BIOINFORMATICS REVIEW

Vol. 19 no. 17 2003, pages 2176–2190
DOI: 10.1093/bioinformatics/btg309

Early bioinformatics: the birth of a discipline—a personal view

Christos A. Ouzounis[1,*] and Alfonso Valencia[2]

[1]Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK, [2]Protein Design Group, National Center for Biotechnology, CNB-CSIC Campus U. Autonoma Cantoblanco, Madrid 28049, Spain

Received on December 13, 2002; revised on May 25, 2003; accepted on March 28, 2003

ABSTRACT

Motivation: The field of bioinformatics has experienced an explosive growth in the last decade, yet this 'new' field has a long history. Some historical perspectives have been previously provided by the founders of this field. Here, we take the opportunity to review the early stages and follow developments of this discipline from a personal perspective.

Results: We review the early days of algorithmic questions and answers in biology, the theoretical foundations of bioinformatics, the development of algorithms and database resources and finally provide a realistic picture of what the field looked like from a resources and finally provide a realistic picture of what the field looked like from a practitioner's viewpoint 10 years ago, with a perspective for future developments.

Contact: ouzounis@ebi.ac.uk

PRELUDE

The recent revolution in genomics and bioinformatics has caught the world by storm. From company boardrooms to political summits, the issues surrounding the human genome, including the analysis of genetic variation, access to genetic information and the privacy of the individual have fueled public debate and extended way beyond the scientific and technical literature. During the past few years, bioinformatics, defined as the computational handling and processing of genetic information, has become one of the most highly visible fields of modern science. Yet, this 'new' field has a long, even humble, history, along with the triumphs of molecular genetics and cell biology of the last century.

Taking a historical perspective, we will examine the birth of this discipline, and some of the factors that shaped it into one of the hottest areas of frantic scientific research and technical development. First, we will attempt to describe briefly some key developments for computational biology, from the very early days to the close of the century. Second, we will compare some 'early' bioinformatics activities of just ten years ago with today's field, hoping that we provide a perspective for the future. Clearly, our account is a personal perspective and by no means an objective treatise on the history of bioinformatics. Yet, we hope that this will provide a basis for further discussion and debate, enriched by personal interviews, a detailed citation analysis and a more wide coverage of the different areas within a field. For instance, we have not covered sufficiently entire areas of biological computation, such as structural bioinformatics (X-ray crystallography, electron microscopy and nuclear magnetic resonance), modelling and dynamics, including image and signal analysis (regulatory and gene networks, physiological simulations, metabolic control theory, tissue visualization via tomography and nuclear magnetic imaging) or neurobiology and neuroinformatics (neural networks, control theory). These fields are outside the scope of our review and at the borders of biological computing with other important areas of research. We would like to make clear that we focus on our own area of expertise and discuss the milestones of the field of protein sequence and structure analysis while attempting to provide a general overview of the major achievements in bioinformatics. We list a number of institutions and key papers (Tables 1 and 2) that were influential in our own intellectual development and thus should not be considered as an objectively derived 'hall of fame' in this field. We hope that this treatise will inspire other scientists to take an opportunity and provide their own perspectives for the history of computational biology.

THE PRE-70'S: PIONEERING COMPUTATIONAL STUDIES

It could be argued that some of the most fundamental problems in the early days of molecular biology presented some formidable algorithmic problems. In that sense, the structure of DNA (Watson and Crick, 1953), the encoding of genetic information for proteins (Gamow *et al.*, 1956), the

*To whom correspondence should be addressed.

Table 1. Ten institutions that pioneered and fostered computation in biology

Institutions	Country
Birkbeck College, University of London	UK
Boston University	USA
European Molecular Biology Laboratory (EMBL)	DE and EMBL states
Institute of Protein Research, Academy of Sciences, Puschino	Former USSR
Laboratory of Molecular Biology (LMB), MRC Cambridge	UK
Los Alamos National Laboratory (LANL)	USA
National Biomedical Research Foundation (NBRF), Georgetown U	USA
Stanford University	USA
University of California San Francisco (UCSF)	USA
University College, University of London (UCL)	UK

factors governing protein structure (Anfinsen, 1973; Pauling *et al*., 1951), the structural properties of protein molecules (Anfinsen and Scheraga, 1975; Crick, 1953; Pauling and Corey, 1953; Szent-Györgyi and Cohen, 1957), the evolution of biochemical pathways (Horowitz, 1945) and gene regulation (Britten and Davidson, 1969), and the chemical basis for development (Turing, 1952) all contain seeds of some of the problems that were possible to address by computation in the following decades. In parallel, much of fundamental computer science, including the theory of computation (Chaitin, 1966) and information theory (Shannon and Weaver, 1962), the definition of grammars (Chomsky, 1959) and random strings (Martin-Löf, 1966), the theory of games (Neumann and Morgenstern, 1953) and cellular automata (Neumann, 1966) emerged during the 1950s and 1960s.

These early approaches had already been combining computational and experimental information to better understand biological macromolecules, and insights were gained on the evolution of genes and proteins (Ingram, 1961; Margoliash, 1963; Zuckerkandl and Pauling, 1965b), the issues of molecular homology (Florkin, 1962; Zuckerkandl and Pauling, 1965a), the analysis of molecules to unveil evolutionary patterns (Zuckerkandl and Pauling, 1965b), the structural constraints of polypeptide chains (Ramachandran *et al*., 1963), the informational properties of DNA (Gatlin, 1966) and protein sequences (Nolan and Margoliash, 1968), the origins of the genetic code (Crick, 1968; Woese, 1970), its coding capacity (Alff-Steinberger, 1969) and the accuracy of the translation process (Crick, 1966), the construction of phylogenetic trees (Fitch and Margoliash, 1967), the use of molecular graphics (Katz and Levinthal, 1966), properties of protein sequence alignment (Cantor, 1968) and the processes of molecular evolution (Kimura, 1968; Nei, 1969).

This era can be considered as the birth of computational biology, with a number of key developments appearing: the first sequence alignment algorithms (Gibbs and McIntyre, 1970; Needleman and Wunsch, 1970), models for selection-free molecular evolution (King and Jukes, 1969), the preferential substitution of amino acid residues in protein sequences (Clarke, 1970; Epstein, 1967), formal studies of protein primary structure (Krzywicki and Slonimski, 1967), derivation of preferences for amino acid residues in secondary structures (Pain and Robson, 1970; Ptitsyn, 1969), the invention of the helical wheel representation for protein sequences (Dunnill, 1968; Schiffer and Edmundson, 1967), the widespread use of molecular data in evolutionary studies (Fitch and Margoliash, 1970; Jukes, 1969), the origins of life (West and Ponnamperuma, 1970) and the theory of evolution by gene duplication (Ohno, 1970). In 1970, the central dogma had also been conceived (Crick, 1970), after the seminal discoveries of the processes of RNA transcription and translation.

