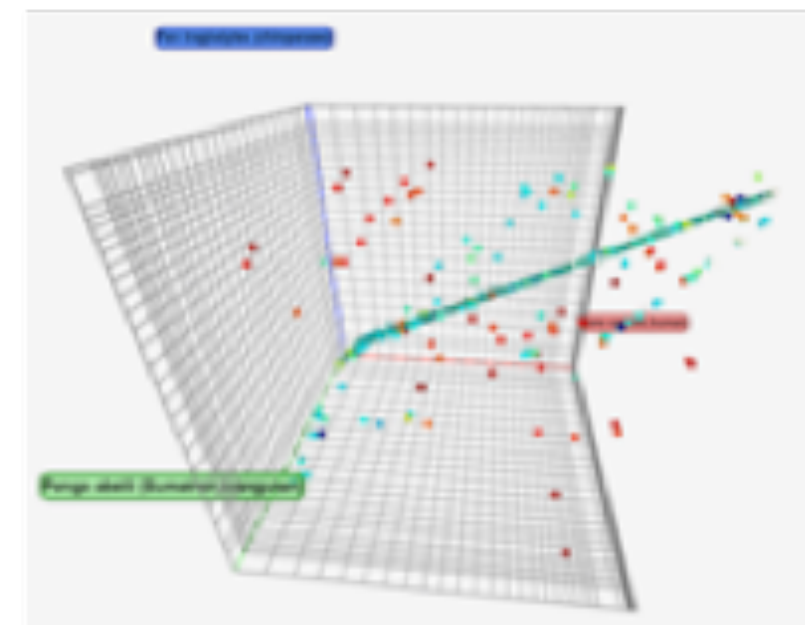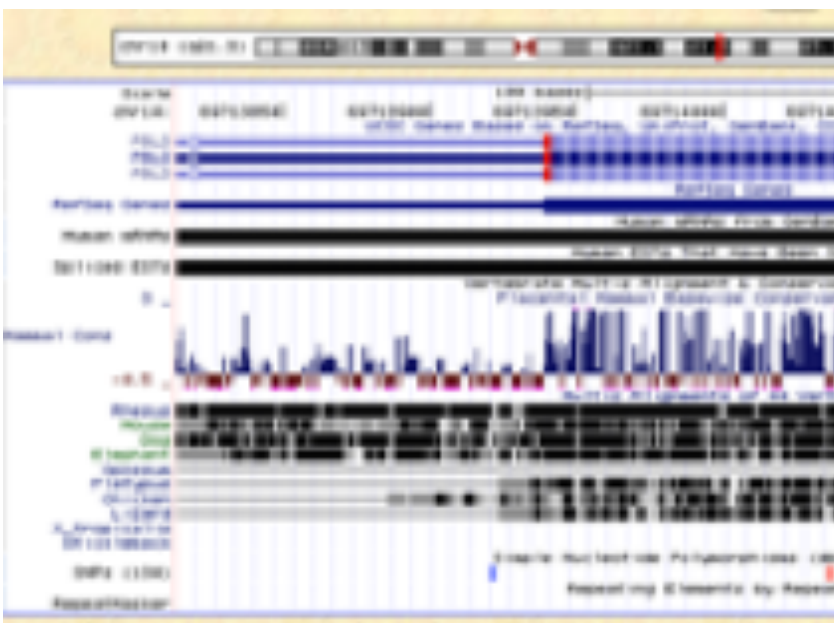# Computational Genomics

History of Bioinformatics

The origins of bioinformatics

TIMELINE

The origins of bioinformatics

Joel B. Hagen

All slide that follow are based on the timeline article "*The Origins of Bioinformatics*" linked above

The origins of bioinformatics

- **Computers emerged as important tools in molecular biology during the early 1960s.**

- **Computers were important tools in molecular biology a decade before DNA sequencing became feasible.**

- **Access to computers was SLOW, by the early 1960's the scarcity of computers was no longer a serious stumbling block for the development of computational biology.**

