solution had rested on the comparison of many homologous sequences from a variety of organisms, including humans, pigs, sheep, oxen, horses, sperm whales, finback whales, humpback whales, seals, salmon, chickens, turkeys, silkworms, frogs, rabbits, bacteria, and viruses. But the impression that this feat came about solely as a result of Nirenberg and Mattaei's experimental and "exemplary" approach, was reinforced by the fact that it took only five years to determine the remaining sixty-three codons composing the genetic code. This speed can be explained, however, by the fact that a number of protein sequences from mutants and various species were already available for comparative studies by the time the first codon was determined experimentally. Thus the history of the genetic code should be told not only as a victory of the experimental approach adopted by Nirenberg and Mattaei over the theoretical approach of Gamow and others, but as an illustration of the combination of exemplary and comparative approaches.

BEYOND MOLECULAR BIOLOGY

Although in this paper we have limited our analysis to the 'classical' period of molecular biology, comparative work practices as well as collections continued to play a key role in later projects undertaken by molecular biologists. Once data about the three-dimensional structure of proteins, obtained by crystallographic methods, began to accumulate, researchers compared them extensively in an attempt to solve the "protein folding problem". Although it was clear from the early 1960s that protein sequences uniquely determined the conformation of proteins, researchers were unable to predict how a given amino acid chain would fold into a functional protein. Even the most powerful computers of the day (and this is still true today) failed to solve this computationally intensive problem. It was in order to find other ways to predict protein folding, as well as to help determine the structure of new proteins, that researchers established the Protein Data Bank at Brookhaven National Laboratory in 1972. Just like Dayhoff's *Atlas of protein sequence and structure*, a decade earlier, the computerized collection of three-dimensional protein structures offered numerous possibilities to compare data and draw inferences about the structure, function, and evolutionary history of proteins.

In 1977, the development of new methods finally made DNA sequencing technically feasible. As a result, DNA sequence data began to accumulate at an even faster rate than protein sequences. Most of these DNA sequences, however, were at first meaningless to researchers, who then sought to understand their functions through extensive comparisons with other sequences, especially homologous sequences in other organisms. They reasoned that a high similarity between two sequences might indicate that the resulting proteins performed a similar function. Researchers then attempted to confirm these tentative functional assignments experimentally in the laboratory. To further this aim, in 1982, the European Molecular Biology Laboratory and the National Institutes of Heath (NIH) established public databases for nucleic acid sequences (the EMBL Nucleotide Sequence Library and GenBank respectively).[64] With the use of increasingly sophisticated computer algorithms these databases made systematic sequence comparisons, over a wide range of sequences from many species, possible and accessible to a wide community of researchers. By 2011, the databases contained sequences from over two hundred thousand different species. This information supported extensive comparative studies, both in support of theoretical research and in combination with experimental work following the "exemplary" approach.

CONCLUSIONS

Pickstone's call to identify different ways of knowing within scientific disciplines proves fruitful in the case of molecular biology. It makes visible a set of practices, based on the collection and comparison of data, which differs from the exemplary approach that is most often associated with molecular biology. This calls for a historical narrative about the rise of molecular biology which is not centred on powerful

instruments and a few model organisms. A focus on comparative molecular analysis opens new directions for understanding the history of molecular biology, such as the influence of comparative biochemistry and comparative pathology, fields which have largely been overlooked so far. More unexpected connections, not explored in this paper, are also made apparent by emphasizing comparative forms of analysis in the life sciences. As this approach grew in importance with the accumulation of molecular data, those who collected and compared data, such as Dayhoff, turned to the new library and information sciences as a source of inspiration for the management and analysis of molecular data.[65] Bringing the rise of molecular biology into historical perspective might thus require to look beyond the impact of other natural sciences, such as physics and chemistry, and take the social sciences into consideration.

