BIOINFORMATICS

Vol. 16 no. 1 2000
Pages 10–15

A historical perspective on gene/protein functional assignment

T. C. Hodgman

GlaxoWellcome Medicines Research Centre, Stevenage SG1 2NY, UK

Abstract

Sequence determination and analysis began on proteins in the 1950s, with RNA starting about a decade later and DNA a similar period later still. Hence many of the concepts for function prediction were first developed by looking at amino acid sequences. Over time these methods have become much more sophisticated, allowing better discrimination of only weak similarities. The most recent developments concern an examination of contextual information, such as operon structure, metabolic reconstruction or co-expression profiles.

Contact: *ch18380@ggr.co.uk*

Introduction

The history of sequence management and analysis began on paper, but has always made use of (if not required) the latest developments in computer technology. This work will focus on the biological applications of these techniques because it is closer to my own expertise: other articles in this volume go into these more technical aspects. When citing early examples of applications or techniques, I am not implying that these are the first occasions, but rather the earliest that I have found or remembered.

Sequence determination

Fred Sanger has rightly been awarded for his truly outstanding work in discovering that biological macromolecules are chemically distinct entities having specific orders for the components residues, i.e. *sequences*, and in the development of sequencing techniques (Sanger, 1952). Analysis of proteins came first because they are more chemically and structurally stable, and easier to purify in large quantities. Several approaches were introduced, specifically, separating and purifying the different subunits of the protein, breaking up and then purifying the polypeptide subfragments generated by different proteases, using special chemistry to determine the residue order of the subfragments, finding the overlaps, and then assembling the 'contiguous' subsequences into the complete sequence.

Nucleotide sequences are usually longer and certainly less stable, making it difficult to produce specific subfragments small enough (10–20 bases were recommended in Galibert *et al.*, 1974) and in sufficient quantities. The next developments were to radiolabel the oligonucleotide so that less material was needed, make a nested set of fragments by incomplete digestion which were separated by appropriate 2D electrophoresis into a readable sequence directly off the autoradiograph (Sanger *et al.*, 1965; Szekely and Sanger, 1969; Ziff *et al.*, 1974). Read lengths of 20–30 bases were considered good. Finally, DNA sequencing became the norm in the late 1970s with the advent of *in vitro* DNA cloning, interrupted replication synthesis using a DNA polymerase (Sanger and Coulson, 1975), modified chain-terminating nucleotides (Sanger *et al*., 1977) and thin gels allowing much longer reads (Sanger and Coulson, 1978). The chemical sequencing method of DNA (Maxam and Gilbert, 1977) has not stood the test of time.

It was the analysis of protein and RNA sequence data that introduced most of the concepts that are now the standard tools for functional identification, though many awaited developments in mathematics and computer technology to realize their full potential.

Pairwise sequence alignment

Many aspects of biological research rely on comparison of one or more biological objects with a reference set of others; for example, species identification for ecology, palaeontology and evolutionary biology, or tissue status determination from histological or pathological specimens. Historically, people have carried such 'databases' in their heads or textbooks. Gene/protein function assignment uses similar principles; and nowadays where the pairwise percentage identity is high enough, the assignment is accepted rather than waiting for biochemical confirmation. Long before sequences of unknown function became widespread, the issues of how to make and judge biologically meaningful alignments were being addressed.

The most widely regarded early algorithm came from Needleman and Wunsch (1970). Sequence comparison and multiple alignment, particularly of proteins whose crystal structure was known, helped determine the parameters by which alignments should be judged. It also

quickly became apparent that percentage identity alone was too crude and that matrices defining the relatedness of each pair of amino acids should be considered. For many years the PAM matrices defined by Dayhoff's group were the parameters of choice (Dayhoff and Eck, 1967/8; Dayhoff, 1978). It was not until recent years that the re-evaluation leading to the BLOSUM matrices (Henikoff and Henikoff, 1992) was carried out and came into common use.

