nomic environment that includes various social values, research practices and business pressures. We are mindful that, in some situations, modifying patent law may reduce one problem (such as permitting more competition), while magnifying others (such as reducing incentives to conduct research and development). Nevertheless, although care must be taken, this debate needs to progress to ensure that patenting practices, as applied to genetic material, fulfil the ultimate objective of encouraging the development of genetic technologies into products for the public's good.

Update – note added in proof

The recent announcement that scientists will share a patent over a disease-related gene[43] with a patient advocacy group, who provided the researchers with blood and tissue samples[44], is a positive sign that researchers take seriously their moral responsibility to donors. Such steps are in agreement with recent policy statements issued by HUGO[45]. Binding legal measures would help to ensure that researchers and companies who comply with this type of ethical norm do not face unfair competition from those who do not.

Timothy Caulfield is at the Health Law Institute, University of Alberta, Edmonton, Alberta, T6G 2H5, Canada. E. Richard Gold is at the Faculty of Law, University of Western Ontario, London Ontario, N6A 3K7 , Canada. Mildred K. Cho is at the Center for Biomedical Ethics, Stanford University, Stanford, California 94304 , USA. Correspondence to : tcaulfld@law.ualberta.ca

Links

DATABASE LINKS BRCA1 | APOE

FURTHER INFORMATION **American College of Medical Genetics | Unesco's 1997 Universal Declaration on the Human Genome and Human Rights | European Patent Office | United States Patent Office | Canadian Patent Office | Japanese Patent Office | patent on the 'onco-mouse' | European Patent Convention | Incyte Pharmaceuticals | United States Supreme court case of Diamond versus Chakrabarty | United States Patent Office's recent interim guidelines**

1. Rifkin, J. *The Biotech Century* (Penguin Putnam, New York, 1998).
2. American College of Medical Genetics, Position Statement on Gene Patents and Accessibility of Gene Testing (1999). www.faseb.org/genetics/acmg/pol-34.htm
3. Sarma, L. Biopiracy: Twentieth century imperialism in the form of international agreements. *Temple International and Comparative Law Journal* **13**, 107–136 (1999).
4. Thomas, S. *et al.* Ownership of the human genome. *Nature* **380**, 387–388 (1996).
5. Thomas, S. in *The Commercialization of Genetic Research: Ethical, Legal and Policy Issues* (eds Caulfield, T. & Williams-Jones, B.) 55–62 (Kluwer Academic/Plenum Publishing, New York, 1999).
6. Nau, J. Y. Brevetabilité des gènes humains: le comité d'éthique en désaccord avec la directive européenne. *Le Monde* 15 June (2000).
7. Kolata, G. Special Report: Who owns your genes? *New York Times* 15 May (2000).
8. Ramirez, A. School given patent to clone humans. *National Post* 16 May (2000).
9. Sagar, A., Daemmrich, A. & Ashiya, M. The tragedy of commoners: biotechnology and its publics. *Nature Biotechnol.* **18**, 2–4 (2000).
10. Angell, M. Is academic medicine for sale? *N. Engl. J. Med.* **20**, 1516–1518 (2000).
11. Pottagem, A. The inscription of life in law: gene, patents, and bio-politics. *The Modern Law Review* **61**, 740–765 (1998).
12. Gold, E. R. *Body Parts: Property Rights and the Ownership of Human Biological Materials* (Georgetown Univ. Press, Washington DC, 1996).
13. Caulfield, T. & Gold, E. R. Whistling in the wind: reframing the genetic patent debate. *Forum for Applied Research and Public Policy* **15**, 75–79 (2000).
14. Ernst and Young's Fourth Report on the Canadian Biotechnology Industry. *Can. Biotechol. '97: Coming of Age* (Ernst and Young, 1997).
15. *President and Fellows of Harvard v. Commissioner of Patents* (August 3, 2000) No. A-334–398 (Fed. Crt of Appeals).
16. Nottingham, S. *Eat Your Genes* (St. Martin's, New York, 1999).
17. Roberts, T. Why not patent plants? *Patent World* **113**, 14–16 (1999).
18. Schehr, R. & Fox, J. Human genome bombshell. *Nature Biotechnol.* **18**, 365 (2000).
19. Marcus, A. Owning a gene: patent pending. *Nature Med.* **2**, 728–729 (1996).
20. Nelkin, D. & Andrews, L. *Homo economicus:* Commercialization of body tissue in the age of biotechnology. *Hastings Center Report* **28**, 30–39 (1998).
21. Heller, M. & Eisenberg, R. Can patents deter innovation? The anticommons in biomedical research. *Science* **280**, 698–701 (1998).
22. Knoppers, B. M. Status, sale and patenting of human genetic material: an international survey. *Nature Genet.* **22**, 23–26 (1999).
23. Bunk, S. Researchers feel threatened by disease gene patents. *The Scientist* **13**, 7 (1999)
24. Academy of Clinical Laboratory Physicians and Scientists. *ACLPS Resolution: Exclusive Licenses for Diagnostic Tests Approved by the ACLPS Executive Council 06/03/99* (1999). http://depts.washington.edu/labweb.aclps/license/htm
25. Cho, M. K. in *Preparing for the Millennium: Laboratory Medicine in the 21st Century, December 4–5, 1998,* 2nd edn 47–53 (AACC, Washington DC, 1998).
26. Caulfield, T. & Gold, E. R. Genetic testing, ethical concerns, and the role of patent law. *Clin. Genet.* **57**, 370–375 (2000).
27. Bruzzone, L. The research exemption: a proposal. *Am. Intell. Prop. Law Assoc. QL* **21**, 52 (1993).
28. Parker, D. Patent infringement exemptions for life science research. *Houston J. Intl Law* **16**, 615 (1994).
29. Gold, E. R. in *Commercialization of Genetic Research: Ethical, Legal and Policy Issues* (eds Caulfield, T. & Williams–Jones, B.) 63–78 (Plenum, New York, 1999).
30. Schissel, A., Merz, J. F. & Cho, M. K. Survey confirms fear about licensing of genetic tests. *Nature* **402**, 118 (1999).
31. Blumenthal, D. *et al.* Withholding Research Results in Academic Life Science: Evidence From a National Survey of Faculty *J. Am. Med. Assoc.* **277**, 1224 (1997).
32. Caulfield, T. The commercialization of human genetics: a discussion of issues relevant to Canadian consumers. *J. Consumer Policy* **21**, 483–526 (1998).
33. Packer, K. & Webster, A. Patenting culture in science: reinventing the scientific wheel of credibility. *Science, Technology and Human Values* **21**, 425–445 (1996).
34. Blumenthal, D. Academic–industry relationships in the life sciences. *J. Am. Med. Assoc.* **268**, 3344 (1992).
35. Straus, J. Intellectual property issues in genome research. *Genome Digest* **3**, 1–2 (1996).
36. Barton, J. Reforming the patent system. *Science* **287**, 1933–1934 (2000).
37. United States Patent and Trade Mark Office. *Interim Utility Guidelines* (1999).
38. Holtzman, N. Are genetic tests adequately regulated? *Science* **286**, 409 (1999).
39. Kodish, E. Commentary: Risks and benefits, testing and screening, cancer, genes and dollars. *J. Law Med. Ethics* **25**, 252–255 (1997).
40. Brower, V. News: Testing, testing, testing? *Nature Med.* **3**, 131–132 (1997).
41. Weiss, R. Genetic testing's human toll. *Washington Post* 21 July (1999).
42. Cowan, D. Tort liability of patentee licensors. *J. Patent Office Soc.* **64**, 87–104 (1982).
43. Le Saux, O. *et al.* Mutations in a gene encoding an ABC transporter cause pseudoxanthoma elasticum. *Nature Genet.* **25**, 223–227 (2000).
44. Smaglik, P. Tissue donors use their influence in deal over gene patent terms. *Nature* **407**, 821 (2000).
45. Human Genome Organization Ethics Committee. Genetic benefit sharing. *Science* **290**, 49 (2000).