Overall, molecular biology appears to have much in common with other approaches to understand life, including the natural historical and the clinical traditions — which is a useful counterpoint to the usual historical emphasis on physical and chemical experimental methods. Those who collected and compared sequences of proteins to shed light on their function took advantage of the fact that their objects of study were the product of biological evolution, not just of physicochemical laws of nature. The fact that natural selection acts on function but is blind to structures offered the possibility to examine variations in structure (in different species or mutants), and draw inferences about the location of the functionally relevant parts of macromolecules. The fact that this approach has been largely ignored in accounts about the history of molecular biology results, perhaps, from the fact that molecular biologists, in their quest for disciplinary authority, were eager to erase any connection they had to more 'traditional' biological practices, especially those reminiscent of the natural history which they so enjoyed deriding.

Comparative analysis has grown in importance since the early days of molecular biology and today is driving the production of knowledge in the life sciences. Databases of protein sequences (the Protein Information Resource, building on Dayhoff's *Atlas*), protein structures (the Protein Data Bank) and nucleic acid sequences (GenBank and EMBL Nucleotide Sequence Database) now play a central role for the collection and comparison of experimental data. In many ways, they serve the same function as the zoological museums and botanical collections of the nineteenth century, which were key places where the collections could be subjected to comparative analysis. Yet, we should not assume that nineteenth-century zoologists and twenty-first-century bioinformatics researchers have adopted identical practices just because they seem to share a similar comparative approach. The fact that naturalists collected things (fossils, bones, skins, plants) while present-day life scientists collect data makes a profound difference in terms of where knowledge is produced and who can produce and access it — especially in the age of digital data. The potential for decentralizing the production of comparative knowledge, made possible by the wide access to databases through the internet, is far greater than for any earlier material collection. This too shows how we can expand our understanding of ways of knowing to ways of doing that are shared with other cultural domains, as Pickstone

aspires. The construction, management and uses of knowledge databases plays a role far beyond biology — think of Wikipedia — and may well define the working of current and future societies.

The two approaches we have distinguished may be seen as divisions of Pickstone's analysis way of knowing, and we agree that analysing things according to their parts, rather that trying to grasp their essence or synthesizing them, constitutes an essential epistemic practice across the sciences. However, we hope we have also shown the benefits of distinguishing between exemplary and comparative forms of analysis, not least so one can show their interactions. We believe that a typology of ways of knowing must make this distinction apparent. Whether our exemplary *v.* comparative distinction holds more generally, beyond analysis, requires further investigation.

Similarly, the chronology of comparative approaches in biology requires further study.[66] Pickstone's ways of knowing come of age at successive times giving a sense of overall progress, or at least of increasing complexity. The analytic way of knowing, in which comparative approaches figure so prominently, was essential to nineteenth-century natural and social sciences. Even though Pickstone makes clear that older ways of knowing remain present alongside more recent ones, they still dominate successive time periods. As our example from the life sciences seems to indicate, comparative analysis, applied to molecular data of all kinds, might become as important for the life sciences of the twenty-first century as it was for nineteenth-century biology. Seen in this light, the exemplary approach, with its focus on single model systems, may come to look like an interlude, although not a trivial one, in the long history of the life sciences dominated by comparative epistemic practices.

REFERENCES

1. This narrative was at the core of William R. Coleman's and Garland E. Allen's seminal work; see William R. Coleman, *Biology in the nineteenth century: Problems of form, function and transformation* (Cambridge, 1971), and Garland E. Allen, *Life science in the twentieth century* (Cambridge and London, 1978). It has served as a model for later work, for example Lois N. Magner, *A history of the life sciences*, 3rd edn (New York, 2002); Peter J. Bowler and Iwan Rhys Morus, *Making modern science: A historical survey* (Chicago, 2005). For a richer description of the practices of natural historians that complicates the neat contraposition between descriptive and experimental life sciences see for instance Joel B. Hagen, "Experimental taxonomy, 1920–1950: The impact of cytology, ecology, and genetics on the ideas of biological classification", Ph.D. dissertation, Oregon State University, 1984; Lynn K. Nyhart, "Natural history and the 'new' biology", in Nicholas Jardine, James A. Secord, and Emma C. Spary (eds), *Cultures of natural history* (London, 1996), 426–43; Robert E. Kohler, *Landscapes and labscapes: Exploring the lab–field border in biology* (Chicago, 2002); Sharon E. Kingsland, *The evolution of American ecology, 1890–2000* (Baltimore, 2005); Erika Lorraine Milam, "'The experimental animal from the naturalist's point of view': Behavior and evolution at the American Museum of Natural History, 1928–1954", in Joe Cain and Michael Ruse (eds), *Descended from Darwin: Insights into the history of evolutionary studies, 1900–1970* (Philadelphia, 2009), 157–78; Mary Terrall, "Following insects around: Tools and techniques of eighteenth-century natural history", *The British journal for the history of science*, xliii (2010), 573–88; Mary Terrall, "Frogs on the mantelpiece: The practice of observation in daily life", in Lorraine Daston and Elizabeth Lunbeck