THE 70'S: THE THEORETICAL FOUNDATIONS

As a consequence of the above, an agenda for computational problems in molecular biology had already been formulated. Studies of substitution mutation rates (Koch, 1971), the calculation of solvent accessibility on protein structures (Lee and Richards, 1971), the parsimonial determination of tree topology (Fitch, 1971), RNA structure prediction (Tinoco *et al*., 1971) and more methods for sequence alignment (Beyer *et al*., 1974; Gibbs *et al*., 1971; Grantham, 1974; Sackin, 1971; Sellers, 1974a; Wagner and Fischer, 1974) have appeared. One of the most prominent theoretical advancements of this time was the merging of classical population genetics with molecular evolution (Kimura, 1969; Ohta and Kimura, 1971), to produce the theory of neutral evolution (Kimura, 1983) and the constancy of the evolutionary rate of proteins (Jukes and Holmquist, 1972), also known as the molecular clock hypothesis (Kimura and Ohta, 1974). Another area of intensifying research was the string comparison problem in computer science (Levin, 1973; Sankoff and Sellers, 1973; Wagner and Fischer, 1974) (or 'sequence alignment' in biology), developed hand-in-hand with applications to biological macromolecules (Beyer *et al*., 1974; Gordon, 1973; Kimura and Ohta, 1972; Sankoff, 1972; Sankoff and Cedergren, 1973; Sellers, 1974b). At the same time, the first phylogenetic analyses of macromolecular families (Wu *et al*., 1974), including immunoglobulins (Novotny, 1973) and transfer RNA (Holmquist *et al*., 1973), were emerging. Moreover, refined attempts to define sequence patterns that influence protein structure continued to propagate (Kabat and Wu, 1973; Liljas and Rossman, 1974; Richards, 1974; Robson, 1974; Schulz *et al*., 1974; Wetlaufer, 1973).

By the mid-1970s, a pretty clear picture has been devised for the theory and practice of sequence alignment, the process of molecular evolution, the quantification of nucleotide and

Table 2. Twenty Publications that influenced our view of bioinformatics

Publication	Comments
Zuckerkandl and Pauling, 1965b	First use of molecular sequences for evolutionary studies
Fitch and Margoliash, 1967	Use of molecular sequences to build trees
Needleman and Wunsch, 1970	First implementation of dynamic programming for protein sequence comparison
Lee and Richards, 1971	Calculation of accessibility on protein structures
Chou and Fasman, 1974	First secondary structure prediction method
Tanaka and Scheraga, 1975	Simulation of protein folding
Dayhoff, 1978	First collection of protein sequences
Hagler and Honig, 1978	One of the first explicit attempts to simulate protein folding
Doolittle, 1981	Seminal paper examining divergence and convergence in protein evolution
Felsenstein, 1981	One of the first statistical treatments of evolutionary tree construction
Richardson, 1981a	The most comprehensive description of protein structure to that date
Kabsch and Sander, 1984	Discovery with profound implications for model building by homology and structure prediction
Novotny *et al*., 1984	The inability of distinguishing correct from incorrect structures threw back structure prediction approaches for a long while
Chothia and Lesk, 1986	Examination of divergence between sequence and structure
Doolittle, 1986	Influential book on sequence analysis
Feng and Doolittle, 1987	The first approach for an efficient multiple sequence alignment procedure, later implemented in CLUSTAL
Lathrop *et al*., 1987	One of the first applications of Artificial Intelligence in protein structure analysis and prediction
Ponder and Richards, 1987	The very first threading approach, using sequence enumeration
Altschul *et al*., 1990	The implementation of a sequence matching algorithm based on Karlin's statistical work
Bowie *et al*., 1991	The first implementation of protein structure prediction using threading

aminoacid substitution rates, the construction of evolutionary trees, and secondary/tertiary protein structure analysis. In certain ways, a lot of the problems that would occupy the computational biologists of the future had been defined during those early years. What was missing is central reference data and software resources and the means to access them, a significant trend that would emerge very prominently during the next decade.

In the last years of that decade, a flurry of activity occurred in the development of string and sequence alignment theory (Aho *et al*., 1976; Chvátal and Sankoff, 1975; Delcoigne and Hansen, 1975; Hirschberg, 1975; Lowrance and Wagner, 1975; Okuda *et al*., 1976; Waterman *et al*., 1976) and evolutionary tree analysis and construction (Felsenstein, 1978; Klotz *et al*., 1979; Sattath and Tvertsky, 1977; Waterman and Smith, 1978a; Waterman *et al*., 1977), as well as the description, visualization, analysis and prediction of protein structure, in an attempt to crack the 'second genetic code', the protein folding problem (Chothia, 1975; Chothia *et al*., 1977; Chou and Fasman, 1978; Crippen, 1978; Garnier *et al*., 1978; Hagler and Honig, 1978; Jones, 1978; Kabsch, 1976; Karplus and Weaver, 1976; Kuntz, 1975; Levitt, 1976, 1978; Levitt and Chothia, 1976; Levitt and Warshel, 1975; Lifson and Sander, 1979; Matthews, 1975; Nagano and Hasegawa, 1975; Richards, 1977; Richardson, 1977; Rose, 1979; Rossmann and Argos, 1976; Schulz, 1977; Schulz and Schirmer, 1979; Sternberg and Thornton, 1978; Tanaka and Scheraga, 1975; Ycas *et al*., 1978), including the first algorithms for secondary structure prediction (Chou and Fasman, 1974; Lim, 1974), the invention of distance geometry for the calculation of structure from distance constraints (Crippen, 1977) and further use of specialized systems for molecular graphics and modelling (Feldmann, 1976). An interesting by-product in this area were the evolutionary 'stories' for specific protein families, such as the selection-dependent evolution of haemoglobins (Goodman *et al*., 1975), the dehydrogenases and kinases (Eventoff and Rossman, 1975), cytochrome *c* (Fitch, 1976) and the first analyses of metabolism, such as the loss of metabolic capacities (Jukes and King, 1975), the evolution of catalytic efficiency (Albery and Knowles, 1976), the evolution of energy metabolism (Dickerson *et al*., 1976) and the simulation of metabolic regulation (Heinrich and Rapoport, 1977). Other emerging problems were the exon–intron question (Gilbert, 1978), the evolution of the bacterial genome (Riley and Anilionis, 1978), RNA structure prediction (Waterman and Smith, 1978b), deep phylogeny (Schwartz and Dayhoff, 1978) and the complex control of morphogenesis (Savageau, 1979a,b).

One key development towards the end of that decade regarding public resources was the compilation of computer archives for the storage, curation and distribution of protein sequence (Dayhoff, 1978) and structure (Bernstein *et al*., 1977) information, a trend that would be amplified enormously in the immediate future.

THE 80'S: MORE ALGORITHMS AND RESOURCES

The following decade was in effect the time when the field of computational biology took shape as an independent discipline, with its own problems and achievements. For the first time, efficient algorithms were developed to cope with an increasing volume of information, and their computer implementations were made available for the wider scientific community. Some commercial activity around software development has already been observed (Devereux *et al*., 1984). Due to the vast volume of literature, we will only cite a limited number of significant papers that represent key developments in computational biology. We will also break down the field into four subfields: (i) sequence analysis, (ii) molecular databases, (iii) protein structure prediction and (iv) molecular evolution.