TIMELINE

The origins of bioinformatics

Joel B. Hagen

Bioinformatics is often described as being in its infancy, but computers emerged as important tools in molecular biology during the early 1960s. A decade before DNA sequencing became feasible, computational biologists focused on the rapidly accumulating data from protein biochemistry. Without the benefits of supercomputers or computer networks, these scientists laid important conceptual and technical foundations for bioinformatics today.

It is tempting to trace the origins of bioinformatics to the recent convergence of DNA sequencing, large-scale genome projects, the internet and supercomputers[1–3]. However, some scientists who claim that bioinformatics is in its infancy acknowledge that computers were important tools in molecular biology a decade before DNA sequencing became feasible[4]. Although the pioneers of computational biology did not use the term 'bioinformatics' to describe their work, they had a clear vision of how computer technology, mathematics and molecular biology could be fruitfully combined to answer fundamental questions in the life sciences.

Three important factors facilitated the emergence of computational biology during the early 1960s. First, an expanding collection of amino-acid sequences provided both a

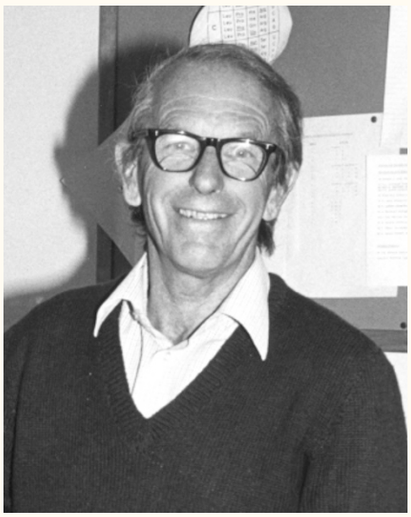


Figure 1 | **Frederick Sanger at the Nobel prize ceremony in 1980.**
(Photograph kindly provided by the MRC, Laboratory of Molecular Biology, Cambridge, UK.)

source of data and a set of interesting problems that were infeasible to solve without the number-crunching power of computers. Second, the idea that macromolecules carry information became a central part of the conceptual framework of molecular biology. Although some historians and philosophers have questioned the theoretical significance of this idea for modern molecular biology[5–7], it seems likely that thinking in terms of macromolecular information provided an important conceptual link between molecular biology and the computer science from which formal information theory had arisen. Third, high-speed digital computers, which had developed from weapons research programmes during the Second World War, finally became widely available to academic biologists. Not all biologists had — or wanted to have — access to these machines but, by 1960, scarcity of computers was no longer a serious stumbling block for the development of computational biology.