(eds), *Histories of scientific observation* (Chicago, 2011), 185–205.

2. Robert E. Kohler, *Partners in science: Foundations and natural scientists* (Chicago, 1991); Lily E. Kay, *The molecular vision of life: Caltech, the Rockefeller Foundation and the rise of the new biology* (New York, 1993).
3. Lily E. Kay, "Life as technology, representing, intervening and molecularizing", in Sahotra Sarkar (ed.), *The philosophy and history of molecular biology: New perspectives* (Dordrecht, 1996), 87–100; Angela N. H. Creager, *The life of a virus: Tobacco mosaic virus as an experimental model, 1930–1965* (Chicago, 2002); Nicolas Rasmussen, *Picture control: The electron microscope and the transformation of biology in America, 1940–1960* (Stanford, 1997); Soraya de Chadarevian, *Designs for life: Molecular biology after World War II* (Cambridge, 2002); Bruno J. Strasser, *La fabrique d'une nouvelle science: La biologie moléculaire à l'âge atomique (1945–1964)* (Florence, 2006).
4. On phage, Ernst P. Fischer and Carol Lipson, *Thinking about science: Max Delbrück and the origins of molecular biology* (New York, 1988); on TMV, Creager, *The life of a virus* (ref. 3); on moulds, Paul Berg and Maxine Singer, *George Beadle, an uncommon farmer: The emergence of genetics in the 20th century* (Cold Spring Harbor, NY, 2003); on flies, Robert E. Kohler, *Lords of the fly: Drosophila genetics and the experimental life* (Chicago, 1994); on worms, Soraya de Chadarevian, "Of worms and programmes: *Caenorhabditis elegans* and the study of development", *Studies in history and philosophy of biological and biomedical sciences*, xxix (1998), 81–105 and Rachel A. Ankeny, "The natural history of *Caenorhabditis elegans* research", *Nature reviews genetics*, ii (2001), 474–79; on mice, Karen A. Rader, *Making mice: Standardizing animals for American biomedical research, 1900–1955* (Princeton, 2004); on plants, Sabina Leonelli, "Arabidopsis, the botanical Drosophila: From mouse cress to model organism", *Endeavour*, xxxi (2007), 34–8; and more generally on model organisms and systems, Angela N. H. Creager, Elizabeth Lunbeck, and M. Norton Wise, *Science without laws: Model systems, cases, exemplary narratives* (Durham, 2007).
5. John V. Pickstone, "Ways of knowing: Towards a historical sociology of science, technology and medicine", *The British journal for the history of science*, xxvi (1993), 433–58; John V. Pickstone, "Museological science? The place of the analytical/comparative in nineteenth-century science, technology and medicine", *History of science*, xxxii (1994), 111–38; John V. Pickstone, *Ways of knowing: A new history of science, technology and medicine* (Manchester, 2000); John V. Pickstone, "Working knowledges before and after circa 1800: Practices and disciplines in the history of science, technology and medicine", *Isis*, xcviii (2007), 489–516.
6. For an early use of the expression, see J. M. Mitchison, *The biology of the cell cycle* (Cambridge, 1971), 6. We thank Rachel Ankeny for providing this reference.
7. See for example the naturalist E. O. Wilson's reaction to the attacks of molecular biologists such as James D. Watson in Edward O.Wilson, *Naturalist* (Washington, DC, 1994), chap. 12; and Michael R. Dietrich, "Paradox and persuasion: Negotiating the place of molecular evolution within evolutionary biology", *Journal of the history of biology*, xxxi (1998), 85–111; Joel B. Hagen, "Naturalists, molecular biology, and the challenge of molecular evolution", *Journal of the history of biology*, xxxii (1999), 321–41; Erika Lorraine Milam, "The equally wonderful field: Ernst Mayr and organismic biology", *Historical studies in the natural sciences*, xl (2010), 279–317.
8. On other attempts to challenge this distinction, see Milam, "'The experimental animal from the naturalist's point of view'" (ref. 1); Kohler, *Landscapes and labscapes* (ref. 1); Nyhart, "Natural history and the 'new' biology" (ref. 1); Bruno J. Strasser, "Laboratories, museums, and the comparative perspective: Alan A. Boyden's serological taxonomy, 1925–1962", *Historical studies in the natural sciences*, xl (2010), 149–82. On collecting as an essential practice in a wide spectrum of sciences, including genomics, see Pickstone, *Ways of knowing* (ref. 5); Bruno J. Strasser, "Collecting and experimenting: The moral economies of biological research,

