

## Earliest pages of bioinformatics

Edward N. Trifonov

Department of Structural Biology, The Weizmann Institute of Science,  
Rehovot 76100, Israel

### Abstract

*This review is a brief outline of the chronology and essence of early events in bioinformatics, covering the period from 1869 (discovery of DNA by Miescher) to 1980–1981 (beginning of massive sequencing). For the purpose of this review, bioinformatics is understood as a chapter of molecular biology dealing with the amino acid and nucleotide sequences and with the information they carry.*

**Contact:** [edward.trifonov@weizmann.ac.il](mailto:edward.trifonov@weizmann.ac.il)

### The nature of hereditary material

The role of proteins in life was already understood at the beginning of the 19th century when Berzelius coined the very term ‘protein’ (Greek ‘*proteios*’), as basic elementary material of living matter. A linear text-like primary structure of proteins was originally hypothesized by Curtius (1883). A more advanced formulation of this ‘peptide theory’ can be found in Hofmeister (1902) and Fischer (1902). With progress in polymer science this idea eventually developed into a broadly accepted view, but only in 1947, with the first sequencing of a pentapeptide of gramicidine S by Consden *et al.* (1947) and reconstruction of partial 30 residue sequence of insulin by Sanger and Tuppy (1951), did the linear, sequential structure of proteins become an established fact.

The hereditary functions were initially ascribed by various schools to proteins and chromatin, at that time of unclear composition. The genetic function of DNA was already appreciated by Miescher in 1869 when he discovered DNA (published with a delay, in 1871; see references in Davidson, 1968). In 1892 he also expressed the idea that the genetic information may exist in the form of a mere molecular text, a linear sequence of chemical symbols (as referred to in Judson, 1979). He thought that a few small chemical units connected in large molecules may play hereditary role ‘just as the words and concepts of all languages can find expression in twenty-four to thirty letters of the alphabet’. Fifty years later Astbury and Bell (1938), who discovered a conspicuous 3.3 Å periodicity in the fiber X-ray diffraction of DNA suggestive of a linear stacking of 3.3 Å thick flat DNA bases, came to the same conclusion. They hypothesized that the bases ‘form the long scroll on which is written the pattern of life’. These were, however, isolated statements, and DNA was

not broadly recognized as genetic material until the early 1950s. For a long time (1906–1948) DNA was viewed as a monotonous repetition of identical tetranucleotide units (Steudel, 1906; Levene and Simms, 1925). The A=T and G=C rule discovered by Chargaff *et al.* (1949) was clear evidence that the DNA chains are not as primitive. The experiment of Avery *et al.* (1944) did establish the exclusive genetic role of DNA, although it took several years before this fact was independently confirmed and accepted. Interestingly, the tetranucleotide theory that denied any hereditary role for DNA continued to reverberate till at least 1963 when Salvador Dali completed his work that he called ‘GALACIDALACIDEOXYRIBUNUCLEICACID (homage to Crick and Watson)’, where the tetranucleotides were featured.

When the first sequences of nucleic acids appeared in the literature the genetic role of the DNA and RNA sequences was already fully recognized. First short sequences, of fragments of transfer RNA, appeared in the literature in 1961. The repertoire of the fragments gradually accumulated to the extent that complete sequences of tRNA could be assembled. That was done first by American teams, closely followed by German and Russian laboratories: Ala tRNA (Holley *et al.*, 1965), Tyr tRNA (Madison *et al.*, 1966), Ser tRNA (Zachau *et al.*, 1966), and Val tRNA (Baev *et al.*, 1967). The first short sequence of DNA, the restriction site GTYRAC recognized by the endonuclease R of *Hemophilus influenzae* (HindII), was derived by Kelly and Smith (1970). At about the same time the short sequence of sticky ends of the bacteriophage lambda DNA was published (Wu and Taylor, 1971).