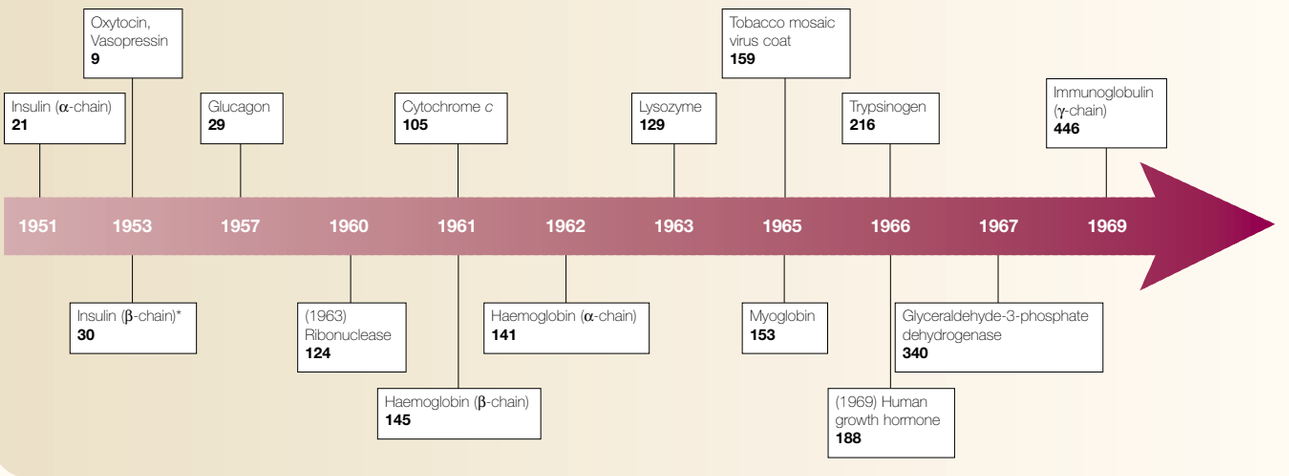
Sequencing proteins

The idea that proteins carry information encoded in linear sequences of amino acids is commonplace today, but it has a relatively short history. This idea first emerged during the decades following the Second World War, a time that one main participant, Emil Smith, later described as a "heroic period" in protein biochemistry[8]. The watershed event of this period was the first successful sequencing of a complete protein, INSULIN, by Frederick Sanger and his colleagues[8–10] at Cambridge University during the decade 1945–1955 (FIG. 1).

Sanger's achievement, for which he was awarded the 1958 Nobel Prize in chemistry, firmly established the polypeptide theory of protein structure. First formulated in 1902, this theory had faced considerable scepticism and competition from alternative theories[9] (FIG. 2). Analytical techniques in protein biochemistry had improved greatly during the 1930s and 1940s, but before Sanger's work, practically nothing was known about the order of amino acids in any protein. One could, therefore, still cling to the belief that proteins were structurally simple or even that they had no definite structure at all. As the biochemist Paul Zamecnik later recalled, these lingering concerns were "blown away" by Sanger's work, which quickly dispelled any doubts that each protein was characterized by a unique primary structure[11].

Sequencing insulin was a case of problem solving by a master chemist who used great scientific skill in separating and identifying the fragments of protein degradation[12]. At the same time, however, other biochemists were developing more refined methods that would transform the laborious analytical process used by Sanger and his co-workers. The Edman degradation reaction, by which biochemists could sequentially remove and identify individual amino acids from the amino terminus of a short peptide, was a great improvement over the more tedious methods used by Sanger[8,9]. The use of ion exchange columns and other innovations in CHROMATOGRAPHY and electrophoresis also made sequencing more efficient. Just as significantly, the entire process of separating and identifying amino acids was rapidly becoming automated. Using semi-automated techniques, researchers led by Stanford Moore and William Stein at the Rockefeller Institute were able to sequence the 124 amino acids in RIBONUCLEASE in about half the time that Sanger's group had spent deciphering the sequence of the 51 amino acids in insulin[13,14]. Automation sent a shock wave through the biochemical community, because it promised to transform sequencing into a routine procedure carried out, not by master chemists, but by competent laboratory technicians[8]. By the late 1960s, Pehr Edman had designed the 'sequenator', a fully automated sequencing machine that implemented his already widely used degradation reaction[15].

Timeline | **Some early milestones in protein and peptide sequencing**



*The complete primary structure of insulin, including the positions of the disulphide bonds, was published in 1955.
(Dates in parentheses are for revisions of the originally published sequences; numbers in bold are the numbers of amino acids.)
Source: L.R. Croft, *Handbook of Protein Sequence Analysis: A Compilation of Amino Acid Sequences of Proteins with an Introduction to the Methodology* (John Wiley, Chichester, 1980).