By 1980, it had already become clear that computer analysis of nucleotide sequences was essential for the better understanding of biology (Gingeras and Roberts, 1980). Sequence comparison continued to benefit from parallel developments in computer science (Hall and Dowling, 1980). The dot-matrix model of sequence comparison was well developed at that time (Maizel and Lenk, 1981). The genome hypothesis for preferential codon usage was formulated on the basis of computer analysis (Grantham *et al*., 1980). Progress in DNA (Trifonov and Sussman, 1980) and RNA (Nussinov and Jacobson, 1980) structure analysis prediction was also reported. Other theoretical work at the turn of that decade included key analyses of the evolution of prokaryotes with the identification of the Archaea as a separate domain of life (Fox *et al*., 1980), the notion of selfish DNA (Doolittle and Sapienza, 1980) and variable modes of molecular evolution (Dover and Doolittle, 1980). Other fields with influence on computational biology were neural networks (Hopfield, 1982), molecular computing (Conrad, 1985), nanotechnology (Drexler, 1981), complexity and cellular automata (Burks and Farmer, 1984; Reggia *et al*., 1993; Wolfram, 1984) and the theory of clustering (Shepard, 1980), all of which had a direct impact on protein structure prediction and design as well as sequence database searching and clustering.

(i) Theoretical developments in sequence analysis, for example the computation of evolutionary distances (Sellers, 1980) or approximate string matching (Ukkonen, 1985), were followed by the development of key algorithms, such as the Smith–Waterman dynamic programming sequence alignment algorithm (Smith and Waterman, 1981a,b) and the FASTA family of algorithms for database searching (Lipman and Pearson, 1985; Wilbur and Lipman, 1983). Similarly, analysis of repeats in theoretical computer science (Guibas and Odlyzko, 1980; Steele, 1982) was followed by parallel analyses for biological sequences (DeWachter, 1981; Martinez, 1983; Nussinov, 1983). Matrix-based models of sequence comparison continued to be developed (Fristensky, 1986; Novotny, 1982), as well as the first integrated sequence analysis systems (Brutlag *et al*., 1982; Lyall *et al*., 1984; Pustell and Kafatos, 1984; Staden, 1982). Two major developments were the automation and wide use of multiple sequence alignment (Carrillo and Lipman, 1988; Feng *et al*., 1985; Hogeweg and Hesper, 1984; Murata *et al*., 1985; Sankoff and Cedergren, 1983), especially the tree-based alignment method (Feng and Doolittle, 1987; Higgins and Sharp, 1988), and sequence profile analysis (Gribskov *et al*., 1987, 1988). One of the first applications of sequence analysis to the discovery of important protein motifs was the identification of the ATP-binding motif in various functionally unrelated proteins (Walker *et al*., 1982), the zinc-finger motif (Klug and Rhodes, 1987), the leucine-zipper motif (Landschulz *et al*., 1988), the homology of bacterial sigma factors (Gribskov and Burgess, 1986) and the nature of signal sequences (Heijne, 1981, 1985). Other studies included optimality in sequence alignment (Altschul and Erickson, 1986; Fickett, 1984; Fitch and Smith, 1983; Waterman, 1983), rigorous statistical approaches in sequence analysis (Arratia *et al*., 1986; Arratia and Waterman, 1985a,b; Karlin *et al*., 1983; Tavaré, 1986; Wilbur and Lipman, 1984), pattern recognition in several sequences and consensus generation (Abarbanel *et al*., 1984; Sellers, 1984; Waterman *et al*., 1984) random sequences (Fitch, 1983), sequence logos (Schneider *et al*., 1986), and syntactic analysis (Ebeling and Jiménez-Montaño, 1980; Jiménez-Montaño, 1984). One issue was the performance of these computation-intensive programs on small computer systems (Gotoh, 1987; Korn and Queen, 1984). Algorithms for the prediction of antigenic determinants (Hopp and Woods, 1981), the detection of open reading frames (Fickett, 1982; Shepherd, 1981; Staden and McLachlan, 1982) and translation initiation sites (Stormo *et al*., 1982), the computation of RNA folding (Dumas and Ninio, 1982; Turner *et al*., 1988) and the calculation of evolutionary trees (Felsenstein, 1982) were also invented. The first reviews (Goad, 1986; Hodgman, 1986; Jungck and Friedman, 1984; Kruskal, 1983; Kruskal and Sankoff, 1983) and books (Doolittle, 1986; Heijne, 1987; Rawlings, 1986) on sequence analysis and comparison also appeared at this time.

(ii) The initial phase of database development for data quality control and collection rapidly progressed (Kelly and Meyer, 1983; Orcutt *et al*., 1983), with the appearance of at least two major resources for nucleotide data submission (Philipson, 1988), GenBank (Bilofsky *et al*., 1986) and the EMBL Data Library (Hamm and Cameron, 1986). Proposals for computer networks that ensured availability and facilitated distribution (Lesk, 1985; Lewin, 1984) were materialized, with initiatives such as EMBNET (Lesk, 1988) and BIONET (Kristofferson, 1987; Smith *et al*., 1986). Archives of molecular biology software also appeared, for example the LiMB software catalog (Burks *et al*., 1988; Lawton *et al*., 1989). Various reviews summarizing strategies for sequence database searching were published (Cannon, 1987; Davison, 1985; Henikoff and Wallace, 1988; Lawrence *et al*., 1986; Orcutt and

Barker, 1984; Thornton and Gardner, 1989), indicating that distributed computing for the wider community was coming of age (Heijne, 1988). Entire programs in various institutes such as EMBL formed the very first departments exclusively devoted to computational biology (Lesk, 1987). Finally, experimentation with various dedicated hardware platforms for more efficient analysis of biological sequences emerged (Collins and Coulson, 1984; Core *et al*., 1989; Edmiston *et al*., 1988; Gotoh and Tagashira, 1986; Huang, 1989; Lopresti, 1987) along with relational database technology that facilitated querying (Islam and Sternberg, 1989; Rawlings, 1988), as databases continued to grow at an exponential rate (DeLisi, 1988).

(iii) The field of protein structure analysis and prediction experienced a significant growth in that decade. Various approaches to protein structure representation and visualization were explored, including the derivation of coordinates from stereo diagrams (Rossmann and Argos, 1980), domain definitions (Rashin, 1981), hydrophobicity plots (Kyte and Doolittle, 1982; Sweet and Eisenberg, 1983) and moments (Eisenberg *et al*., 1984), automatic structure drawing (Lesk and Hardman, 1982), fractal surfaces (Brooks and Karplus, 1983), signed distance maps (Braun, 1983), solvent accessible surfaces (Connolly, 1983), vector representations of protein sequences (Swanson, 1984) and structures (Yamamoto and Yoshikura, 1986), substructure dictionaries (Jones and Thirup, 1986), amino acid conservation patterns (Taylor, 1986), differential geometry (Rackovsky and Goldstein, 1988) sequence motifs (Rooman and Wodak, 1988) and building blocks (Unger *et al*., 1989). Interactive computer graphics were introduced as well, with programs such as FRODO (Jones, 1985) and RIBBON (Priestle, 1988). Structure comparison was further developed, with new analyses and algorithms (Cohen and Sternberg, 1980a; McLachlan, 1982; Sippl, 1980; Taylor and Orengo, 1989). Class prediction as a filtering step in protein structure prediction was also invented at that time (Klein, 1986; Klein and DeLisi, 1986; Nishikawa *et al*., 1983a,b). Molecular modelling was developed (Greer, 1981), further validated with dictionaries of peptides (Kabsch and Sander, 1984) [and ultimately fully automated (Holm and Sander, 1992; Levitt, 1992) in the 1990s]. The problem of threading sequences to structures was also introduced (Ponder and Richards, 1987). Descriptive studies deriving architectural principles of protein structure (Chothia, 1984; Richardson, 1981b) from statistical analysis of specific families and folds continued to increase in quantity and sophistication (Brändén, 1980; Janin and Chothia, 1980; Lifson and Sander, 1980; Ptitsyn and Finkelstein, 1980; Weber and Salemme, 1980)—examples include analyses of disulfide bridges (Thornton, 1981), beta-sheet sandwiches (Cohen *et al*., 1981), helix packing patterns (Chothia *et al*., 1981) and beta-sheets (Chothia and Janin, 1981), beta-hairpins (Sibanda and Thornton, 1985), beta-barrels (Lasters *et al*., 1988), loops (Leszczynski and Rose, 1986) and coiled-coils (Cohen and Parry, 1986). The recent discovery of exons led to their mapping on known protein structures (Craik *et al*., 1982, 1983; Gô, 1981, 1983, 1985). The development of NMR allowed the solution of protein structures (Wüthrich, 1989), and presented new problems (Braun, 1987), the calculation of 3D coordinates from distance data: distance geometry (Gower, 1982, 1985) and molecular dynamics (Brünger *et al*., 1986) came to the rescue. These methods were previously used to approach the protein folding problem as prediction methods, with the use of distance constraints (Cariani and Goel, 1985; Cohen and Sternberg, 1980b; Galaktionov and Rodionov, 1981; Goel *et al*., 1982; Goel and Ycas, 1979; Kuntz *et al*., 1976; Wako and Scheraga, 1981, 1982) and the prediction of residue contacts (Miyazawa and Jernigan, 1985; Warme and Morgan, 1978) as well as restrained energy minimization and molecular dynamics (Levitt, 1983). Development of distance geometry continued (Braun, 1987; Braun and Gô, 1985; Crippen, 1987; Easthope and Havel, 1989; Hadwiger and Fox, 1989; Havel *et al*., 1983a,b; Havel and Wüthrich, 1984; Metzler *et al*., 1989; Sippl and Scheraga, 1985).