### Translation (triplet) code

Long before the nucleotide sequences became available the question arose as to how the four-letter alphabet nucleotide sequences can encode the 20-letter alphabet protein sequences. The simple thought that at least some codons and, perhaps, all should contain three letters (Gamov, 1954) was inevitable. If this is a triplet code then what would be the explanation for the difference between the number of the amino acids and the total assortment of the triplets (20 and 64, respectively)? It suggested that many of the possible 64 triplets could be

synonymous, degenerate. Gamov described two ingenious schemes where 20 unique sets of three-letter combinations of four bases could be put in correspondence with 20 amino acids—the ‘diamond’ code and the ‘triangle’ code (Gamov, 1954, 1955). Crick *et al.* (1957) suggested yet another possible way of selection of unique 20 triplets, the ‘comma-free’ code. It had an important property of turning every triplet of the sequence into a nonsense when the reading frame changes. These schemes, however, did not survive long, having been definitely dismissed after spectacular unraveling of the actual triplet code, in 1961–1966, by Nirenberg, Ochoa, Khorana and their respective teams (Khorana *et al.*, 1966; Nirenberg *et al.*, 1966; Speyer *et al.*, 1963). The key discovery and the first codon assignment that led to the avalanche of all 64 codon assignments was the observation by Nirenberg and Matthaei (1961) that in a cell-free system poly-U directs synthesis of polyphenylalanine, thus suggesting the correspondence of UUU/phe.

The code indeed turned out to be highly degenerate, as Gamov expected. On the impressive background of the complete repertoire of the triplet assignments the *de facto* degeneracy of the code had now to be accepted as granted, no urgent explanation needed. Interestingly, the triplet code has been deciphered without any use of natural mRNA sequences, which were not available at that time. One only did what one could, and Nirenberg and Ochoa could *synthesize* the artificial mRNA sequences, rather than sequence the natural ones. The synthetic approach became possible due to the discovery of RNA-synthesizing enzyme by Grunberg-Manago *et al.* (1956). Only in 1969, with the advance in RNA sequencing the first comparison of a mRNA sequence with respective protein sequence also available at that time (the coat protein of bacteriophage R17) became possible, with spectacular confirmation of mRNA/protein co-linearity and of 30 of 64 codon assignments (Adams *et al.*, 1969).

The degeneracy problem was theoretically treated first by Schaap (1971) who suggested that apart from the translational triplet code there are, presumably, other codes, not related to the translation. Due to the degeneracy, various codes could co-exist in one and the same sequence by making suitable adjustments in the sequence allowed by the degenerate use of alternative triplets. One possible example of such overlapping has been mentioned *en passant* by R.Holliday in 1968: sequence-specific recombination signals in yeast, presumably located within the protein-coding sequences. Further developments on the additional codes and overlapping messages have been abandoned for a long time apparently due to the discovery of the phenomenon of gene splicing (see below). The seemingly meaningless intervening sequences that separate the protein-coding sequence sections in eukaryotes were unfortunately condemned right away to a non-functionality

(Doolittle and Sapienza, 1980; Orgel and Crick, 1980), denying the genomes any sophistication beyond their protein-coding capacity. Surely, if the cell could afford to carry such excessive ‘junk’, the information compression and superposition of various codes would not make sense. It is worth noting in this connection that Woese (1967) warned against the ‘popular misconception that ... a ‘cracking’ of the code—is all there is to the genetic code’.

### Sequencing of the nucleic acids

The first DNA and RNA sequencing efforts were a painful combination of chemical, enzymatic and spectral analysis techniques applied to every single base step, in order to identify and verify it unequivocally. These early composite techniques of sequencing, thus, could not be realistically extended to the genome scale. One of the earliest universal DNA sequencing techniques which consisted of rather limited number of elementary operations was based on repair synthesis (Wu and Kaiser, 1968). The genomic era of massive DNA sequencing started soon after, in 1975–1977, with the introduction of two rapid ‘read-the-gel’ techniques. The sequencing method of Sanger (Sanger and Coulson, 1975) is based on template DNA synthesis interrupted in random positions at one of four letters of choice, by limited supply of the respective dNTPs. The technique of Maxam and Gilbert (1977) uses specific chemical reactions to break the end-labeled DNA molecule at one of four bases. Both techniques generate four series of single-stranded DNA fragments with common start, ending at respective four bases wherever the bases are encountered in the sequence. The comparatively large sequence fragments (up to a few hundred bases) could be literally read from the sequence of bands in the gel. Each band would thus correspond to a specific subfragment of the sequence with a specific base at the end, indicated by the location of the band in one of four lanes of the gel. Both techniques, modified and automated, are intensively used today, sequencing thousands of bases daily by one machine. Considering ongoing rapid progress in the sequencing, this should not impress the reader. The rate may quickly change. In 1978 R.Wu wrote: ‘Today a DNA sequence of 200 nucleotides can be determined within a month’.