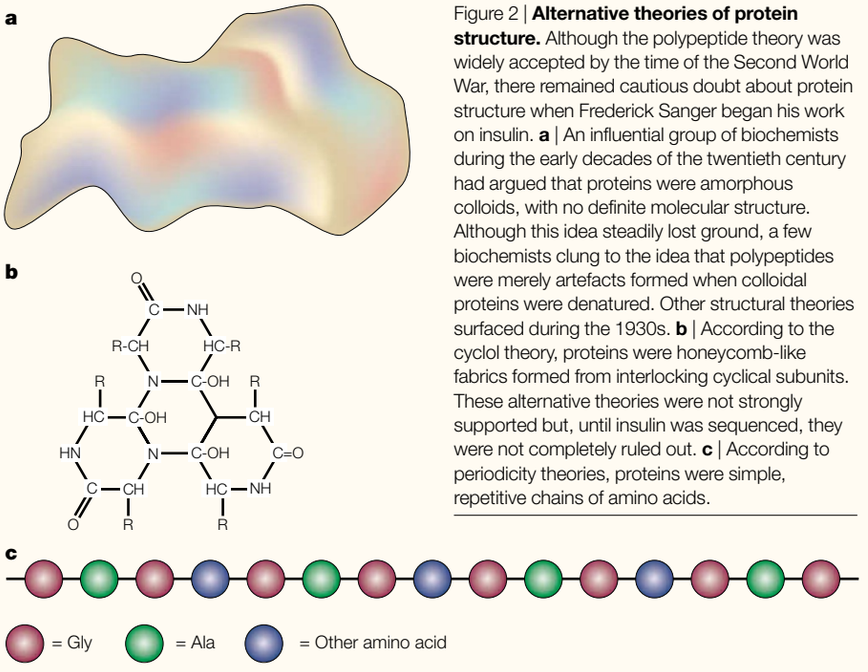


Figure 2 | **Alternative theories of protein structure.** Although the polypeptide theory was widely accepted by the time of the Second World War, there remained cautious doubt about protein structure when Frederick Sanger began his work on insulin. **a** | An influential group of biochemists during the early decades of the twentieth century had argued that proteins were amorphous colloids, with no definite molecular structure. Although this idea steadily lost ground, a few biochemists clung to the idea that polypeptides were merely artefacts formed when colloidal proteins were denatured. Other structural theories surfaced during the 1930s. **b** | According to the cyclol theory, proteins were honeycomb-like fabrics formed from interlocking cyclical subunits. These alternative theories were not strongly supported but, until insulin was sequenced, they were not completely ruled out. **c** | According to periodicity theories, proteins were simple, repetitive chains of amino acids.

These innovations encouraged many laboratories to begin sequencing proteins and rapidly expanded the library of amino-acid sequences (TIMELINE).

Macromolecular information

Once the polypeptide theory became firmly established and methods for sequencing proteins were readily available, the idea of proteins as information-carrying macromolecules became widespread. This general idea developed within three broadly overlapping contexts: the genetic code, the three-dimensional structure of a protein in relation to its function, and protein evolution.

The genetic code. Concurrent developments in the molecular biology of the gene provided a compelling theoretical context for discussing how genetic information was transferred from a sequence of nucleotides to a sequence of amino acids. However, the sequencing of DNA and RNA presented formidable technical hurdles that were not fully overcome until the early 1970s (REF. 16). So, although molecular biologists learned a great deal about the genetic code, the actual nucleotide sequences of genes remained largely unknown during the 1960s. With a growing collection of amino-acid sequences, the idea of molecular information could therefore be explored with proteins in ways not applicable to nucleic acids.

Protein structure. From a purely biochemical perspective, one could ask about the causal relationship between the information carried in the primary structure of a protein and the three-dimensional configuration of the active molecule. Experiments carried out by Christian Anfinsen and his colleagues at the National Institutes of Health in the late 1950s showed that, after being denatured, ribonuclease spontaneously refolded to regain its original enzymatic activity[17]. This was taken as compelling evidence that the sequence of amino acids completely specified the three-dimensional structure of the protein. Of course, in practical terms, knowing the sequence did not necessarily allow biochemists to correctly predict the secondary and tertiary structures of a protein. But sequence data played a key role in interpreting the X-ray diffraction images used by John Kendrew and Max Perutz (FIG. 3) to determine the three-dimensional structures of MYOGLOBIN and HAEMOGLOBIN[18]. Combining the biochemical techniques of sequence analysis with the biophysical techniques of X-RAY CRYSTALLOGRAPHY seemed to hold the key to understanding how the molecular information in a sequence of amino acids causes a protein to fold into a specific, often highly complex, three-dimensional configuration[13,14,17].

Protein evolution. The idea that linear information could determine the structure and function of proteins fits squarely within a dominant tradition in twentieth-century biochemistry — the 'protein paradigm'[19]. For several decades before 1960, biochemists had focused their efforts on providing mechanistic explanations for how enzymes, protein hormones, antibodies and respiratory pigments worked. This episode has attracted considerable historical interest[9,12,20], but historians have paid much less attention to how macromolecular information could be thought of in an explicitly phylogenetic context. This phylogenetic approach was a significant departure from traditional biochemical thinking about structure and function, which had largely ignored evolutionary questions. Indeed, historians have often stressed the conflicts between evolutionary biology and the more experimental sciences such as biochemistry. However, during the 1960s, biochemists and molecular biologists were increasingly drawn to evolutionary questions. For example, Emile Zuckerkandl and Linus Pauling referred to proteins and nucleic acids as "semantides", whose information-carrying sequences of subunits could be used to document evolutionary history[21]. Derived from 'semanteme', the fundamental unit of meaning used by linguists to study human speech, semantides were to be the analogous biochemical units (hence the chemical suffix — ide) for evolutionary studies.