The early 1980s saw the beginning of the exponential growth of the sequence databases, forcing the necessity of storing and querying of the data on computers, which no longer consisted of many representatives of a limited set of protein families. The chances of finding something in the protein database which matched a novel sequence were slim, and clues for functional relatedness came from distantly related sequences. However, the most sensitive algorithms (such as Smith and Waterman, 1981) were unacceptably slow on the hardware of the day, so appropriate shortcuts were sought. The algorithm of Wilbur and Lipman (1983) gave way to the FAST series (Lipman and Pearson, 1985; Pearson and Lipman, 1988) and later the BLAST series (Altschul *et al.*, 1990, 1997); while the introduction of machines with parallel CPU architectures enabled special implementations of the more sensitive algorithms (Collins *et al.*, 1987).

Function prediction by sequence database searching finally acquired international acclaim in 1983, when the v-*sis* oncogene was shown to be related to platelet-derived growth factor (Doolittle *et al.*, 1983; Waterfield *et al.*, 1983). Although there may be earlier examples, this observation marked the beginning of assignment by similarity, and the first practical guides to function prediction appeared a few years later (Doolittle, 1986; Hodgman, 1986). During that decade, it was still uncommon to find strong (i.e. $>$ 30% identity) matches, and function prediction relied heavily upon multiple alignments and sequence signatures (for example, McGeoch and Davison, 1986). However, there has been a marked shift in this decade towards finding good sequence matches routinely, but potential difficulties arise from the match being of unknown function or its annotation has been misleading (for example, Richards *et al.*, 1995)

Simple pairwise alignment of DNA has never really been the method of choice for functional assignment of protein-coding genes because its alphabet is smaller, making the possibility of misalignment more likely, and information about synonymous codons cannot easily be taken into account. The general rule remains: translate the DNA then search the protein databases. This is made easier by algorithms like BLASTX which translate the DNA automatically. Introns may complicate these searches and users should look for the appropriate DNA sequence signatures, especially when good similarity at the protein level suddenly stops.

However, nucleotide database searching is still a valid exercise because some sequences fall outside protein coding regions, and some genes code for RNAs that have a structural function (rRNA, tRNA, rnp RNAs – Dreyfuss *et al.*, 1988), catalytic activity (ribozymes including RNaseP), or a hybrid function such as transfer-messenger RNA (Muto *et al.*, 1998). Some genome annotators have not included certain of the RNA gene classes, presumably because they were not looking for them. Furthermore, searching for strong matches in the database of expressed sequence tags (Boguski *et al.*, 1993) can be a very effective way of identifying transcribed regions, even though the function of that gene product is unknown.

Sequence signatures

The work done in the 1950s and 1960s led to the publication of the *Atlas of Protein Sequences and Structure* and its subsequent extra volumes (Dayhoff and Eck, 1967/8; Dayhoff, 1978). These books presented multiple sequence alignments (even of the available nucleotide sequences!) and crystal structure figures, and showed how patterns of conservation can be seen which arise from the structural or functional role played by the residue at each alignment position. Short diagnostic peptides were observed, which have been named *motifs*. Some motifs are regular expressions (a computing term), as exemplified by the N-glycosylation signal: [Asn]-[not Pro]-[Ser or Thr] (Marshall, 1972). Because the protein databases were expanding and computational power increasing, it was inevitable that a more concerted effort would be made to construct motif databases, and in 1989 a bibliographic 200-motif resource (Hodgman, 1989), a database of weight matrices (later absorbed into the Leeds University Protein Engineering Suite) and the now standard resource PROSITE (Bairoch, 1991) all appeared.

The techniques used to define these sequence signatures and matching them to query sequences has become increasingly sophisticated over time. Regular expressions (as in the early PROSITE releases) gave way to weight matrices (Staden, 1984), perceptrons (Stormo *et al.*, 1982), profiles (Gribskov *et al.*, 1987), neural networks (Qian and Sejnowski, 1988), hidden Markov models (Churchill, 1989), the use of Dirichlet mixtures (Brown *et al.*, 1993), Bayesian statistics (Liu *et al.*, 1995), computational linguistics (Searls, 1997) and sampling techniques (Lawrence *et al.*, 1993; Neuwald and Green, 1994). The historical development of these is much better described elsewhere (Baldi and Brunak, 1998).