Using the ‘read-the-gel’ techniques virtually length-independent sequencing became possible, and very soon the first viral genome had been completed, that of RNA bacteriophage MS2 (3569 bases) (Fiers *et al.*, 1976). Several more viral genomes closely followed: bacteriophage  $\phi$ X174, single-stranded DNA of 5375 bases (Sanger *et al.*, 1977); bacteriophage fd, 6408 bases (Beck *et al.*, 1978); bacteriophage G4, 5577 bases (Godson *et al.*, 1978); and of simian virus 40, 5256 base pairs of double-stranded DNA (Fiers *et al.*, 1978; Reddy *et al.*,

1978). The sequencing of the  $\phi$ X174 genome revealed with certainty that overlapping of various messages does indeed occur. In two cases the same sequence within the genome encoded two different proteins, in different reading frames.

### Sequence research

With the emergence of the sequences, they themselves became a rich source of new information by the application of various computational techniques. Similarity of related or homologous protein sequences was immediately noticed. An important connection of the variability of the sequences with their evolutionary relations was made in 1962 by Zuckerkandl and Pauling, who initiated a whole new field—molecular evolution. Sequence comparisons and establishment of their functional and taxonomic relatedness (molecular clock) soon became a major type of sequence analysis (see, for example, Nolan and Margoliash, 1968) routinely applied now to every new sequence.

The high degree of divergence of the sequences that parted long ago in evolution makes it difficult to quantitatively establish their relatedness and evaluate the time elapsed since their separation. As M. Dayhoff originally observed (1969) many amino acids are replaced in evolution not in a random way but with rather specific preferences. For example, hydrophobic residues are more frequently replaced by other hydrophobic residues. Introduction of the Dayhoff PAM (Point Accepted Mutations) matrices is considered to substantially improve the sequence comparisons. Another important seminal contribution to the problem of sequence comparisons is the work of Needleman and Wunsch on sequence alignment (1970).

The first information-theoretical treatment of the sequences was offered by Gatlin (1972). Starting with Shannon's definition of information, Gatlin introduced logarithmic measures of sequence divergence  $D_1$  from equiprobability and divergence  $D_2$  from independence between neighboring bases. This approach provided the first quantitative evidence that natural sequences are highly non-random.

### Gene regulation

A general logic of the sequence organization in prokaryotes became clear after experimental studies that led to the discovery of regulatory components of lac operon of *Escherichia coli* and their interactions (Jacob and Monod, 1961). The suggested regulatory scheme was later fully confirmed, also at the sequence level, and thus the theory of Jacob and Monod became a major landmark in molecular biology. One important result that emerged from this theory and from respective experimental work is the dis-

covery of several classes of regulatory sequences (operators, promoters, terminators) which are physically separate from the protein-coding sequences. These prokaryotic regulatory sequences are the first obvious demonstration that the 'non-coding' sequences are not necessarily 'junk'.

An important step in the studies on gene regulation in bacteriophages was the discovery of the timing program in the transcription—early and late mRNA syntheses (Khesin *et al.*, 1962). This initiated numerous works on viral strategies. The phenomenon of the transcription timing turned out later to be truly general. That is, it is typical not only for viruses, but for all prokaryotes and eukaryotes as well.

The earliest general theoretical treatment of the molecular genetic regulatory systems was carried out by Ratner (1974). This work was largely based on ideas of Jacob and Monod. It was not known at that time that eukaryotic regulatory systems are rather different.

### Gene splicing

A surprising and, more than 20 years later, still enigmatic discovery was made in 1976–1977, when several research groups almost simultaneously observed genes interrupted by inclusions that did not appear in the final processed (spliced) RNA. These are interrupted genes for eukaryotic rRNA (Glover and Hogness, 1977; Pellegrini *et al.*, 1977; Wellauer and Dawid, 1977), for mRNA of eukaryotic viruses (Aloni *et al.*, 1977; Berget *et al.*, 1977; Celma *et al.*, 1977; Chow *et al.*, 1977; Dunn and Hassell, 1977; Klessig, 1977) and for tRNA genes (Goodman *et al.*, 1977). Later a massive mRNA splicing was also found in eukaryotes in general. The prophecy of Monod that 'what was true for *E. coli*, would also be true for the elephant' (as referred to in Judson, 1979, p. 613) lost its attractive generality, since the 'elephants', as it turned out, have principally different gene structure compared to prokaryotes.