How would the evolutionary information carried by semantides be used? Zuckerkandl and Pauling imagined a new field of 'paleogenetics' that would use the rigorous laboratory techniques of biochemistry and molecular biology to answer evolutionary questions traditionally studied by palaeontologists and naturalists. Paleogeneticists, who soon

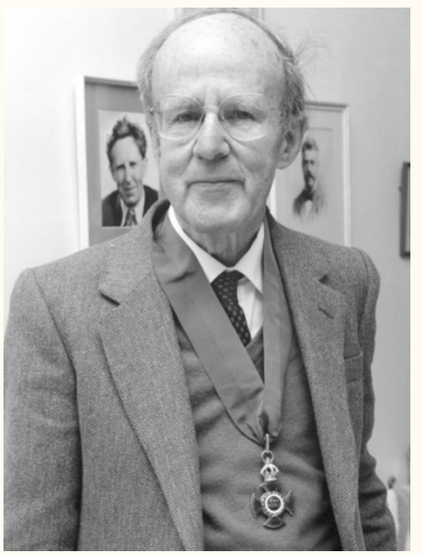


Figure 3 | **Max Perutz, who shared the 1962 Nobel prize in chemistry with John Kendrew.** (Photograph kindly provided by the MRC, Laboratory of Molecular Biology, Cambridge, UK.)

became more commonly referred to as molecular evolutionists, had several approaches at their disposal. Comparisons of similar proteins, such as myoglobin and haemoglobin, provided evidence for molecular evolution by gene duplication. Comparison of homologous proteins drawn from various species could be used to trace phylogenetic relationships among both the proteins themselves and the species that carried them. In some cases, such comparisons could also be used to recreate the ancestral proteins from which present-day molecules evolved. Assuming that amino-acid substitution rates were relatively constant within a given protein, paleogeneticists had a 'MOLECULAR CLOCK' by which evolutionary events might be reliably dated.

These claims, particularly the idea of a molecular clock, were enormously controversial and provided a source of conflict between molecular evolutionists and traditional naturalists[22–24]. Sequence analysis also had to compete with well-established molecular techniques, such as the immunological measures used by Morris Goodman, Allan Wilson, Vincent Sarich and others to unravel phylogenetic relationships[23]. Encounters among competing groups of biologists at professional meetings could be bruising, but these confrontations should not eclipse the important synthesis of evolutionary biology, protein biochemistry and computer science that was beginning to emerge during the early 1960s, which laid an evolutionary foundation for the bioinformatics of today[24–26].

Emergence of computational biology

Historical studies of protein biochemistry have emphasized the importance of instrumentation[8,9,19] but, with the exception of John Kendrew's use of computers for elucidating the three-dimensional structure of myoglobin[18,27], the historical role of digital computers during the 1960s has been virtually ignored. Even in the case of Kendrew, computers have not been viewed as contributing decisively to the discovery process. Nonetheless, digital computers were well suited to deal with the types of numerical data that protein biochemists were generating in growing amounts.

By the early 1960s, computers were becoming widely available to academic researchers. According to surveys conducted at the beginning of the decade, 15% of colleges and universities in the United States had at least one computer on campus, and most principal research universities were purchasing so-called 'second generation' computers, based on transistors, to replace the older vacuum-tube models[28]. The first high-level programming language, FORTRAN (formula translation), was introduced by the International Business Machines (IBM) corporation in 1957. It was particularly well suited to scientific applications, and compared with the earlier machine languages, it was relatively easy to learn. For the first time, detailed knowledge of computer architecture was not needed to write a computer program. This important innovation in computer software encouraged the growth of computational biology. At the same time, there was a concerted effort by governmental agencies and the computer industry to foster the development of academic computing in the life sciences[29,30].

The attraction of computers is well illustrated by the career of Margaret Oakley Dayhoff[31,32]. Trained in quantum chemistry and mathematics, she became interested in proteins and molecular evolution around 1960. As associate director of the newly established National Biomedical Research Foundation, an organization founded specifically to encourage the development of computer applications, Dayhoff was well situated to explore mathematical approaches for analysing amino-acid sequence data (FIG. 4).

Continuously funded by grants from the National Institutes of Health throughout the 1960s and with further support from the National Science Foundation, the National Aeronautics and Space Administration, and the IBM corporation, Dayhoff moved on several fronts. Her initial project was writing a series of FORTRAN programs to determine the amino-acid sequences of protein molecules[33,34]. Taking the overlapping peptide fragments from the partial digestion of a protein, the programs deduced all of the possible sequences that were consistent with the data. Conceptually similar to the puzzle-solving approach that biochemists claimed to have used in the early sequencing investigations of insulin and ribonuclease[10,13,14,35], Dayhoff's computer programs arrived at the correct sequence for a small protein (ribonuclease) within a few minutes. The same feat had taken a team of humans many months to accomplish. Similar programs written by other computational biologists at about the same time claimed to successfully sequence hypothetical proteins up to 750 amino acids in length[36]. Significantly, even during the early development of these programs, Dayhoff and her contemporaries realized that the logic of sequence analysers could also be directly applied to nucleic acids when gene sequences finally became available.