Many have relied upon pre-existing alignments, though some do not and can even be used to generate alignments from a divergent protein set. Some of these algorithms

have resulted in 'next generation' databases, such as BLOCKS (Henikoff and Henikoff, 1994), PRINTS (Attwood and Beck, 1994) and PFAM (Sonnhammer *et al.*, 1997).

The inverse of the above approach can be used for families of unknown function. An alignment can be used to identify motifs which are then used to search the protein sequence databases. In this way, the helicase superfamily and the function of a protein domain in RNA viruses were identified simultaneously (Hodgman, 1988). Software tools were developed to assist this procedure, notably SCRUTINEER (Sibbald and Argos, 1990) and PROFILESEARCH (now an integral part of the GCG software). PSIBLAST now attempts to automate this process into a single step (Altschul *et al.*, 1997).

Signatures have been defined for nucleotide sequences. Those of gene length have been defined by computational grammars (Guigo *et al.*, 1992), and use of RNA secondary/tertiary structure (Eddy and Durbin, 1994; Gorodkin *et al.*, 1997). Signatures for functional subsequences have been identified for transcriptional control signals, exon boundaries, mRNA localization (Duret *et al.*, 1993; Steward and Singer, 1997), and matrix attachment sites (Singh *et al.*, 1997). However, the small alphabet generates a poor signal/noise ratio, which can only be reduced by taking position-specific effects or supplementary biological information into account such as distances between motifs (Staden, 1988). Many of these nucleotide motifs have also been collated into databases (notably by Prestridge, 1996; Heinemeyer *et al.*, 1998; Perier *et al.*, 1998).

Single sequence analyses

When the above techniques failed, there were still analyses of the sequence itself which could provide pointers to function. Most easily identified through dot-matrix plots, the presence of (imperfect) repeats (Gibbs and McIntyre, 1970; Staden, 1982), might indicate that the protein plays a structural role (such as keratin or myosin tails), involved in protein–protein interactions (for example the leucine zipper or tower helices of haemagglutinin) or ligand-binding domains of receptors (Doolittle, 1985; Sudhoff *et al.*, 1985). Regions of biased amino acid composition, depicted by plots of the sequence against amino acid abundance or some biophysical characteristic (for example, hydrophobicity or charge), have been useful for identifying functional segments such as transmembrane sequences, metal or nucleic acid binding sites, and regions prone to cause inherited disease (Ashley and Warren, 1995). Yet other techniques attempt to elucidate secondary structural elements as a guide to function.

With the exception of some theoretical work (Pauling, 1951), protein structure prediction began in the mid-1960s (Davies, 1964). Subsequently, information theory techniques (Garnier *et al.*, 1978) performed somewhat better than early statistical approaches (Chou and Fasman, 1974) which have since been tacitly falsified (Rooman and Wodak, 1988). Another early technique concerned drawing protein subsequences on helical wheels (Schiffer and Edmundson, 1967) and then looking for characteristic hydrophobic patches. This was later developed further into hydrophobic cluster analysis (Henrissat *et al.*, 1990), hydrophobic moment plots (Eisenberg, 1984), and the use of multiple alignments (Hodgman and Ellar, 1990). Significant steps forward in prediction accuracy came from the use of machine learning (King and Sternberg, 1990), neural networks (Rost *et al.*, 1993) and threading techniques (Bryant and Lawrence, 1993). However, these rely on the algorithms having 'background knowledge' of known structures which potentially makes them analogous to sequence signatures (more correctly structure signatures). On current trends, these techniques should become extremely useful when we obtain a more comprehensive set of domain folds.