Two theories were suggested at that time that provide some explanation for the phenomenon of gene splicing. One, by Gilbert (1978), suggests that the intervening sequences have been introduced at some moment in evolution in order to facilitate recombinational transfer of the coding sections (exons), that is, deletions and reinsertions of exons to new locations ('exon shuffling'). Another theory (Zuckerkandl, 1981) is based on the recognition of the important roles chromatin structure plays in the eukaryotic cell. It argues that the primary function of the intervening sequences is the organization of the chromatin structure of the gene, while the exons are primarily geared to the protein-coding function. These two functions would interfere and compromise each other if not spatially separated.

## Conclusion

The outline above ends with the reference to the work of 1981, time-wise the last one in the list of references below. The author's intention was to describe only the very first pages of the history of molecular biology that led to bioinformatics, the intellectual and experimental landmarks in developments of all major aspects of molecular biology and genetics of those early days, relevant to modern bioinformatics. Since then this field expanded immensely, triggered by the onset of large scale sequencing in 1980–1981. A review on modern bioinformatics, in all its walks, is rather an impossible task, in fact, several reviews would be required.

A few omissions of important names and developments are, perhaps, inevitable in such reviews. If, indeed, there are some, they are unintentional.

For additional reading about early stages of this discipline the following sources are suggested: Crick (1966); Woese (1967); Davidson (1968); Hess (1970); Olby (1974); Wu (1978); Judson (1979) and Smith (1979).

## References

- Adams,J.M., Jeppesen,P.G.N., Sanger,F. and Barrell,B.G. (1969) Nucleotide sequence from the coat protein cistron of R17 bacteriophage RNA. *Nature*, **223**, 1009–1114.
- Aloni,Y., Dhar,R., Laub,O., Horowitz,M. and Khoury,G. (1977) Novel mechanism for RNA maturation: the leader sequences of simian virus 40 mRNA are not transcribed adjacent to the coding sequences. *Proc. Natl. Acad. Sci. USA*, **74**, 3686–3690.
- Astbury,W.T. and Bell,F.O. (1938) Some recent developments in the x-ray study of proteins and related structures. *Cold Spring Harb. Symp. Quant. Biol.*, **6**, 109–118.
- Avery,O.T., MacLeod,C.M. and McCarty,M. (1944) Studies on the chemical transformation of pneumococcal type. *J. Exp. Med.*, **79**, 137–158.
- Baev,A.A., Venkster,T.V., Mirzabekov,A.D., Krutilina,A.I., Li,L. and Axelrod,V.D. (1967) Primary structure of valine transfer RNA 1 of baker's yeast. *Mol. Biol. (Russ.)*, **1**, 754–766.
- Beck,E., Sommer,R., Auerswald,E.A., Kurz,Ch., Zink,B., Osterburg,G., Schaller,H., Sugimoto,K., Sugisaki,H., Okamoto,T. and Takanami,M. (1978) Nucleotide sequence of bacteriophage fd DNA. *Nucleic Acids Res.*, **5**, 4495–4503.
- Berget,S.M., Moore,C. and Sharp,P.A. (1977) Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. USA*, **74**, 3171–3175.
- Celma,M.L., Dhar,R., Pan,J. and Weissman,S.M. (1977) Comparison of the nucleotide sequence of the messenger RNA for the major structural protein of SV40 with the DNA sequence encoding the amino acids of the protein. *Nucleic Acids Res.*, **4**, 2549–2559.