(iv) Protein evolution had also become a key area of research (Bajaj and Blundell, 1984; Dayhoff *et al*., 1983; Doolittle, 1981), with a number of interesting discoveries such as the coordinated changes of key residues (Altschuh *et al*., 1988), the relationship between the divergence of sequence and structure (Chothia and Lesk, 1986), the properties of similarity matrices (Wilbur, 1985), the influence of amino acid composition (Graur, 1985), the definition of homology (Reeck *et al*., 1987), the detection of protein fold determinants (Bashford *et al*., 1987) and the identification of sequence similarities due to convergence (Doolittle, 1988; Fitch, 1988). Key analyses of individual protein families with wider implications for protein sequence/structure relationships included the analysis of the globins (Lesk and Chothia, 1980), the blue-copper proteins (Chothia and Lesk, 1982), the immunoglobulins (Lesk and Chothia, 1982), the proteases (Neurath, 1984), the cytochromes (Mathews, 1985), the bacterial ferredoxins (George *et al*., 1985), the superoxide dismutases (Getzoff *et al*., 1989; Lee *et al*., 1985), the phosphorylases (Hwang and Fletterick, 1986), the ribonucleases (Beintema *et al*., 1988), the crystallins (Lubsen *et al*., 1988; Piatigorsky and Wistow, 1989) and other various case studies (Brenner, 1988; Doolittle, 1985; Goldfarb, 1988). Correspondingly, the analysis of phylogenetic markers such as rRNA (Rothschild *et al*., 1986; Sogin *et al*., 1986), exons and introns (Gilbert, 1985) and various genome segments (Brutlag, 1980) resulted in significant discoveries for genome evolution, such as the relationships of life forms (Cedergren *et al*., 1988; Iwabe *et al*., 1989; Pace *et al*., 1986; Woese, 1987), the dynamics of DNA (Breslauer *et al*., 1986) and genome structure (Blake and Earley, 1986; Loomis and Gilpin, 1986; Ohta, 1987; Reanney, 1986; Sankoff and Goldstein, 1989), the evolution of splicing (Sharp, 1985), exons (Bulmer, 1987; Naora and Deacon, 1982), introns (Gilbert et al., 1986; Senapathy,

1986), intron-encoded proteins (Perlman and Butow, 1989) and non-coding sequences (Naora *et al*., 1987), the origins of retroviruses (Doolittle *et al*., 1986), the salient features of substitution rates (Britten, 1986; Ochman and Wilson, 1987) and the effect of codon usage on gene expression (Grantham *et al*., 1981). Finally, the theory and practice of evolutionary tree computation came into maturity (Felsenstein, 1981, 1985, 1988b), culminated by the widely used program PHYLIP (Felsenstein, 1988a).

TEN YEARS AGO, WITH HINDSIGHT

Here is a pretty realistic picture of a computational biologist working back in 1992. In terms of generic computing tools, there had been access to the InterNet, mostly through services like (bitnet) e-mail, gopher/ftp and the first web browser, Mosaic (http protocol), allowing access to a little more than 100 or so(!) web sites. Computer systems were quite heterogeneous, including VAX/VMS machines and Unix workstations (and another dozen of less widely known operating systems). In addition, in academic environments Apple Macintosh systems were abundant, thanks to their groundbreaking icon-based user interface and word-processing or desktop publishing capabilities. There has been distributed databases, such as GenBank and MedLine, but their availability was limited, mostly through CD-ROMs. CD drives were just being made available and the first version of X-windows was launched (graphical user interfaces were still in their infancy). About that time the first interpreted languages appeared, inspired by the Unix utility awk and quickly followed by perl and python.

In terms of scientific toolkits, BLAST was just made available (Altschul *et al*., 1990), including sequence masking procedures, such as XNU (Claverie and States, 1993). RasMol (Sayle and Milner-White, 1995) and Kinemage (Richardson and Richardson, 1992) were making headlines in terms of protein structure visualization. The Genetics Computer Group (GCG) software was available on VMS and in wide use—along with many other popular sequence analysis packages for the Macintosh. The first sophisticated gene prediction programs were also appearing (Brunak *et al*., 1990; Fickett and Tung, 1992; Guigo *et al*., 1992; Mural and Uberbacher, 1991; States and Botstein, 1991). In protein structure prediction, the second-generation secondary structure prediction algorithms based on multiple sequence alignment (Rost and Sander, 1993), by then also widely available, indicated significant progress in the field. Excitement was in the air (Thornton *et al*., 1992) because of the first successful results in protein docking (Walls and Sternberg, 1992) and protein sequence threading (Bowie *et al*., 1991; Jones *et al*., 1992; Ouzounis *et al*., 1993) (problems still remaining unsolved today). High-throughput sequence similarity runs were being explored, with the clustering of the full protein sequence database (Gonnet *et al*., 1992). This activity denoted the beginning of the genome informatics era, celebrated by the computational re-annotation of the first ever entire chromosome sequence, yeast chromosome III (Bork *et al*., 1992). The rest, as they say, is history.

TODAY AND THE FUTURE

Given this short and rather subjective account on the development of bioinformatics, it is fair to ask what is the value of this kind of historical perspective. Two good reasons come to mind: first, it is important to both appreciate and understand the first steps into the unknown taken by a number of pioneers to open up a field that would later become a discipline with far-reaching implications for biological sciences; second, through this discursive history, it is evident that this field has grown and become an independent discipline with solutions of biological problems but with its own problems, solutions and further directions. Bioinformatics has become an independent scientific discipline, as old as computer science itself. Despite common perceptions, it is not 'just' a technology platform for genomics and systems biology, although its impact on those disciplines should not be underestimated. These data-driven fields, however, provide novel types of data which result in new kinds of problems and expanded horizons both for genomics and bioinformatics, in a healthy and fascinating interplay. Despite the fact that the actual origin of the term 'bioinformatics' still eludes us, it is clear that this discipline will continue to evolve rapidly into the 21st century, perhaps to a point beyond recognition. Merging with nanotechnology, computing with biological matter is expected to transform our own lives, in particular, and life on earth, in general. One day we may look back and understand how computation and experimentation with biological systems blurred the divide and allowed the 'great crossing' between the inanimate and the animate worlds.