The origins of bioinformatics

- **A breakthrough event was the first successful sequencing of the protein INSULIN (51 aar), by Frederick Sanger and his colleagues at Cambridge University during the decade 1945–1955.**
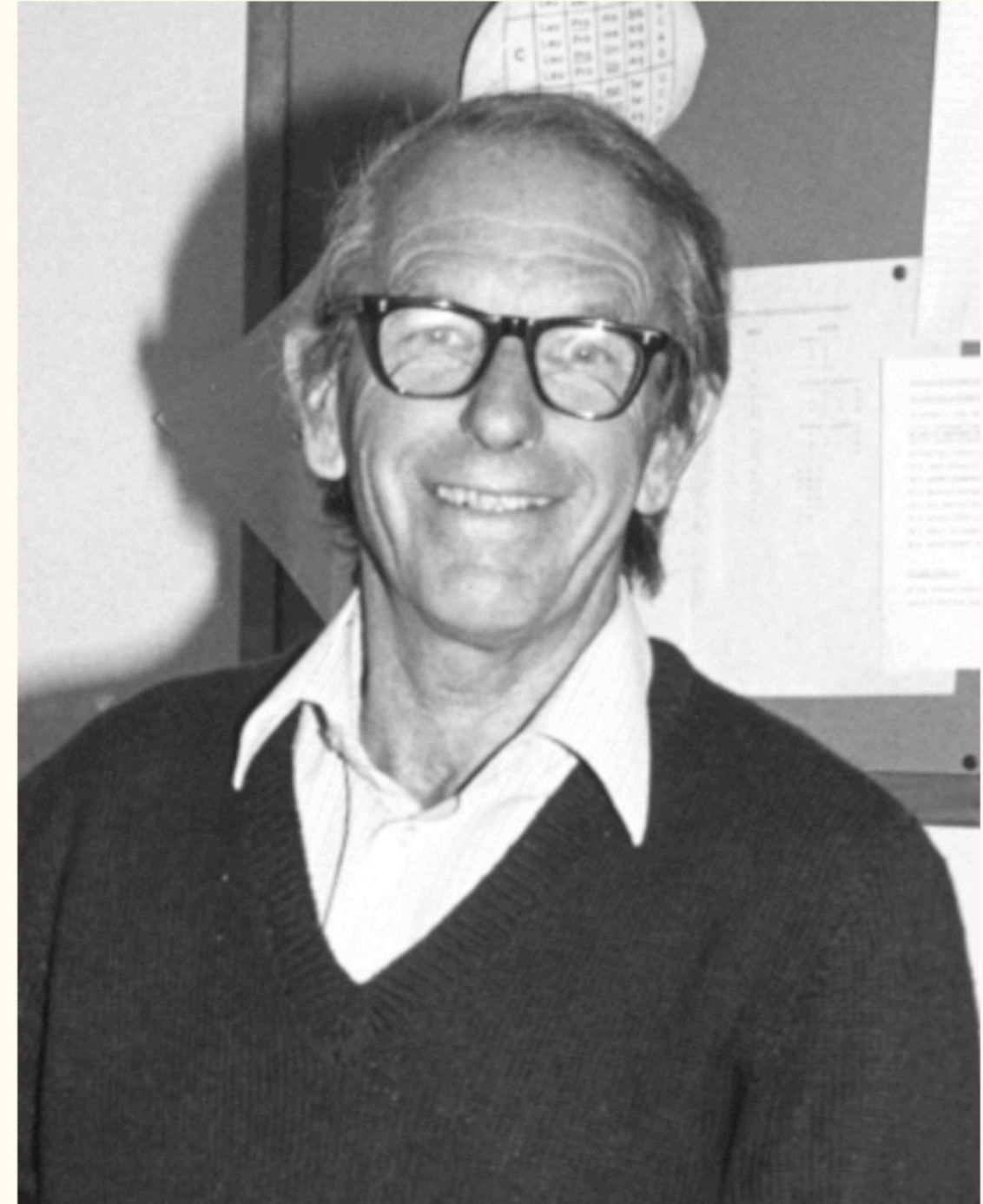