- Chargaff,E., Vischer,E.M., Doniger,R., Green,C. and Misani,F. (1949) The composition of the desoxyribose nucleic acids of thymus and spleen. *J. Biol. Chem.*, **177**, 405–416.
- Chow,L.T., Gelin,R.E., Broker,T.R. and Roberts,R.J. (1977) An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*, **12**, 1–8.
- Conden,R., Gordon,A.H., Martin,A.J.P. and Synge,R.L.M. (1947) Gramicidine S: the sequence of the amino-acid residues. *Biochem. J.*, **41**, 596–602.
- Crick,F.H.C. (1966) The genetic code—yesterday, today and tomorrow. *Cold Spring Harb. Symp. Quant. Biol.*, **31**, 3–9.
- Crick,F.H.C., Griffith,J.S. and Orgel,L.E. (1957) Codes without commas. *Proc. Natl. Acad. Sci. USA*, **43**, 416–421.
- Curtius,T. (1883) Ueber das Glycocoll. *Chem. Ber.*, **16**, 753–757.
- Davidson,J.N. (1968) Nucleic acids—the first hundred years. *Prog. Nucl. Acid Res. Mol. Biol.*, **8**, 1–6.
- Dayhoff,M. (1969) *Atlas of Protein Sequence and Structure 1969*. vol. 4, National Biomedical Research Foundation, Silver Spring, Maryland.
- Doolittle,W.F. and Sapienza,C. (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature*, **284**, 601–603.
- Dunn,A.R. and Hassell,J.A. (1977) A novel method to map transcripts: evidence for homology between an adenovirus mRNA and discrete multiple regions of the viral genome. *Cell*, **12**, 23–36.
- Fiers,W., Contreras,R., Duerinck,F., Haegeman,G., Iserentant,D., Merregaert,J., Min Jou,W., Molemans,F., Raeymaekers,A., Van Den Berghe,A., Volckaert,G. and Ysebaert,M. (1976) Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, **260**, 500–507.
- Fiers,W., Contreras,R., Haegeman,G., Rogiers,R., Van de Voorde,A., Van Heuverswyn,H., Van Herreweghe,J., Volckaert,G. and Ysebaert,M. (1978) Complete nucleotide sequence of SV40 DNA. *Nature*, **273**, 113–120.
- Fischer,E. (1902) Ueber die Hydrolyse der Proteinstoffe. *Chemiker-Zeitung*, **26**, 939–940.
- Gamov,G. (1954) Possible relation between deoxyribonucleic acid and protein structure. *Nature*, **173**, 318.
- Gamov,G. (1955) On information transfer from nucleic acids to proteins. *Kgl. Danske Videnskab Selskab Biol. Medd.*, **22**, 3–7.
- Gatlin,L.L. (1972) *Information Theory and the Living System*. Columbia University Press, New York.
- Gilbert,W. (1978) Why genes in pieces? *Nature*, **271**, 501–501.
- Glover,D.M. and Hogness,D.S. (1977) A novel arrangement of the 18S and 28S sequences in a repeating unit of *Drosophila melanogaster* rDNA. *Cell*, **10**, 167–176.
- Godson,G.N., Barrell,B.G., Staden,R. and Fiddes,J.C. (1978) Nucleotide sequence of bacteriophage G4 DNA. *Nature*, **276**, 236–247.
- Goodman,H.M., Olson,M.V. and Hall,B.D. (1977) Nucleotide sequence of a mutant eukaryotic gene: the yeast tyrosine-inserting ochre suppressor SUP4-o. *Proc. Natl. Acad. Sci. USA*, **74**, 5453–5457.
- Grunberg-Manago,M., Ortiz,P.J. and Ochoa,S. (1956) Enzymic synthesis of polynucleotides. I. Polynucleotide phosphorylase of *Azotobacter vinelandii*. *Biochim. Biophys. Acta*, **20**, 269–285.
- Hess,E.L. (1970) Origins of molecular biology. *Science*, **168**, 664–669.
- Hofmeister,F. (1902) Über Bau und Gruppierung der Eiweisskörper. *Ergeb. Physiol.*, **1**, 759–802.
- Holley,R.W., Apgar,J., Everett,G.A., Madison,J.T., Marquisee,M., Merrill,S.H., Penswick,J.R. and Zamir,A. (1965) Structure of a ribonucleic acid. *Science*, **147**, 1462–1465.
- Holliday,R. (1968) Genetic recombination in fungi. In Peacock,W.J. and Brock,R.D. (eds), *Replication and Recombination of*