ACKNOWLEDGEMENTS

Sincere apologies for omitting many citations due to space limitations. Thanks to Antoine Danchin, Arthur Lesk, Chris Sander, Janet Thornton, Anna Tramontano and referees for comments.

REFERENCES

Abarbanel,R.M., Wieneke,P.R., Mansfield,E., Jaffe,D.A. and Brutlag,D.L. (1984) Rapid searches for complex patterns in biological molecules. *Nucleic Acids Res.*, **12**, 263–280.

Aho,V.A., Hirschberg,D.S. and Ullman,J.D. (1976) Bounds on the complexity of the longest common subsequences problem. *J. ACM*, **23**, 1–12.

Albery,W.J. and Knowles,J.R. (1976) Evolution of enzyme function and the development of catalytic efficiency. *Biochemistry*, **15**, 5631–5640.

Alff-Steinberger,C. (1969) The genetic code and error transmission. *Proc. Natl Acad. Sci. USA*, **64**, 584–591.

Altschuh,D., Vernet,T., Berti,P., Moras,D. and Nagai,K. (1988) Coordinated amino acid changes in homologous protein families. *Protein Eng.*, **2**, 193–199.

Altschul,S.F. and Erickson,B.W. (1986) Optimal sequence alignment using affine gap costs. *Bull. Math. Biol.*, **48**, 603–616.

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.

Anfinsen,C.B. (1973) Principles that govern the folding of protein chains. *Science*, **181**, 223–230.

Anfinsen,C.B. and Scheraga,H.A. (1975) Experimental and theoretical aspects of protein folding. *Adv. Protein Chem.*, **29**, 205–300.

Arratia,R., Gordon,L. and Waterman,M. (1986) An extreme value theory for sequence matching. *Ann. Stat.*, **14**, 971–993.

Arratia,R. and Waterman,M.S. (1985a) Critical phenomena in sequence matching. *Ann. Prob.*, **13**, 1236–1249.

Arratia,R. and Waterman,M.S. (1985b) An Erdös–Rényi law with shifts. *Adv. Math.*, **55**, 13–23.

Bajaj,M. and Blundell,T. (1984) Evolution and the tertiary structure of proteins. *Ann. Rev. Biophys. Bioeng.*, **13**, 453–492.

Bashford,D., Chothia,C. and Lesk,A.M. (1987) Determinants of a protein fold: unique features of the globin amino acid sequences. *J. Mol. Biol.*, **196**, 199–216.

Beintema,J.J., Schüller,C., Irie,M. and Carsana,A. (1988) Molecular evolution of the ribonuclease superfamily. *Prog. Biophys. Mol. Biol.*, **51**, 165–192.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D. *et al*. (1977) The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.

Beyer,W.A., Stein,M.L., Smith,T.F. and Ulam,S.M. (1974) A molecular sequence metric and evolutionary trees. *Math. Biosci.*, **19**, 9–25.

Bilofsky,H.S., Burks,C., Fickett,J.W., Goad,W.B., Lewitter,F.I., Rindone,W.P., Swindell,C.D. and Tung,C.S. (1986) The GenBank genetic sequence data bank. *Nucleic Acids Res.*, **14**, 1–4.

Blake,R.D. and Earley,S. (1986) Distribution and evolution of sequence characteristics in the *E.coli* genome. *J. Biomol. Struct. Dyn.*, **4**, 291–307.

Bork,P., Ouzounis,C., Sander,C., Scharf,M., Schneider,R. and Sonnhammer,E. (1992) What's in a genome? *Nature*, **358**, 287–287.

Bowie,J.U., Luethy,R. and Eisenberg,D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.

Brändén,C.-I. (1980) Relation between structure and function of α/β proteins. *Qu. Rev. Biophys.*, **13**, 317–338.

Braun,W. (1983) Representation of short- and long-range handedness in protein structures by signed distance maps. *J. Mol. Biol.*, **163**, 613–621.

Braun,W. (1987) Distance geometry and related methods for protein structure determination from NMR data. *Qu. Rev. Biophys.*, **19**, 115–157.

Braun,W. and Gô,N. (1985) Calculation of protein conformations by proton–proton distance constraints. A new efficient algorithm. *J. Mol. Biol.*, **186**, 611–626.

Brenner,S. (1988) The molecular evolution of genes and proteins: a tale of two serines. *Nature*, **334**, 528–530.

Breslauer,K.J., Frank,R., Blöcker,H. and Marky,L.A. (1986) Predicting DNA duplex stability from the base sequence. *Proc. Natl Acad. Sci. USA*, **83**, 3746–3750.

Britten,R. (1986) Rates of DNA sequence evolution differ between taxonomic groups. *Science*, **231**, 1393–1398.

Britten,R.J. and Davidson,E.H. (1969) Gene regulation for higher cells: a theory. *Science*, **165**, 347–357.

Brooks,B. and Karplus,M. (1983) Fractal surfaces of proteins. *Proc. Natl Acad. Sci. USA*, **80**, 6571–6575.

Brunak,S., Engelbrecht,J. and Knudsen,S. (1990) Neural network detects errors in the assignment of mRNA splice sites. *Nucleic Acids Res.*, **18**, 4797–4801.

Brünger,A.T., Clore,M.G., Gronenborn,A.M. and Karplus,M. (1986) Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: application to crambin. *Proc. Natl Acad. Sci. USA*, **83**, 3801–3805.

Brutlag,D.L. (1980) Molecular arrangement and evolution of heterochromatic DNA. *Ann. Rev. Genet.*, **14**, 121–144.

Brutlag,D.L., Clayton,J., Friedland,P. and Kedes,L.H. (1982) SEQ: a nucleotide sequence analysis and recombination system. *Nucleic Acids Res.*, **10**, 279–294.

Bulmer,M. (1987) A statistical analysis of nucleotide sequence of introns and exons in human genes. *Mol. Biol. Evol.*, **4**, 395–405.

Burks,C. and Farmer,D. (1984) Towards modeling DNA sequences as automata. *Physica* D, **10**, 157–167.

Burks,C., Lawton,J.R. and Bell,G.I. (1988) The LiMB database. *Science*, **241**, 888–888.

Cannon,G.C. (1987) Sequence analysis on microcomputers. *Science*, **238**, 97–103.

Cantor,C.R. (1968) The occurrence of gaps in protein sequences. *Biochem. Biophys. Res. Comm.*, **31**, 410–416.

Cariani,P. and Goel,N.S. (1985) On the computation of the tertiary structure of globular proteins—IV. Use of secondary structure information. *Bull. Math. Biol.*, **47**, 367–407.

Carrillo,H. and Lipman,D.J. (1988) The multiple sequence alignment problem in biology. *SIAM J. Appl. Math.*, **48**, 1073–1082.

Cedergren,R., Gray,M.W., Abel,Y. and Sankoff,D. (1988) The evolutionary relationships among known life forms. *J. Mol. Evol.*, **28**, 98–112.