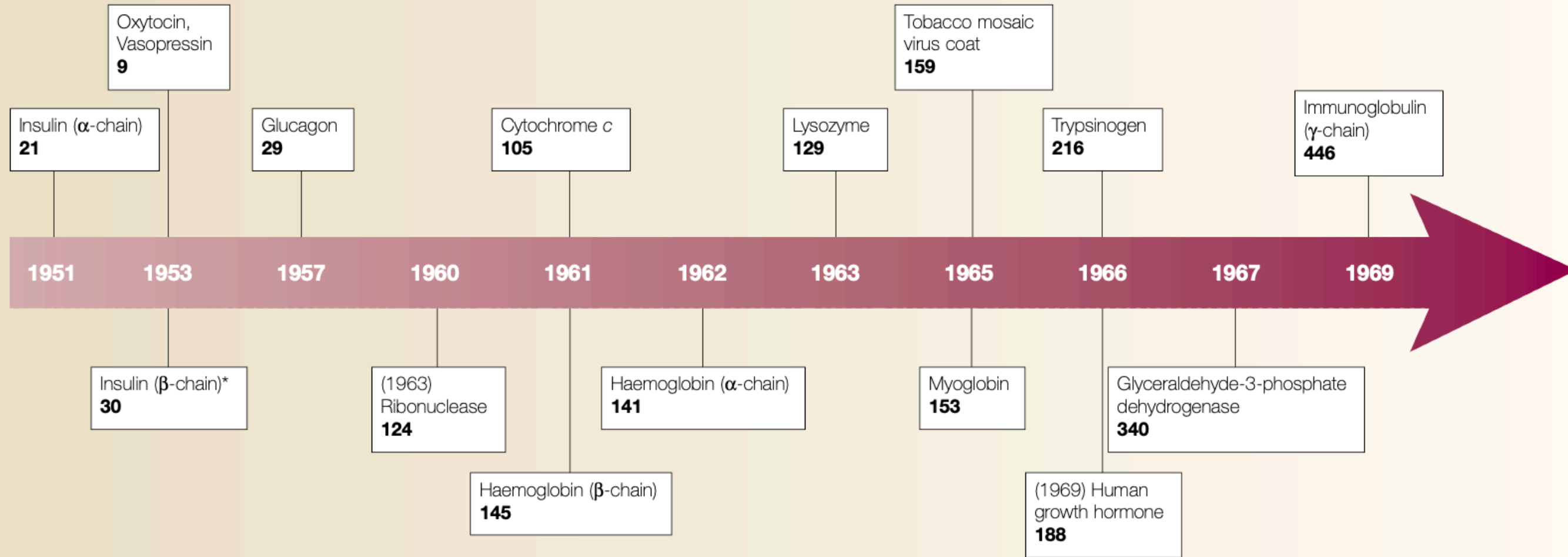Figure 1 | **Frederick Sanger at the Nobel prize ceremony in 1980.**
(Photograph kindly provided by the MRC, Laboratory of Molecular Biology, Cambridge, UK.)

The origins of bioinformatics

- **Using semi-automated techniques, researchers led by Stanford Moore and William Stein at the Rockefeller Institute were able to sequence the 124 aar in RIBONUCLEASE in about half the time that Sanger's group had spent deciphering the sequence of the 51 aar in insulin.**

- **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.**

- **By the late 1960s, Pehr Edman had designed the 'sequenator', a fully automated sequencing machine.**

The origins of bioinformatics

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

Oxytocin, Vasopressin
9

Tobacco mosaic virus coat
159

Insulin (α-chain)
21

Glucagon
29

Cytochrome *c*
105

Lysozyme
129

Trypsinogen
216

Immunoglobulin (γ-chain)
446

1951 — 1953 — 1957 — 1960 — 1961 — 1962 — 1963 — 1965 — 1966 — 1967 — 1969

Insulin (β-chain)*
30

(1963) Ribonuclease
124

Haemoglobin (α-chain)
141

Myoglobin
153

Glyceraldehyde-3-phosphate dehydrogenase
340

Haemoglobin (β-chain)
145

(1969) Human growth hormone
188

*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).

The origins of bioinformatics

- **Protein sequencing firmly established the polypeptide theory.**

- **This opened the following areas:**

  - **The relationship of protein structure and the genetic code.**

  - **Relationship between sequence to structure to function.**

  - **Relationship between the three dimensional structure of a protein in relation to its function.**

    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.

The origins of bioinformatics

- Sequence data played a key role in interpreting the X-ray diffraction images used by John Kendrew and Max Perutz to determine the three-dimensional structures of Myoglobin and Hemoglobin.