Computer programs for sequence analysis followed the principles initiated by the automatic amino-acid analyser used by Stein and Moore[37,38]. In both cases, the objective was to develop quickly a library of sequences that could be used for studies in comparative biochemistry and molecular evolution. To promote this end, Dayhoff established the *Atlas of Protein Sequence and Structure*, an annual publication that attempted to catalogue all known amino-acid sequences. Although rudimentary by today's standards, the *Atlas* served as the first database for molecular biology, and it became an indispensable resource for early computational research[25,31,32,39]. It eventually evolved into a major online database, the Protein Information Resource (PIR), established in 1983, and it provided an important point of departure for other com-

Glossary

CHROMATOGRAPHY
A chemical analysis technique that uses a process of separating gases, liquids or solids from mixtures or solutions by selective adsorption.

CYTOCHROMES
Proteins whose function is to carry electrons or protons (hydrogen ions) by virtue of the reversible charging/discharging of an iron atom or iron/sulphur atoms in the centre of the protein. Cytochromes are central molecules of electron transport in the process known as oxidative phosphorylation. Cytochromes are divided into four groups (*a*, *b*, *c*, *d*) according to their ability to absorb or transmit certain colours of light.

HAEMOGLOBIN
Protein present in red blood cells that reversibly binds oxygen for transport to tissues.

INSULIN
A protein hormone secreted by β cells of the pancreas. Insulin is important in the regulation of glucose metabolism, generally promoting the cellular use of glucose. It is also an important regulator of protein and lipid metabolism. Insulin is used as a drug to control insulin-dependent diabetes mellitus.

MOLECULAR CLOCK
The hypothesis that, in any given gene or DNA sequence, mutations accumulate at an approximately constant rate in all evolutionary lineages as long as the gene or the DNA sequence retains its original function.

MYOGLOBIN
An oxygen-carrying muscle protein that makes oxygen available to the muscles for contraction.

RIBONUCLEASE
A enzyme that hydrolyses RNA.

X-RAY CRYSTALLOGRAPHY
Study of the molecular structure of crystalline compounds through X-ray diffraction techniques. When an X-ray beam bombards a crystal, the atomic structure of the crystal causes the beam to scatter (diffract) in a specific pattern. X-ray crystallography provides information on the positions of individual atoms in the crystal, the distances between atoms, the angles of the atomic bonds and other features of molecular geometry.

Figure 4 | **The IBM 7090 computer, which Margaret Dayhoff used for her early work.** This famous computer was one of the first solid-state machines and was used widely in business and defence settings, as well as scientific applications.
(Photograph courtesy of IBM archives.)

putational biologists, who soon began building their own molecular databases[40].

Critics have pointed out that most of the entries in Dayhoff's *Atlas* were interspecific variations of a few well-studied proteins such as cytochrome *c*, and that the number of known protein sequences remained fairly small throughout the 1960s (REF. 40). What should not be overlooked, however, is how the early comparative studies of homologous proteins opened up the field of molecular evolution. Although phylogenetic analysis of amino-acid sequences could be done by hand[41], computers proved immensely valuable in this regard. From the beginning, theoretical biologists realized that in most cases the number of possible phylogenetic trees was so great that it would be infeasible for a human to evaluate even a small fraction of them. If every amino acid in even a small protein was to be considered a separate character, then finding the most likely tree was clearly an appropriate task for digital computers. Early molecular evolutionists such as Russell Doolittle, who began studying protein phylogenies without computers, quickly added them to their research programmes by the late 1960s.

The potential for using computers for phylogenetic analysis was dramatically demonstrated for cytochrome *c*, the respiratory pigment found in all aerobic cells. By the mid-1960s, the protein had been sequenced for a wide variety of plants, animals, fungi and microbes. In a now classic article, Walter Fitch and Emanuel Margoliash showed how this data could be used to build a phylogenetic tree that was remarkably similar to those based on more traditional taxonomic characters[42]. In this, and in the similar computer programs concurrently devised by Dayhoff's team[43,44], pairwise comparisons were made among homologous amino-acid sites on CYTOCHROMES drawn from 20 species. The computer calculated the mutation distances, or the minimum number of steps required to convert one cytochrome to another. Starting with a simple three-branched tree, subsequent branches were added sequentially in a way that would minimize the mutation distances. These early computer programs did not attempt exhaustive searches for the simplest phylogenetic tree, but left this partly to human intuition. The investigator chose a different initial subset of three proteins, then the computer constructed a second tree. The new tree was discarded if it turned out to be less satisfactory than the original one. During the research leading to their article, Fitch and Margoliash examined 40 trees in an attempt to find the optimal one.

The molecular phylogenies constructed by early computational biologists rested on the assumption that the proteins being compared were homologous. In cytochrome *c* trees, the proteins from various species were so similar that there was no question that they shared a close common ancestor. However, detecting homology, and distinguishing it from chance similarity, in more distantly related proteins was recognized as an important problem by molecular evolutionists. During the late 1960s, several biologists developed computer algorithms for determining sequence homology and aligning related sequences to account for deletions or insertions[43,45,46].