Chaitin,G.J. (1966) On the length of programs for computing finite binary sequences. *J. ACM*, **13**, 547–569.

Chomsky,N. (1959) On certain formal properties of grammar. *Inform. Control*, **2**, 137–167.

Chothia,C. (1975) Structural invariants in protein folding. *Nature*, **254**, 304–308.

Chothia,C. (1984) Principles that determine the structures of proteins. *Ann. Rev. Biochem.*, **53**, 537–572.

Chothia,C. and Janin,J. (1981) Relative orientations of close-packed β-pleated sheets in proteins. *Proc. Natl Acad. Sci. USA*, **78**, 4146–4150.

Chothia,C. and Lesk,A.M. (1982) Evolution of proteins formed by β-sheets. I. Plastocyanin and azurin. *J. Mol. Biol.*, **160**, 309–323.

Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.

Chothia,C., Levitt,M. and Richardson,D. (1977) Structures of proteins: packing of alpha-helices and pleated sheets. *Proc. Natl Acad. Sci. USA*, **74**, 4130–4134.

Chothia,C., Levitt,M. and Richardson,D. (1981) Helix to helix packing in proteins. *J. Mol. Biol.*, **145**, 215–250.

Chou,P.Y. and Fasman,G.D. (1974) Prediction of protein conformation. *Biochemistry*, **13**, 222–244/225.

Chou,P.Y. and Fasman,G.D. (1978) Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol.*, **47**, 45–148.

Chvátal,V. and Sankoff,D. (1975) Longest common subsequences of two random sequences. *J. Appl. Prob.*, **12**, 306–315.

Clarke,B. (1970) Selective constraints on amino-acid substitution during the evolution of proteins. *Nature*, **228**, 159–160.

Claverie,J.-M. and States,D.J. (1993) Information enhancement methods for large scale sequence analysis. *Comput. Chem.*, **17**, 191–201.

Cohen,C. and Parry,D.A.D. (1986) a-Helical coiled coils—a widespread motif in proteins. *Trends Biochem. Sci.*, **11**, 245–248.

Cohen,F.E. and Sternberg,M.J.E. (1980a) On the prediction of protein structure: the significance of the root-mean-square deviation. *J. Mol. Biol.*, **138**, 321–333.

Cohen,F.E. and Sternberg,M.J.E. (1980b) On the use of chemically derived distance constraints in the prediction of protein structure with myoglobin as an example. *J. Mol. Biol.*, **137**, 9–22.

Cohen,F.E., Sternberg,M.J.E. and Taylor,W.R. (1981) Analysis of the tertiary structure of protein β-sheet sandwiches. *J. Mol. Biol.*, **148**, 253–272.

Collins,J.F. and Coulson,A.F.W. (1984) Applications of parallel processing algorithms for DNA sequence analysis. *Nucleic Acids Res.*, **12**, 181–192.

Connolly,M.L. (1983) Solvent-accessible surfaces of protein and nucleic acids. *Science*, **221**, 709–713.

Conrad,M. (1985) On design principles for a molecular computer. *Comm. ACM*, **28**, 464–480.

Core,N.G., Edmiston,E.W., Saltz,J.H. and Smith,R.M. (1989) Supercomputers and biological sequence comparison algorithms. *Comput. Biomed. Res.*, **22**, 497–515.

Craik,C.S., Rutter,W.J. and Fletterick,R. (1983) Splice junctions: association with variation in protein structure. *Science*, **220**, 1125–1129.

Craik,C.S., Sprang,S., Fletterick,R. and Rutter,W.J. (1982) Intron–exon splice junctions map at protein surfaces. *Nature*, **299**, 180–182.

Crick,F.H.C. (1953) The packing of α-helices: simple coiled coil. *Acta Cryst.*, **6**, 689–697.

Crick,F.H.C. (1966) Codon–anticodon pairing: the wobble hypothesis. *J. Mol. Biol.*, **19**, 548–555.

Crick,F.H.C. (1968) The origin of the genetic code. *J. Mol. Biol.*, **38**, 367–379.

Crick,F.H.C. (1970) Central dogma of molecular biology. *Nature*, **227**, 561–563.

Crippen,G.M. (1977) A novel approach to the calculation of conformation: distance geometry. *J. Comput. Phys.*, **26**, 449–452.

Crippen,G.M. (1978) The tree structural organization of domains in globular proteins. *J. Mol. Biol.*, **126**, 315–332.

Crippen,G.M. (1987) Why energy embedding works. *J. Phys. Chem.*, **91**, 6341–6343.

Davison,D. (1985) Sequence similarity ('homology') searching for molecular biologists. *Bull. Math. Biol.*, **47**, 437–474.

Dayhoff,M.O. (1978) Atlas of Protein Sequence and Structure, Vol. 4, Suppl. 3. National Biomedical Research Foundation, Washington, D.C., U.S.A.

Dayhoff,M.O., Barker,W.C. and Hunt,L.T. (1983) Establishing homologies in protein sequences. *Meth. Enzymol.*, **91**, 524–545.

Delcoigne,A. and Hansen,P. (1975) Sequence comparison by dynamic programming. *Biometrika*, **62**, 661–664.

DeLisi,C. (1988) Computers in molecular biology: current applications and emerging trends. *Science*, **240**, 47–52.

Devereux,J., Haeberli,P. and Smithies,O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.*, **12**, 387–395.

DeWachter,R. (1981) The number of repeats expected in random nucleic acid sequences and found in genes. *J. Theor. Biol.*, **91**, 71–98.

Dickerson,R.E., Timkovich,R. and Almassy,R.J. (1976) The cytochrome fold and the evolution of bacterial energy metabolism. *J. Mol. Biol.*, **100**, 473–491.

Doolittle,R.F. (1981) Similar amino acid sequences: chance or common ancestry? *Science*, **214**, 149–159.

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, CA.

Doolittle,R.F. (1988) More molecular opportunism. *Nature*, **336**, 18–18.

Doolittle,R.F., Feng,D.-F., Johnson,M.S. and McClure,M.A. (1986) Origins and evolutionary relationships of retroviruses. *Qu. Rev. Biol.*, **64**, 1–30.

Doolittle,W.F. and Sapienza,C. (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature*, **284**, 601–603.

Dover,G. and Doolittle,W.F. (1980) Modes of genome evolution. *Nature*, **288**, 646–647.

Drexler,K.E. (1981) Molecular engineering: an approach to the development of general capabilities for molecular manipulation. *Proc. Natl Acad. Sci. USA*, **78**, 5275–5278.

Dumas,J.-P. and Ninio,J. (1982) Efficient algorithms for folding and comparing nucleic acid sequences. *Nucleic Acids Res.*, **10**, 197–206.

Dunnill,P. (1968) The use of helical net-diagrams to represent protein structures. *Biophys. J.*, **8**, 865–875.

Easthope,P.L. and Havel,T.F. (1989) Computational experience with an algorithm for tetrangle inequality bound smoothing. *Bull. Math. Biol.*, **51**, 173–194.

Ebeling,W. and Jiménez-Montaño,M.A. (1980) On grammars, complexity, and information measures of biological macromolecules. *Math. Biosci.*, **52**, 53–71.

Edmiston,E.W., Gore,N.G., Saltz,J.H. and Smith,R.M. (1988) Parallel processing of biological sequence comparison algorithms. *Int. J. Parallel Program*, **17**, 259–275.