1960s–1980s", *Preprints of the Max-Plack Institute for the History of Science*, no. 310 (2006), 105–23; Robert E. Kohler, "Finders, keepers: Collecting sciences and collecting practice", *History of science*, xlv (2007), 428–54.

9. Pickstone, *Ways of knowing* (ref. 5), chap. 4.
10. Ludwik Fleck, *Genesis and development of a scientific fact* (Chicago, 1979); Thomas S. Kuhn, *The structure of scientific revolutions*, 2nd edn (Chicago, 1970); Michel Foucault and Colin Gordon, *Power/knowledge: Selected interviews and other writings, 1972–1977* (New York, 1980); Gerald James Holton, *Thematic origins of scientific thought: Kepler to Einstein* (Cambridge, MA, 1973); Alistair C. Crombie, *Styles of scientific thinking in the European tradition: The history of argument and explanation especially in the mathematical and biomedical sciences and arts* (London, 1994).
11. Ian Hacking, "'Style' for historians and philosophers", *Studies in history and philosophy of science*, xxiii (1992), 1–20.
12. See also Ian Hacking, "The self-vindication of the laboratory sciences", in Andrew Pickering (ed.), *Science as practice and culture* (Chicago, 1992), 29–64. For a critique of Hacking's notion of "style", see Martin Kusch, "Hacking's historical epistemology: A critique of styles of reasoning", *Studies in history and philosophy of science*, xli (2010), 158–73.
13. Pickstone, "Museological science?" (ref. 5).
14. For a very useful exposition of exemplary cases in the clinic, see Ilana Löwy, "The experimental body", in Roger Cooter and John Pickstone (eds), *Medicine in the twentieth century* (Amsterdam, 2000), 435–49. On model organisms more generally, see ref. 5.
15. Strasser, "Laboratories, museums, and the comparative perspective" (ref. 8).
16. Jacques Monod and François Jacob, "General conclusions: Teleonomic mechanisms in cellular metabolism, growth, and differentiation", *Cold Spring Harbor Symposia on Quantitative Biology*, xxi (1961), 389–401. On the origins of this expression, see Herbert C. Friedmann, "From 'Butyribacterium' to 'E. coli': An essay on unity in biochemistry", *Perspectives in biology and medicine*, xlvii (2004), 47–66. In the earlier part of the twentieth century, knowledge derived from animal model organisms was believed to hold true only for animals, not plants and microbes, as would become the case in molecular biology starting in the late 1950s.
17. On this point see also Pickstone, "Working knowledges before and after circa 1800" (ref. 5), 513–14.
18. Bruno J. Strasser, "A world in one dimension: Linus Pauling, Francis Crick and the central dogma of molecular biology", *History and philosophy of the life sciences*, xxviii (2006), 491–512.
19. The general rule was interpreted less strictly when the idea of "allosteric" proteins was developed a few years later. Angela N. H. Creager and Jean-Paul Gaudillière, "Meanings in search of experiments and vice-versa: The invention of allosteric regulation in Paris and Berkley, 1959–1968", *Historical studies in the physical and biological sciences*, xxvii (1996), 1–89.
20. Francis H. C. Crick, "On protein synthesis", *Symposia of the Society for Experimental Biology*, xii (1958), 138–63.
21. Stéphane Schmitt, "From physiology to classification: Comparative anatomy and Vicq d'Azyr's plan of reform for life sciences and medicine (1774–1794)", *Science in context*, xxii (2009), 145–93; Dorinda Outram, *Georges Cuvier: Vocation, science, and authority in post-revolutionary France* (Manchester, 1984).
22. Soraya de Chadarevian, "Sequences, conformation, information: Biochemists and molecular biologists in the 1950s", *Journal of the history of biology*, xxix (1996), 361–86; Miguel Garcia-Sancho, "A new insight into Sanger's development of sequencing: From proteins to DNA, 1943–1977", *Journal of the history of biology*, xliii (2010), 265–323.
23. Frederick Sanger, "Species differences in insulins", *Nature*, clxiv (1949), 529; H. Brown, Frederick Sanger, and Ruth Kitai, "The structure of pig and sheep insulins", *Biochemical journal*, lx (1955), 556–65; J. Ieuan Harris, Michael A. Naughton, and Frederick Sanger, "Species differences in