Walter Fitch's moving-window approach searched for nonrandom alignments by comparing all possible combinations of sequences of a given length (say 30 amino acids) along two protein molecules. For each of the thousands of comparisons, the computer calculated the minimum number of mutations required to convert one sequence to the other. This was done using a matrix of the number of mutations needed to substitute one amino acid for any other on the basis of the genetic code. The computer was instructed to print out all sequences whose similarity could not be accounted for by chance. Fitch's approach was further elaborated by others, notably Saul Needleman and Christian Wunsch, whose algorithm remains one of the standard methods for sequence alignment[47]. The computational approach used by Dayhoff and Richard Eck was similar, but their MDM (mutation data matrix) or PAM (per cent accepted mutation) matrices were based on the probability of substituting a given amino acid with any other. These probabilities were estimated by counting the occurrence of each amino-acid substitution in families of very similar proteins (for example, cytochromes or haemoglobins) taken from the *Atlas of Protein Sequence and Structure*. The matrices proved useful for more distantly related proteins as well, and quickly became standard tools for detecting homology and aligning sequences. PAM matrices continue to be used today and have stimulated the development of several more refined methods[3].

Even the most challenging computational problems in bioinformatics had precursors in the 1960s. For example, the biophysicist Cyrus Levinthal and a team of researchers[48] used one of the first large, time-sharing computers at the Massachusetts Institute of Technology to construct three-dimensional models of cytochrome *c*. A visual display of the molecule was projected on an oscilloscope screen. Researchers could control the rotation of the model using a hand-operated device similar to a track ball and could manipulate the model using either a keyboard or a light pen. Because of limitations in the speed of even the most powerful computers of the day, this early modelling effort did not have an immediate impact on biochemistry or molecular biology. However, it was an important historical bridge between earlier mechanical and stereoscopic modelling techniques and the advanced computer models of today.

Conclusions

By 1970, computational biologists had developed a diverse set of techniques for analysing molecular structure, function and evolution. Although originally designed for studying proteins, many of these computing techniques could be adapted for studying nucleic acids. Some of these techniques survive today or have lineal descendants that are used in bioinformatics. In other cases, they stimulated the development of more refined techniques to correct deficiencies in the original methods. Although the nascent field was later revolutionized by the advent of genome projects, large-scale computer networks, immense databases, supercomputers and powerful desktop computers, today's bioinformatics also rests on the important intellectual and technical foundations laid by scientists at an earlier period in the computer era.

Joel B. Hagen is at the Department of Biology, Radford University, Radford, Virginia 24142, USA. e-mail: jhagen@radford.edu

Links

DATABASE LINKS insulin | ribonuclease | myoglobin | haemoglobin | Protein Information Resource | cytochrome *c*
FURTHER INFORMATION IBM history | National Biomedical Research Foundation | History of visualization of biological macromolecules
ENCYCLOPEDIA OF LIFE SCIENCES DNA sequencing | Sanger, Frederick | Protein secondary structures: predictions | Molecular clocks | Linus Carl Pauling