Eisenberg,D., Schwarz,E., Komaromy,M. and Wall,R. (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.*, **179**, 125–142.

Epstein,C.J. (1967) Non-randomness of amino-acid changes in the evolution of homologous proteins. *Nature*, **215**, 355–359.

Eventoff,W. and Rossman,M.G. (1975) The evolution of dehydrogenases and kinases. *CRC Crit. Rev. Biochem.*, **3**, 111–140.

Feldmann,R.J. (1976) The design of computing systems for molecular modeling. *Ann. Rev. Biophys. Bioeng.*, **5**, 477–510.
Felsenstein,J. (1978) The number of evolutionary trees. *Syst. Zool.*, **27**, 27–33.
Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.
Felsenstein,J. (1982) Numerical methods for inferring evolutionary trees. *Qu. Rev. Biol.*, **57**, 379–404.
Felsenstein,J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.
Felsenstein,J. (1988a) PHYLIP: phylogeny inference package. *Cladistics*, **5**, 355–356.
Felsenstein,J. (1988b) Phylogenies from molecular sequences: inference and reliability. *Ann. Rev. Genet.*, **22**, 521–565.
Feng,D.-F. and Doolittle,R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, **25**, 351–360.
Feng,D.-F., Johnson,M.S. and Doolittle,R.F. (1985) Aligning amino acid sequences: commonly used methods. *J. Mol. Evol.*, **21**, 112–125.
Fickett,J.W. (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.*, **10**, 5303–5318.
Fickett,J.W. (1984) Fast optimal alignment. *Nucleic Acids Res.*, **12**, 175–179.
Fickett,J.W. and Tung,C.-S. (1992) Assessment of protein coding measures. *Nucleic Acids Res.*, **20**, 6441–6450.
Fitch,W.M. (1971) Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.*, **20**, 406–416.
Fitch,W.M. (1976) The molecular evolution of cytochrome *c* in eukaryotes. *J. Mol. Evol.*, **8**, 13–40.
Fitch,W.M. (1983) Random sequences. *J. Mol. Biol.*, **163**, 171–176.
Fitch,W.M. (1988) Examples, please. *Nature*, **334**, 19–19.
Fitch,W.M. and Margoliash,E. (1967) Construction of phylogenetic trees. *Science*, **155**, 279–284.
Fitch,W.M. and Margoliash,E. (1970) Usefulness of amino acid and nucleotide sequences in evolutionary studies. *Evol. Biol.*, **4**, 67–109.
Fitch,W.M. and Smith,T.F. (1983) Optimal sequence alignments. *Proc. Natl Acad. Sci. USA*, **80**, 1382–1386.
Florkin,M. (1962) Isologie, homologie, analogie et convergence en biochimie comparée. *Bull. Classe Sci. Acad. R. Belg.*, **48**, 819–824.
Fox,G.E., Stackenbrandt,E., Hespell,R.B., Gibson,J., Maniloff,J., Dyer,T.A., Wolfe,R.S., Balch,W.E., Tanner,R.S., Magrum,L.J. *et al.* (1980) The phylogeny of prokaryotes. *Science*, **209**, 457–463.
Fristensky,B. (1986) Improving the efficiency of dot-matrix similarity searches through use of an oligomer table. *Nucleic Acids Res.*, **14**, 597–610.
Galaktionov,S.G. and Rodionov,M.A. (1981) Calculation of the tertiary structure of proteins on the basis of analysis of the matrices of contacts between amino acid residues. *Biophysics*, **25**, 395–403.
Gamow,G., Rich,A. and Ycas,M. (1956) The problem of information transfer from nucleic acids to proteins. *Adv. Biol. Med. Phys.*, **4**, 23–68.
Garnier,J., Osguthorpe,D.J. and Robson,B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, **120**, 97–120.
Gatlin,L.L. (1966) The information content of DNA. *J. Theor. Biol.*, **10**, 281–300.
George,D.G., Hunt,T.L., Yeh,L.-S.L. and Barker,W.C. (1985) New perspectives on bacterial ferredoxin evolution. *J. Mol. Evol.*, **22**, 20–31.
Getzoff,E.D., Tainer,J.A., Stempien,M.M., Bell,G.I. and Hallewell,R.A. (1989) Evolution of CuZn superoxide dismutase and the Greek key β-barrel structural motif. *Proteins*, **5**, 322–336.
Gibbs,A.J., Dale,M.B., Kinns,H.R. and MacKenzie,H.G. (1971) The transition matrix method for comparing sequences; its use in describing and classifying proteins by their amino acid sequences. *Syst. Zool.*, **20**, 417–425.
Gibbs,A.J. and McIntyre,G.A. (1970) The diagram, a method for comparing sequences. *Eur. J. Biochem.*, **16**, 1–11.
Gilbert,W. (1978) Why genes in pieces? *Nature*, **271**, 501–501.
Gilbert,W. (1985) Genes-in-pieces revisited. *Science*, **228**.
Gilbert,W., Marchionni,M. and McKnight,G. (1986) On the antiquity of introns. *Cell*, **46**, 143–147.
Gingeras,T.R. and Roberts,R.J. (1980) Steps toward computer analysis of nucleotide sequences. *Science*, **209**, 1322–1328.
Gô,M. (1981) Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature*, **291**, 90–92.
Gô,M. (1983) Modular structural units, exons, and function in chicken lysozyme. *Proc. Natl Acad. Sci. USA*, **80**, 1964–1968.
Gô,M. (1985) Protein structures and split genes. *Adv. Biophys.*, **19**, 91–131.
Goad,W.B. (1986) Computational analysis of genetic sequences. *Ann. Rev. Bioph. Biophys. Chem.*, **15**, 79–95.
Goel,N.S., Rouyanian,B. and Sanati,M. (1982) On the computation of the tertiary structure of globular proteins. III. Inter-residue distances and computed structures. *J. Theor. Biol.*, **99**, 705–757.
Goel,N.S. and Ycas,M. (1979) On the computation of the tertiary structure of globular proteins. II. *J. Theor. Biol.*, **77**, 253–305.
Goldfarb,P.S. (1988) Evolution of modern proteins. *Nature*, **336**, 429–429.
Gonnet,G.H., Cohen,M.A. and Benner,S.A. (1992) Exhaustive matching of the entire protein sequence database. *Science*, **256**, 1443–1445.
Goodman,M., Moore,G.W. and Masuda,G. (1975) Darwinian evolution in the genealogy of haemoglobin. *Nature*, **253**, 603–608.
Gordon,A.D. (1973) A sequence-comparison statistic and algorithm. *Biometrika*, **60**, 197–200.
Gotoh,O. (1987) Pattern matching of biological sequences with limited storage. *Comput. Appl. Biosci.*, **3**, 17–20.
Gotoh,O. and Tagashira,Y. (1986) Sequence search on a supercomputer. *Nucleic Acids Res.*, **14**, 57–64.
Gower,J. (1985) Properties of Euclidean and non-Euclidean distance matrices. *Linear Algebra Appl.*, **67**, 81–97.
Gower,J.C. (1982) Euclidean distance geometry. *Math. Sci.*, **7**, 1–14.
Grantham,R. (1974) Amino acid difference formula to help explain protein evolution. *Science*, **185**, 862–864.
Grantham,R., Gautier,C., Gouy,M., Jacobzone,M. and Mercier,R. (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.*, **9**, r43–r74.
Grantham,R., Gautier,C., Gouy,M., Mercier,R. and Pavé,A. (1980) Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.*, **8**, r49–r62.