insulin", *Archives of biochemistry and biophysics*, lxv (1956), 427–38; Soraya de Chadarevian, "Protein sequencing and the making of molecular genetics", *Trends in biochemical sciences*, xxiv (1999), 203–6.

24. Harris, Naughton, and Sanger, "Species differences in insulin" (ref. 23); Brown, Sanger, and Kitai, "The structure of pig and sheep insulins" (ref. 23), 565.
25. Harris, Naughton, and Sanger, "Species differences in insulin" (ref. 23), 437. Sequence comparisons did not necessarily regard sequences from other species. They could also concern enzyme families from the same species. This project was pursued for instance by Brian Hartley who later moved together with Sanger to the Laboratory of Molecular Biology in Cambridge. His comparative study of various pancreatic enzymes first served to study the catalytic mechanism of the enzymes, but soon led to comparative evolutionary studies; see de Chadarevian, *Designs for life* (ref. 3), 273.
26. de Chadarevian, "Sequences, conformation, information" (ref. 22).
27. Crick, "On protein synthesis" (ref. 20), 142.
28. *Ibid*.
29. Hans Tuppy and Gerhard Bodo, "Cytochrom c. III: Zur Frage der Artspezifität von Säugetier-Cytochrom c", *Monatshefte für Chemie*, lxxxv (1954), 1182–6.
30. Sven Paléus and Hans Tuppy, "A hemopeptide from a tryptic hydrolysate of *Rhodospirillum rubrum* cytochrome-c", *Acta chemica scandinavica*, xiii (1959), 641–6, p. 642.
31. Hans Tuppy, "Über die Artspezifität der Proteinstruktur", in Albert Neuberger (ed), *Symposium on protein structure* (New York, 1958), 66–76.
32. Birger Blombäck, Margareta Blombäck, and Nils Jakob Grondahl, "Studies on fibrinopeptides from mammals", *Acta chemica scandinavica*, xix (1965), 1789–91, p. 1789.
33. C. B. Anfinsen, S. E. Aqvist, J. P. Cooke and B. Jonsson, "A comparative study of the structures of bovine and ovine pancreatic ribonucleases", *Journal of biological chemistry*, ccxxxiv (1959), 1118–23, p. 1118.
34. *Ibid*., 143.
35. Frederick Sanger, "Sequences, sequences, and sequences", *Annual review of biochemistry*, lvii (1988), 1–28, p. 3.
36. Ernest Baldwin, *An introduction to comparative biochemistry* (Cambridge, 1937; 2nd edn, Cambridge, 1966).
37. Baldwin was a student of Hopkins. On Hopkins, see Robert E. Kohler, *From medical chemistry to biochemistry: The making of a biomedical discipline* (Cambridge, 1982).
38. Baldwin, *An introduction to comparative biochemistry* (ref. 36), p. xiv.
39. de Chadarevian, *Designs for life* (ref. 3), 89.
40. Marcel Florkin, *L'évolution biochimique* (Paris, 1944; Engl. transl., New York, 1949).
41. Florkin, *L'évolution biochimique* (ref. 40), 11, translation ours.
42. Pnina Abir-Am, "The politics of macromolecules: Molecular biologists, biochemists, and rhetoric", *Osiris*, vii (1992), 164–91.
43. Stanford Moore, Darrel H. Spackman, and William H. Stein, "Automatic recording apparatus for use in the chromatography of amino acids", *Federation proceedings*, lvii (1958), 1107–15. Commercial versions only became available in the 1960s.
44. Strasser, "Laboratories, museums, and the comparative perspective" (ref. 8).
45. Bruno J. Strasser, "Collecting, comparing, and computing sequences: The making of Margaret O. Dayhoff's *Atlas of Protein Sequences and Structure*, 1954–1965", *Journal of the history of biology*, xliii (2010), 623–60.
46. J. Lederberg to M. O. Dayhoff, 12 March 1964, National Biomedical Research Foundation Archives, Georgetown, MD.