The above categories of single sequence analysis all have parallels for nucleic acids. Codon usage plots are quite reliable for finding prokaryotic or viral protein coding regions because they are not broken up by large introns (Staden and McLachlan, 1982), and may reveal information about translational pausing (Krasheninnikov *et al.*, 1989, 1991). 'Islands' abundant in the dinucleotide CpG are also indicative of eukaryotic promoter regions and the 5′ end of genes (Bird, 1985). The search for (imperfect) direct or inverted repeats has aided the identification of regulatory elements and insertion sequence boundaries. There is also a long history of predicting RNA secondary and tertiary structure by energy minimization, which has been useful in characterizing translational regulatory elements (Pipas and McMahon, 1975; Westhof *et al.*, 1997). Finally, neural nets have been brought into use for exon identification (well reviewed by Schneider and Stormo, 1997) and hidden Markov models have been generated for miscellaneous applications (Baldi and Brunak, 1998; Durbin *et al.*, 1998).

Genome scale assignment

The 1990s have seen a massive drive to factory-scale automation of DNA sequencing, making Bioinformatics an indispensable part of molecular biology and genetics. Some dozens of genomes have now been completely sequenced which has provided statistically meaningful test sets for functional assignment. Some attempts have been made to automate the assignment process, notably GeneQuiz (Casari *et al.*, 1996) and EcoCyc (Karp *et al.*, 1998). A moderately consistent picture is emerging in which (a) 40–70% of genes can be assigned on the basis

of simple database searches; (b) assignment is an ongoing process as extra sequence and functional data become available, and domain experts clarify ambiguous observations; (c) evidence from operon structure can corroborate otherwise questionable database search results; (d) similar evidence can be provided by metabolic reconstruction, indeed, enzyme functions have been forecast whose gene has not yet been specified (Selkov, personal communication); and (e) 15–30% of genes cannot be assigned because they have no recognizable features or belong to a family of unknown function.

The current challenges

The current challenges in functional prediction concern simplifying the procedures for assignment by domain experts and from corroborative data, which now also including gene expression profiles, on the principle that what is expressed together functions together; and effectively capturing and propagating the experimental results for the genes of unknown function, which are being generated in new, systematic, large-scale functional genomics projects like EUROFAN (Oliver, 1996).

References

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Ashley,C.T. Jr. and Warren,S.T. (1995) Trinucleotide repeat expansion and human disease. *Annu. Rev. Genet.*, **29**, 703–728.

Attwood,T.K. and Beck,M.E. (1994) PRINTS – a protein motif fingerprint database. *Protein Eng.*, **7**, 841–848.

Bairoch,A. (1991) PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res.*, **19,** suppl, 2241–2245.

Baldi,P. and Brunak,S. (1998) *Bioinformatics. The Machine Learning Approach*. MIT Press, Cambridge, Massachussetts.

Bird,A.P. (1985) CpG-rich islands and the function of DNA methylation. *Nature*, **321**, 209–213.

Boguski,M.S., Lowe,T.M.J. and Tolstoshev,C.M. (1993) dbEST – database for 'Expressed Sequence Tags'. *Nat. Genet.*, **4**, 332–333.

Brown,M.P., Hughey,R., Krogh,A., Mian,I.S., Sjolander,K. and Haussler,D. (1993) Using Dirichlet mixture priors to derive hidden Markov models for protein families. In Hunter,L., Searls,D. and Shavlik,J. (eds), *Intelligent Systems in Molecular Biology 1993* AAAI/MIT Press, Menlo Park, California, pp. 47–55.

Bryant,S.H. and Lawrence,C.E. (1993) An empirical energy function for threading protein sequence through the folding motif. *Proteins*, **16**, 92–112.

Casari,G., De Daruvar,A., Sander,C. and Schneider,R. (1996) Bioinformatics and the discovery of gene function. *Trends Genet.*, **12**, 244–245.

Chou,P.Y. and Fasman,G.D. (1974) Prediction of protein conformation. *Biochemistry*, **13**, 222–245.