- Combining sequence analysis with X-Ray Crystallography were key to understand how a sequence of amino acids causes a protein to fold into a specific, highly complex, three-dimensional configuration.
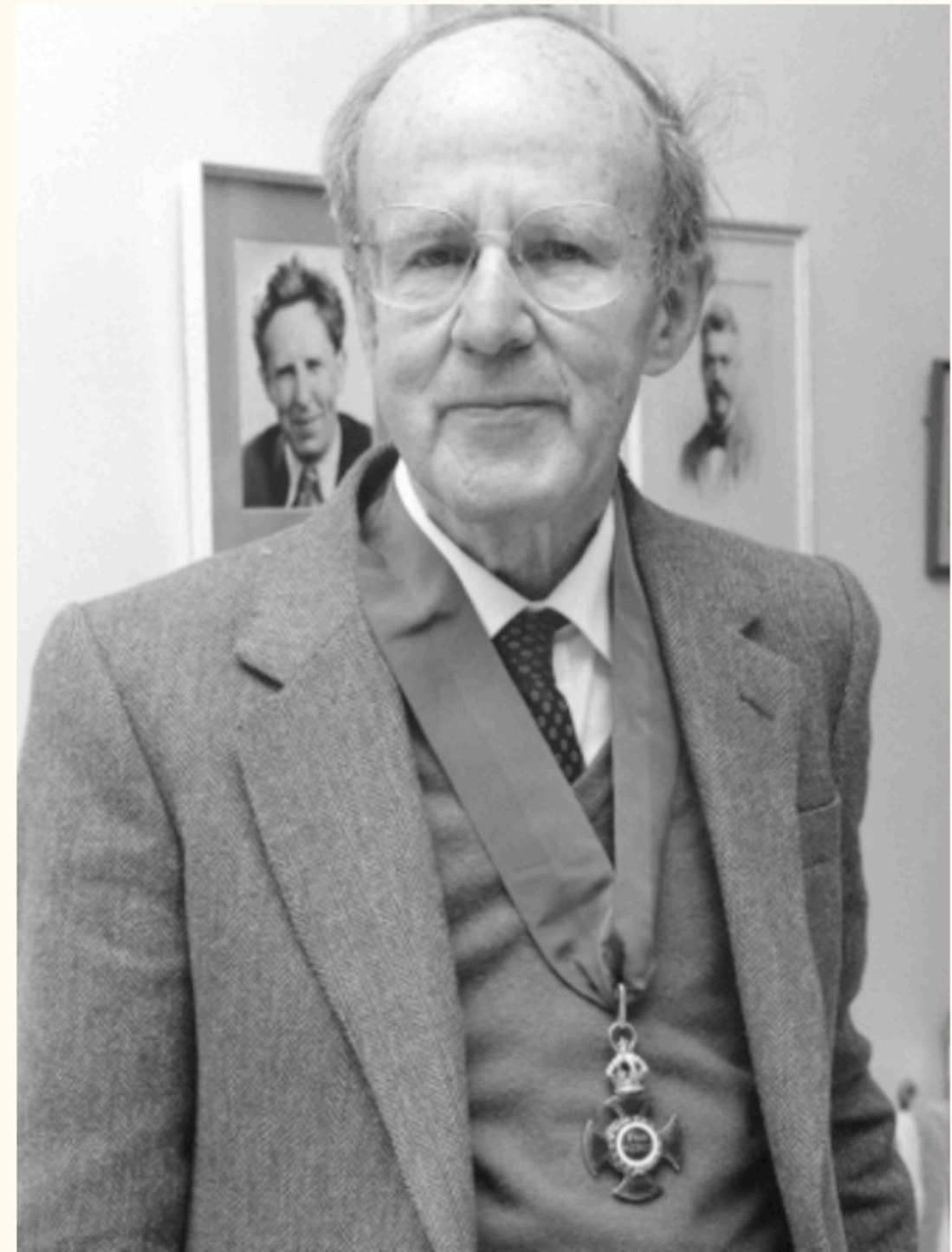
Figure 3 | **Max Perutz, who shared the 1962 Nobel prize in chemistry with John Kendrew.** (Photograph kindly provided by the MRC, Laboratory of Molecular Biology, Cambridge, UK.)

The origins of bioinformatics

- **Protein Evolution**
  - **During the 1960s, biochemists and molecular biologists were increasingly drawn to evolutionary questions.**
  - **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.**
  - **Derived from '*semanteme*', the fundamental unit of meaning used by linguists to study human speech, *semantides* were to be the analogous biochemical units for evolutionary studies.**

The origins of bioinformatics

- **Protein Evolution**
  - **From this, the field of *Paleogenetics* was proposed.**
  - **Paleogeneticists are now know as Molecular Evolutionists.**
  - **Comparisons of similar proteins, such as myoglobin and hemoglobin, provided evidence for molecular evolution by gene duplication.**
  - **It was proposed that 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.**
  - **Thus, the field of Phylogenetic was born.**

The origins of bioinformatics

- **Emergence of computational biology**

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.)

The origins of bioinformatics

- **Emergence of computational biology**

  - 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.

The origins of bioinformatics

- **Emergence of computational biology**

  - **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**

- **Margaret Oakley Dayhoff**

  - **She wrote FORTRAN programs to determine the amino-acid sequences of protein molecules**

  - **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.**