47. de Chadarevian, *Designs for life* (ref. 3) ; Georgina Ferry, *Max Perutz and the secret of life* (New York, 2007).
48. Max F. Perutz, *I wish I'd made you angry earlier: Essays on science, scientists, and humanity* (Plainview, NY, 1998).
49. Soraya de Chadarevian, "Following molecules: Haemoglobin between the clinic and the laboratory", in Soraya de Chadarevian and Harmke Kamminga (eds), *Molecularizing biology and medicine: New practices and alliances, 1910s–1970s* (Amsterdam, 1998), 171–201.
50. Hermann Lehmann, *Man's haemoglobins: Including the haemoglobinopathies and their investigation* (Amsterdam, 1966).
51. Max F. Perutz and Hermann Lehmann, "Molecular pathology of human haemoglobin", *Nature*, ccxix (1968), 902–9; Hermann Morimoto, H. Lehmann, and Max F. Perutz, "Molecular pathology of human haemoglobin: Stereochemical interpretation of abnormal oxygen affinities", *Nature*, ccxxxii (1971), 408–13. Significantly, clinical symptoms included observations obtained at the bedside (e.g. cyanosis, splenomegaly) as well as classifications based on laboratory data (e.g. inclusion body anemia, haemolytic anemia). Even before paper electrophoresis of abnormal haemoglobins became a standard procedure in hematology, hematologists relied heavily on microscopic observation and laboratory tests for the description and classification of diseases.
52. Michel Morange, *A history of molecular biology* (Cambridge, 2000), chap. 12; Lily E. Kay, *Who wrote the book of life?: A history of the genetic code* (Stanford, 2000).
53. George Gamow, "Possible relation between desoxyribonucleic acid and protein structure", *Nature*, clxxiii (1954), 318; George Gamow and Nicholas Metropolis, "Numerology of polypeptide chains", *Science*, cxx (1954), 779–80; Francis H. C. Crick, *What mad pursuit* (New York, 1990), 94.
54. Kay, *Who wrote the book of life?* (ref. 52).
55. Strasser, "Collecting, comparing, and computing sequences" (ref. 45).
56. See de Chadarevian, "Protein sequencing and the making of molecular genetics" (ref. 23).
57. For examples of this approach, see George Gamow, Alexander Rich, and Martynas Yčas, "The problem of information transfer from the nucleic acids to proteins", *Advances in biological and medical physics*, iv (1956), 23–68; Martynas Ycas, "The protein text", in Hubert P. Yockey (ed.), *Symposium on information theory in biology* (New York, 1958), 70–102; Martynas Ycas, "Replacement of amino acids in proteins", *Journal of theoretical biology*, i (1961), 244–57. On the genesis of Gamow's review, see Kay, *Who wrote the book of life?* (ref. 52), 148. On the impossibility of overlapping codes, see Sydney Brenner, "On the impossibility of all overlapping triplet codes in information transfer from nucleic acid to proteins", *Proceedings of the National Academy of Sciences of the United States of America*, xliii (1957), 687–94. In fact, Brenner only showed that codes overlapping by 2 nucleotides out of 3 were impossible.
58. Kay points out that this information was used to confirm the code, but she does not raise the point that it was produced with the coding problem in mind; see Kay, *Who wrote the book of life?* (ref. 52), 187–9. However, Angela Creager makes this point; see Creager, *The life of a virus* (ref. 3), 303–11.
59. For example, if UUU coded for the amino acid phenylalanine, as Nirenberg and Matthaei had established, and phenylalanine were replaced by another amino acid in a mutant, scientists could deduce that this amino acid was coded by one of only nine different codons (all including two Us), thus excluding forty-four other possible combinations.
60. Joseph F. Speyer *et al*., "Synthetic polynucleotides and the amino acid code. II", *Proceedings of the National Academy of Sciences of the United States of America*, xlviii (1962), 63–8; Peter Lengyel *et al.*, "Synthetic polynucleotides and the amino acid code. III", *Proceedings of the National Academy of Sciences of the United States of America*, xlviii (1962), 282–4; J. F. Speyer *et al.*, "Synthetic polynucleotides and the amino acid code. IV", *Proceedings of the National Academy*