1. Lake, J. A. & Moore, J. E. Phylogenetic analysis and comparative genomics. *Trends Guide Bioinformatics* 22–23 (1998).
2. Howard, K. The bioinformatics gold rush. *Sci. Am.* **283**, 58–63 (2000).
3. Rashidi, H. H. & Buehler, L. K. *Bioinformatics Basics: Applications in Biological Science and Medicine* (CRC Press, Boca Raton, 2000).
4. Boguski, M. S. Bioinformatics — a new era. *Trends Guide Bioinformatics* 1–3 (1998).
5. Kay, L. Who wrote the book of life? Information and the transformation of molecular biology, 1945–1955. *Science Cont.* **8**, 609–634 (1995).
6. Kay, L. Cybernetics, information, life: The emergence of scriptural representations of heredity. *Configurations* **5**, 23–91 (1997).
7. Sarkar, S. in *The Philosophy and History of Molecular Biology: New Perspectives* (ed. Sarkar, S.) 187–231 (Kluwer Academic, Dordrecht, 1996).
8. Smith, E. L. in *The Origins of Modern Biochemistry: A Retrospective on Proteins* (eds Srinivasan, P. R., Fruton, J. S. & Edsall, J. T.) 107–118 (New York Academy of Sciences, New York, 1979).
9. Fruton, J. S. *Proteins, Enzymes, Genes: The Interplay of Chemistry and Biology* (Yale Univ. Press, New Haven, 1999).
10. Sanger, F. Chemistry of insulin. *Science* **129**, 1340–1344 (1959).
11. Zamecnik, N. H. in *The Origins of Modern Biochemistry: A Retrospective on Proteins* (eds Srinivasan, P. R., Fruton, J. S. & Edsall, J. T.) 269–294 (New York Academy of Sciences, New York, 1979).
12. Fruton, J. S. *A Skeptical Biochemist* (Harvard Univ. Press, Cambridge, Massachusetts, 1992).
13. Stein, W. H. & Moore, S. The chemical structure of proteins. *Sci. Am.* **204**, 81–92 (1961).
14. Moore, S. & Stein, W. H. Chemical structures of pancreatic ribonuclease and deoxyribonuclease. *Science* **180**, 458–464 (1973).
15. Edman, P. & Begg, G. A protein sequenator. *Eur. J. Biochem.* **1**, 80–91 (1967).
16. Sanger, F. Sequences, sequences, and sequences. *Annu. Rev. Biochem.* **57**, 1–28 (1988).
17. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **161**, 223–230 (1973).
18. Olby, R. C. The 'Mad Pursuit': X-Ray crystallographers' search for the structure of hemoglobin. *Hist. Phil. Life Sci.* **7**, 171–193 (1985).
19. Kay, L. *The Molecular Vision of Life* (Oxford Univ. Press, New York, 1993).
20. Srinivasan, P. R., Fruton, J. S. & Edsall, J. T. (eds) *The Origins of Modern Biochemistry: A Retrospective on Proteins* (New York Academy of Sciences, New York, 1979).
21. Zuckerkandl, E. & Pauling, L. Molecules as documents of evolutionary history. *J. Theor. Biol.* **8**, 357–366 (1965).
22. Zuckerkandl, E. On the molecular evolutionary clock. *J. Mol. Evol.* **26**, 34–64 (1987).
23. Dietrich, M. Paradox and persuasion. Negotiating the place of molecular evolution within evolutionary biology. *J. Hist. Biol.* **31**, 85–111 (1998).
24. Hagen, J. B. Naturalists, molecular biologists, and the challenges of molecular evolution. *J. Hist. Biol.* **32**, 321–341 (1999).
25. Jungck, J. R. & Friedman, R. M. Mathematical tools for molecular genetics data: An annotated bibliography. *Bull. Math. Biol.* **46**, 699–744 (1984).
26. Hagen, J. B. The introduction of computers into systematic research in the United States during the 1960s. *Studies Hist. Phil. Biol. Biomed. Sci.* (in the press).
27. Perutz, M. Early days of protein crystallography. *Meth. Enzymol.* **114**, 3–18 (1985).
28. Anonymous. Computing in the university. *Datamation* **8**, 27–30 (1962).
29. Ledley, R. S. Digital electronic computers in biomedical sciences. *Science* **130**, 1225–1234 (1959).
30. Ledley, R. S. *Use of Computers in Biology and Medicine* (McGraw–Hill, New York, 1965).
31. Hunt, L. T. Margaret O. Dayhoff, 1925–1983. *DNA* **2**, 87–98 (1983).
32. Hunt, L. T. Margaret Oakley Dayhoff, 1925–1983. *Bull. Math. Biol.* **46**, 467–472 (1984).
33. Dayhoff, M. O. & Ledley, R. S. Comprotein: A computer program to aid primary protein structure determination. *Proc. Fall Joint Comp. Conf.* **22**, 262–274 (1962).
34. Dayhoff, M. O. Computer aids to protein sequence determination. *J. Theor. Biol.* **8**, 97–112 (1965).
35. Thompson, E. O. The insulin molecule. *Sci. Am.* **192**, 36–41 (1955).
36. Bernhard, S. A., Bradley, D. F. & Duda, W. L. Automatic determination of amino acid sequences. *IBM J. Res. Dev.* **7**, 246–251 (1963).
37. Spackman, D. D., Stein, W. H. & Moore, S. Automatic recording apparatus for use in the chromatography of amino acids. *Anal. Chem.* **30**, 1190–1206 (1958).
38. Mason. E. E. & Bulgren, W. G. *Computer Applications in Medicine* (Charles C. Thomas, Springfield, Illinois, 1964).
39. Fitch, W. M. Book review of M. O. Dayhoff, Atlas of Protein Sequence and Structure. *Syst. Zool.* **22**, 196 (1972).
40. Doolittle, R. F. Some reflections on the early days of sequence searching. *J. Mol. Med.* **75**, 239–241 (1997).
41. Doolittle, R. F. & Blombäck, B. Amino-acid sequence investigations of fibrinopeptides from various mammals: Evolutionary implications. *Nature* **202**,147–152 (1964).
42. Fitch, W. M. & Margoliash, E. Construction of phylogenetic trees. *Science* **155**, 279–284 (1967).
43. Eck, R. V. & Dayhoff, M. O. *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Silver Spring, Maryland, 1966).
44. Dayhoff, M. O. Computer analysis of protein evolution. *Sci. Am.* **221**, 87–95 (1969).
45. Fitch, W. M. An improved method of testing for evolutionary homology. *J. Mol. Biol.* **16**, 9–16 (1966).
46. Dayhoff, M. O. & Eck, R. *Atlas of Protein Sequence and Structure 1967–1968* (National Biomedical Research Foundation, Silver Spring, Maryland, 1968).
47. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
48. Levinthal, C. Molecular model-building by computer. *Sci. Am.* **214**, 42–52 (1966).

TIMELINE

Genetic disease since 1945

M. Susan Lindee

Although hereditary disease has been recognized for centuries, only recently has it become the prevailing explanation for numerous human pathologies. Before the 1970s, physicians saw genetic disease as rare and irrelevant to clinical care. But, by the 1990s, genes seemed to be critical factors in virtually all human disease. Here I explore some perspectives on how and why this happened, by looking at two genetic diseases — familial dysautonomia and phenylketonuria.

Human hereditary disease has a long recorded history but, for most of the twentieth century, geneticists knew more about hereditary pathology in the mouse or the fruitfly than in human beings. In the past thirty years, however, pathology as related to our genes has become the focus of intense medical, scientific and corporate interest. Many complicated bodily states have been reconfigured as genetic diseases and, by the 1990s, heredity was the prevailing explanation for virtually all disease states. Even infectious disease has been rhetorically subsumed under genetic mechanisms of inherent vulnerability and genetically driven immune system responses. The model in which all bodily states have an underlying genetic cause dominates concepts of disease in the industrialized world, particularly in English-speaking nations and even more so in the United States.

One possible way to understand the new centrality of genetic disease to scientific research, medical education, medical theory and the broader culture would be to