Churchill,G.A. (1989) Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.*, **51**, 79–94.

Collins,J.F., Coulson,A.F.W. and Lyall,A. (1987) *Comput. J.*, **30**, 420–424.

Davies,D.R. (1964) A correlation between amino acid composition and protein structure. *J. Mol. Biol.*, **9**, 605–609.

Dayhoff,M.O. (1978) *Atlas of Protein Sequences and Structure*. Vol. 5, suppl. 3., NBRF, Washington, DC.

Dayhoff,M.O. and Eck,R.V. (1967/8) *Atlas of Protein Sequences and Structure*. NBRF, Washington, DC.

Doolittle,R.F. (1985) The genealogy of some recently evolved vertebrate proteins. *Trends Biochem. Sci.*, **10**, 233–237.

Doolittle,R.F. (1986) *Of URFs and ORFs. A Primer on How to Analyze Derived Amino Acid Sequences*. University Science Books, Mill Valley, California.

Doolittle,R.F., Hunkapiller,M.W., Hood,L.E., Devare,S.G., Robbins,K.C., Aaronson,S.A. and Antoniades,H.N. (1983) Simian sarcoma *onc* gene, v-*sis*, is derived from the gene (or genes) encoding a platelet-derived growth factor. *Science*, **221**, 275–277.

Dreyfuss,G., Philipson,L. and Mattaj,I.W. (1988) Ribonucleoprotein particles in cellular processes. *J. Cell Biol.*, **106**, 1419–1425.

Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.

Duret,L., Dorkeld,F. and Gautier,C. (1993) Strong conservation of non-coding sequences during vertebrate evolution: potential involvement in post-translational regulation of gene expression. *Nucleic Acids Res.*, **21**, 2315–2322.

Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.

Eisenberg,D. (1984) Three-dimensional structure of membrane and surface proteins. *Annu. Rev. Biochem.*, **53**, 595–623.

Galibert,F., Sedat,J. and Ziff,E. (1974) Direct determination of DNA nucleotide sequences: structure of a fragment of bacteriophage phiX172 DNA. *J. Mol. Biol.*, **87**, 377–407.

Garnier,J., Osguthorpe,D.J. and Robson,B. (1978) Analysis of the accuracy and implications of simple methods for the predicting secondary structure of globular proteins. *J. Mol. Biol.*, **120**, 97–120.

Gibbs,A.J. and McIntyre,G.A. (1970) The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *Eur. J. Biochem.*, **16**, 1–11.

Gorodkin,J., Heyer,L.J. and Stormo,G.D. (1997) Finding common sequence and structure motifs in a set of RNA sequences. *Intell. Sys. Mol. Biol.*, **5**, 120–123..

Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA*, **84**, 4355–4358.

Guigo,R., Kundsen,S., Drake,N. and Smith,T. (1992) Prediction of gene structure. *J. Mol. Biol.*, **226**, 141–157.

Heinemeyer,T., Wingender,E., Reuter,I., Hermjakob,H., Kel,A.E., Kel,O.V., Ignatieva,E.V., Ananko,E.A., Podkolodnaya,O.A., Kolpakov,F.A., Podkolodny,N.L. and Kolchanov,N.A. (1998) Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res.*, **26**, 362–367.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution

matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, **89**, 10915–10919.

Henikoff,S. and Henikoff,J.G. Protein family classification based on searching a database of blacks. *Genomics*, **19**, 97–107.

Henrissat,B., Raimbaud,E., Tran,V. and Mornon,J.P. (1990) HCA-BAND: a computer program for the 2D-helical representation of protein sequences. *Comput. Applic. Biosci.*, **6**, 3–5.

Hodgman,T.C. (1986) The elucidation of protein function from its amino acid sequence. *Comput. Applic. Biosci.*, **2**, 181–187.

Hodgman,T.C. (1988) A superfamily of replicative proteins. *Nature*, **333**, 22,23,578.