of Sciences of the United States of America, xlviii (1962), 441–8; Carlos Basilio *et al.*, "Synthetic polynucleotides and the amino acid code. V", *Proceedings of the National Academy of Sciences of the United States of America*, xlviii (1962), 613–16.

61. Creager, *The life of a virus* (ref. 3), 303–11; Kay, *Who wrote the book of life?* (ref. 52), 179–92; Christina Brandt, *Metapher und Experiment: Von der Virusforschung zum genetischen Code* (Göttingen, 2004), chap. 6.
62. Emil L. Smith, "Nucleotide base coding and amino acid replacements in proteins", *Proceedings of the National Academy of Sciences of the United States of America*, xlviii (1962), 677–84; Emil L. Smith, "Nucleotide base coding and amino acid replacements in proteins. II", *Proceedings of the National Academy of Sciences of the United States of America*, xlviii (1962), 859–64; Thomas H. Jukes, "Beta lactoglobulins and amino acid code", *Biochemical and biophysical research communications*, vii (1962), 281–3; Thomas H. Jukes, "Possible base sequences in amino acid code", *Biochemical and biophysical research communications*, vii (1962), 497–502; Thomas H. Jukes, "Relations between mutations and base sequences in amino acid code", *Proceedings of the National Academy of Sciences of the United States of America*, xlviii (1962), 1809–15, Carl R. Woese, "Nature of the biological code", *Nature*, cxciv (1962), 1114–15; Walter M. Fitch, "The probable sequence of nucleotides in some codons", *Proceedings of the National Academy of Sciences of the United States of America*, cxxii (1964), 298–305; Walter M. Fitch, "The relation between frequencies of amino acids and ordered trinucleotides", *Journal of molecular biology*, xvi (1966), 1–8. For a review of the contributions from the comparative approach, see Severo Ochoa, "Chemical basis of heredity, the genetic code", *Experientia*, xx (1964), 57–68; Francis H. C. Crick, "On the genetic code", *Science*, cxxxix (1963), 461–4.
63. Smith, "Nucleotide base coding and amino acid replacements in proteins" (ref. 62); Thomas H. Jukes, "The genetic code", *American scientist*, li (1963), 227–45; D. Beale and H. Lehmann, "Abnormal haemoglobins and the genetic code", *Nature*, ccvii (1965), 259–61.
64. Bruno J. Strasser, "The experimenter's museum: GenBank, natural history, and the moral economies of biomedicine", *Isis*, cii (2011), 60–96.
65. Strasser, "Collecting, comparing, and computing sequences" (ref. 45). As Pickstone pointed out, forms of comparative analysis in the nineteenth century also spanned the natural and the human sciences; Pickstone, *Ways of knowing* (ref. 5), chap. 4.
66. For a history of early comparative approaches in morphology, see Edward Stuart Russell, *Form and function : A contribution to the history of animal morphology* (London, 1916).