- Genetic Material* Australian Academy of Science, Canberra, pp. 157–174.
- Jacob, F. and Monod, J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, **3**, 318–356.
- Judson, H.F. (1979) *The Eighth Day of Creation: Makers of the Revolution in Biology*. Simon and Schuster, New York.
- Kelly, T.J. and Smith, H.O. (1970) A restriction enzyme from *Hemophilus influenzae*. II. Base sequence of the recognition site. *J. Mol. Biol.*, **51**, 393–409.
- Khesin, R.B., Shemiakin, M.F., Gorlenko, J.M., Bogdanova, S.L. and Afanasieva, T.P. (1962) RNA-polymerase in *E. coli* cells infected with T2 phage. *Biokhimiia (Russ.)*, **27**, 1092–1105.
- Khorana, H.G., Büchi, H., Ghosh, H., Gupta, N., Jacob, T.M., Kössel, H., Morgan, R., Narang, S.A., Ohtsuka, E. and Wells, R.D. (1966) Polynucleotide synthesis and the genetic code. *Cold Spring Harb. Symp. Quant. Biol.*, **31**, 39–49.
- Klessig, D.F. (1977) Two adenovirus mRNAs have a common 5' terminal leader sequence encoded at least 10 kb upstream from their main coding regions. *Cell*, **12**, 9–21.
- Levene, P.A. and Simms, H.S. (1925) The dissociation constants of plant nucleotides and nucleosides and their relation to nucleic acid structure. *J. Biol. Chem.*, **65**, 519–534.
- Madison, J.T., Everett, G.A. and Kung, H. (1966) Nucleotide sequence of a yeast tyrosine transfer RNA. *Science*, **153**, 531–534.
- Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA*, **74**, 560–564.
- Miescher, J.F. (1871) *Med.-Chem. Unters.* pp. 441.
- Needleman, S. and Wunsch, C. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
- Nirenberg, M.W. and Matthaei, J.H. (1961) The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. USA*, **47**, 1588–1602.
- Nirenberg, M., Caskey, T., Marshall, R., Brimacombe, R., Kellogg, D.D., Doctor, B., Hatfield, D., Levin, J., Rottman, F., Pestka, S., Wilcox, M. and Anderson, F. (1966) The RNA code and protein synthesis. *Cold Spring Harb. Symp. Quant. Biol.*, **31**, 11–24.
- Nolan, C. and Margoliash, E. (1968) Comparative aspects of primary structures of proteins. *Annu. Rev. Biochem.*, **37**, 727–790.
- Olby, R. (1974) *The Path to the Double Helix*. University of Washington Press, Seattle.
- Orgel, L.E. and Crick, F.H.C. (1980) Selfish DNA: the ultimate parasite. *Nature*, **284**, 604–607.
- Pellegrini, M., Manning, J. and Davidson, N. (1977) Sequence arrangement of the rRNA of *Drosophila melanogaster*. *Cell*, **10**, 213–224.
- Ratner, V.A. (1974) The genetic language. *Prog. Theor. Biol.*, **3**, 143–228.
- Reddy, V.B., Thimmappaya, B., Dhar, R., Subramanian, K.N., Zain, B.S., Pan, J., Ghosh, P.K., Celma, M.L. and Weissman, S.M. (1978) The genome of simian virus 40. *Science*, **200**, 494–502.
- Sanger, F. and Coulson, A.R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, **94**, 441–448.
- Sanger, F. and Tuppy, H. (1951) The amino-acid sequence in the phenylalanyl chain of insulin. *Biochem. J.*, **49**, 463–490.
- Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison III, C.A., Slocumbe, P.M. and Smith, M. (1977) Nucleotide sequence of bacteriophage  $\phi$ X174 DNA. *Nature*, **265**, 687–695.
- Schaap, T. (1971) Dual information in DNA and the evolution of the genetic code. *J. Theor. Biol.*, **32**, 293–298.
- Smith, E.L. (1979) Amino acid sequences of proteins—the beginnings. *Ann. NY Acad. Sci.*, **325**, 107–118.
- Speyer, J.F., Lengyel, P., Basilio, C., Wahba, A.J., Gardner, R.S. and Ochoa, S. (1963) Synthetic polynucleotides and the amino acid code. *Cold Spring Harb. Symp. Quant. Biol.*, **28**, 559–567.
- Steudel, H. (1906) Die Zusammensetzung der Nucleinsäuren aus Thymus und aus Heringsmilch. *Z. Physiol. Chem.*, **49**, 406–409.
- Wellauer, P.K. and Dawid, I.B. (1977) The structural organization of ribosomal DNA in *Drosophila melanogaster*. *Cell*, **10**, 193–212.
- Woese, C. (1967) The present status of the genetic code. *Prog. Nucleic Acid Res. Mol. Biol.*, **7**, 107–172.
- Wu, R. (1978) DNA sequence analysis. *Annu. Rev. Biochem.*, **47**, 607–634.
- Wu, R. and Kaiser, A.D. (1968) Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *J. Mol. Biol.*, **35**, 523–537.
- Wu, R. and Taylor, E. (1971) Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage  $\lambda$  DNA. *J. Mol. Biol.*, **57**, 491–511.
- Zachau, H.G., Dütting, D. and Feldman, H. (1966) Nucleotide sequences of two serine-specific transfer ribonucleic acids. *Angew. Ch.*, **78**, 392.
- Zuckerlandl, E. (1981) A general function of noncoding polynucleotide sequences. *Mol. Biol. Rep.*, **7**, 149–158.
- Zuckerlandl, E. and Pauling, L. (1962) Molecular disease, evolution, and genic heterogeneity. In Kasha, M. and Pullman, B. (eds), *Horizons in Biochemistry* Academic Press, New York, pp. 189–225.