Hodgman,T.C. (1989) The elucidation of protein function by sequence motif analysis. *Comput. Applic. Biosci.*, **5**, 1–13.

Hodgman,T.C. and Ellar,D.J. (1990) Models for the structure and function of the *Bacillus thuringiensis* d-endotoxins determined by compilational analysis. *DNA Seq.*, **1**, 97–103.

Karp,P.D., Riley,M., Paley,S.M., Pelligrini-Toole,A. and Krummenacker,M. (1998) EcoCyc: Encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.*, **26**, 50–53.

King,R.D. and Sternberg,M.J. Machine learning approach for the prediction of protein secondary structure prediction. *J. Mol. Biol.*, **216**, 441–457.

Krasheninnikov,I.A., Komar,A.A. and Adzhubei,I.A. (1989) Role of the code redundancy in determining cotranslational protein folding. *Biokhimiia*, **54**, 187–200.

Krasheninnikov,I.A., Komar,A.A. and Adzhubei,I.A. (1991) Nonuniform size distribution of nascent globin peptides, evidence for pause localisation sites, and a contranslational protein-folding model. *J. Protein Chem.*, **10**, 445–453.

Lawrence,C.E., Altschul,S.F., Boguski,M.S., Lui,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy. *Science*, **262**, 208–214.

Lipman,D.J. and Pearson,W.R. (1985) Rapid and sensitive protein similarity searches. *Science*, **227**, 1435–1441.

Liu,J.S., Neuwald,A.F. and Lawrence,C.E. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.*, **432**, 1156–1170.

McGeoch,D.J. and Davison,A.J. (1986) Alphaherpesviruses possess a gene homologous to the protein kinase gene family of eukaryotes and retroviruses. *Nucleic Acids Res.*, **14**, 1765–1777.

Marshall,R.D. (1972) Glycoproteins. *Annu. Rev. Biochem.*, **41**, 673–702.

Maxam,A.M. and Gilbert,W. (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA*, **74**, 560–564.

Muto,A., Ushida,C. and Himeno,H. (1998) A bacterial RNA that functions as both a tRNA and an mRNA. *Trends Biochem. Sci.*, **23**, 25–29.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Neuwald,A.F. and Green,P. (1994) Detecting patterns in protein sequences. *J. Mol. Biol.*, **239**, 698–712.

Oliver,S. (1996) A network approach to the systematic analysis of yeast gene function. *Trends Genet.*, **12**, 241–242.

Pauling,L. (1951) The structure of proteins: two hydrogen-bonded helical configurations of polypeptide chain. *Proc. Natl. Acad. Sci. USA*, **37**, 205–211.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.

Perier,R.C., Junier,T. and Bucher,P. (1998) The Eukaryotic Promoter Database EPD. *Nucleic Acids Res.*, **26**, 353–357.

Pipas,J.M. and McMahon,J.E. (1975) Method for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. USA*, **72**, 2017–2021.

Prestridge,D.S. (1996) SIGNAL SCAN 4.0: additional databases and sequence formats. *Comput. Applic. Biosci.*, **12**, 157–160.

Qian,N. and Sejnowski,T.J. (1988) Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, **202**, 865–884.

Richards,S.J., Hodgman,C. and Sharpe,M. (1995) Reported sequence homology between Alzheimer amyloid-770 and the MRC OX-2 antigen does not predict function. *Brain Res. Bull.*, **38**, 305–306.

Rooman,M.J. and Wodak,S.J. (1988) Identification of predictive sequence motifs limited by protein structure database size. *Nature*, **335**, 45–49.

Rost,B., Sander,C. and Schneider,R. (1993) PHD – an automatic mail server for protein secondary structure prediction. *Comput. Applic. Biosci.*, **10**, 53–60.

Sanger,F. (1952) The arrangement of amino acids in proteins. *Adv. Protein Chem.*, **7**, 1–67.

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 Coulson,A.R. (1978) The use of thin acrylamide gels for DNA sequencing. *FEBS Letts*, **87**, 107–110.

Sanger,F., Brownlee,G.G. and Barrell,B.G. (1965) A two-dimensional fractionation procedure for radioactive nucleotides. *J. Mol. Biol.*, **13**, 373–398.

Sanger,F., Nicklen,S. and Coulson,A.R. (1973) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.

Schiffer,M. and Edmundson,A.B. (1967) Use of helical wheels to represent the structures of protein and to identify segments with helical potential. *Biophys. J.*, **7**, 121–135.

Schneider,E.E. and Stormo,G.D. (1997) Identifying genes in genomic DNA sequences. In Bishop,M.J. and Rawlings,C.J. (eds), *DNA and Protein Sequence Analysis – a Practical Approach* IRL Press, Oxford, pp. 209–224.

Searls,D.B. (1997) Linguistic approaches to biological sequences. *Comput. Applic. Biosci.*, **13**, 333–344.

Sibbald,P.R. and Argos,P. (1990) Scrutineer: a computer program that flexibly seeks and describes motifs and profiles in protein sequence databases. *Comput. Applic. Biosci.*, **6**, 279–288.

Singh,G.B., Kramer,J.A. and Krawetz,S.A. (1997) Mathematical model to predict regions of chromatin attachment to the nuclear matrix. *J. Mol. Biol.*, **25**, 1419–1425.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Sonnhammer,E.L.L., Eddy,S.R. and Durbin,R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Protein Struct. Funct. Genet.*, **28**, 405–420.

Staden,R. (1982) An interactive graphics program for comparing and aligning nucleic acid and protein sequences. *Nucleic Acids Res.*, **10**, 2951–2961.

Staden,R. (1984) Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.*, **12**, 505–519.

Staden,R. (1988) Methods to define and locate patterns of motifs in sequences. *Comput. Applic. Biosci.*, **4**, 53–60.

Staden,R. and McLachlan,A.D. (1982) Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Res.*, **10**, 141–156.

Steward,O. and Singer,R.H. (1997) The intracellular RNA sorting system: postal zones, zip codes, mail bags and mail boxes. In Harford,J.B. and Morris,D.R. (eds), *mRNA Metabolism and Post-transcriptional Gene Regulation* Wiley-Liss, New York, pp. 127–146.

Stormo,G.D., Schneider,T.D., Gold,L. and Ehrenfeucht,A. (1982) Use of the "perceptron" algorithm to distinguish translational initiation sites in. *E. coli. Nucleic Acids Res.*, **10**, 2997–3011.

Sudhoff,T.C., Goldstein,J.L., Brown,M.S. and Russell,D.W. (1985) The LDL receptor gene: a mosaic of exons sharedwith different proteins. *Science*, **228**, 815–822.

Szekely,M. and Sanger,F. (1969) Use of polynucleotide kinase in fingerprinting non-radioactive nucleic acids. *J. Mol. Biol.*, **43**, 607–617.

Waterfield,M.D., Scrace,G.T., Whittle,N., Stroobant,P., Johnsson,A., Wasteson,Å., Westermark,B., Heldin,C.-H., Huang,J.S. and Deuel,T.F. (1983) Platelet-derived growth factor is structurally related to the putative transforming protein p28sis of simian sarcoma virus. *Nature*, **304**, 35–39.

Westhof,E., Auffinger,P. and Gaspin,C. (1997) DNA and RNA structure prediction. In Bishop,M.J. and Rawlings,C.J. (eds), *DNA and Protein Sequence Analysis – a Practical Approach* IRL Press, Oxford, pp. 255–278.

Wilbur,W.J. and Lipman,D.J. (1983) Rapid similarity searches in nucleic acid and protein databanks. *Proc. Natl. Acad. Sci. USA*, **80**, 726–730.

Ziff,E.B., Sedat,J.W. and Galibert,F. (1974) Determination of the nucleotide sequence of a fragment of bacteriophage phiX174 DNA. *Nat. New Biol.*, **241**, 34–37.