Graur,D. (1985) Amino acid composition and the evolutionary rates of protein-coding genes. *J. Mol. Evol.*, **22**, 53–62.

Greer,J. (1981) Comparative model-building of the mammalian serine proteases. *J. Mol. Biol.*, **153**, 1027–1042.

Gribskov,M. and Burgess,R. (1986) Sigma factors from *E.coli*, *B.subtilis*, phage SP01, and phage T4 are homologous proteins. *Nucleic Acids Res.*, **14**, 6745–6763.

Gribskov,M., Homyak,M., Edenfield,J. and Eisenberg,D. (1988) Profile scanning for three-dimensional structural patterns in protein sequences. *Comput. Appl. Biosci.*, **4**, 61–66.

Gribskov,M., McLachlan,M. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–5358.

Guibas,L.J. and Odlyzko,A.M. (1980) Long repetitive patterns in random sequences. *Z. Wahrschr. verw. Gebiete*, **53**, 241–262.

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

Hadwiger,M.A. and Fox,G.E. (1989) Distances as degrees of freedom. *J. Biomol. Struct. Dyn.*, **7**, 749–771.

Hagler,A.T. and Honig,B. (1978) On the formation of protein tertiary structure on a computer. *Proc. Natl Acad. Sci. USA*, **75**, 554–558.

Hall,P.A.V. and Dowling,G.R. (1980) Approximate string matching. *Comput. Surv.*, **12**, 381–402.

Hamm,G.H. and Cameron,G.N. (1986) The EMBL Data Library. *Nucleic Acids Res.*, **14**, 5–9.

Havel,T.F., Crippen,G.M., Kuntz,I.D. and Blaney,J.M. (1983a) The combinatorial distance geometry method for the calculation of molecular conformation II. Sample problems and computational statistics. *J. Theor. Biol.*, **104**, 383–400.

Havel,T.F., Kuntz,I.D. and Crippen,G.M. (1983b) The theory and practice of distance geometry. *Bull. Math. Biol.*, **45**, 665–720.

Havel,T.F. and Wüthrich,K. (1984) A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear megnetic resonance measurements of intramolecular ^{1}H–^{1}H proximities in solution. *Bull. Math. Biol.*, **46**, 673–698.

Heijne,G.v. (1981) On the hydrophobic nature of signal sequences. *Eur. J. Biochem.*, **116**, 419–422.

Heijne,G.v. (1985) Signal sequences. The limits of variation. *J. Mol. Biol.*, **184**, 99–105.

Heijne,G.v. (1987) *Sequence Analysis in Molecular Biology: Treasure Trove or Trivial Pursuit*. Academic Press, San Diego, CA.

Heijne,G.v. (1988) Getting sense out of sequence data. *Nature*, **333**, 605–607.

Heinrich,R. and Rapoport,T.A. (1977) Metabolic regulation and mathematical models. *Prog. Biophys. Mol. Biol.*, **32**, 1–82.

Henikoff,S. and Wallace,J.C. (1988) Detection of protein similarities using nucleotide sequence databases. *Nucleic Acids Res.*, **16**, 6191–6204.

Higgins,D.G. and Sharp,P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**, 237–244.

Hirschberg,D.S. (1975) A linear space algorithm for computing maximal common subsequences. *Commun. ACM*, **18**, 341–343.

Hodgman,T.C. (1986) The elucidation of protein function from its amino acid sequence. *Comput. Appl. Biosci.*, **2**, 181–188.

Hogeweg,P. and Hesper,B. (1984) The alignment of sets of sequences and the construction of phylogenetic trees. An integrated method. *J. Mol. Evol.*, **20**, 175–186.

Holm,L. and Sander,C. (1992) Fast and simple Monte Carlo algorithm for side chain optimization in proteins: application to model building by homology. *Proteins*, **14**, 213–223.

Holmquist,R., Jukes,T.H. and Pangburn,S. (1973) Evolution of transfer RNA. *J. Mol. Biol.*, **78**, 91–116.

Hopfield,J.J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl Acad. Sci. USA*, **79**, 2554–2558.

Hopp,T.P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl Acad. Sci. USA*, **78**, 3824–3828.

Horowitz,N.H. (1945) On the evolution of biochemical syntheses. *Proc. Natl Acad. Sci. USA*, **31**, 153–157.

Huang,X. (1989) A space-efficient parallel sequence comparison algorithm for a message passing multiprocessor. *Int. J. Parallel Program.*, **18**, 223–239.

Hwang,P.K. and Fletterick,R.J. (1986) Convergent and divergent evolution of regulatory sites in eukaryotic phosphorylases. *Nature*, **234**, 80–83.

Ingram,V.M. (1961) Gene evolution and the haemoglobins. *Nature*, **189**, 704–708.

Islam,S.A. and Sternberg,M.J.E. (1989) A relational database of protein structures designed for flexible enquiries about conformation. *Protein Eng.*, **2**, 431–442.

Iwabe,N., Kuma,K., Hasegawa,M., Osawa,S. and Miyata,T. (1989) Evolutionary relationship of archaebacteria, eubacteria and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl Acad. Sci. USA*, **86**, 9355–9359.

Janin,J. and Chothia,C. (1980) Packing of α-helices onto β-pleated sheets and the anatomy of α/β proteins. *J. Mol. Biol.*, **143**, 95–128.

Jiménez-Montaño,M.A. (1984) On the syntactic structure of protein sequences and the concept of grammar complexity. *Bull. Math. Biol.*, **46**, 641–659.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) A new approach to protein fold recognition. *Nature*, **358**, 86–89.

Jones,T.A. (1978) A graphics model building and refinement system for macromolecules. *J. Appl. Crystallogr.*, **11**, 268–272.

Jones,T.A. (1985) Interactive computer graphics: FRODO. *Meth. Enzymol.*, **115**, 157–171.

Jones,T.A. and Thirup,S. (1986) Using known substructures in protein model building and crystallography. *EMBO J.*, **5**, 819–822.

Jukes,T.H. (1969) Recent advances in studies of evolutionary relationships between proteins and nucleic acids. *Space Life Sci.*, **1**, 469–490.

Jukes,T.H. and Holmquist,R. (1972) Evolutionary clock: nonconstancy of rate in different species. *Science*, **177**, 530–532.

Jukes,T.H. and King,J.L. (1975) Evolutionary loss of ascorbic acid synthesizing ability. *J. Hum. Evol.*, **4**, 85–88.

Jungck,J.R. and Friedman,R.M. (1984) Mathematical tools for molecular genetics data: an annotated bibliography. *Bull. Math. Biol.*, **46**, 699–744.

Kabat,E.A. and Wu,T.T. (1973) The influence of nearest-neighboring amino acid residues on aspects of secondary structure of proteins. Attempts to locate α-helices and β-sheets. *Biopolymers*, **12**, 751–774.

Kabsch,W. (1976) A solution for the best rotation to relate two sets of vectors. *Acta Cryst.* A, **32**, 922–923.

Kabsch,W. and Sander,C. (1984) On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc. Natl Acad. Sci. USA*, **81**, 1075–1078.

Karlin,S., Ghandour,G., Ost,F., Tavaré,S. and Korn,L.J. (1983) New approaches for computer analysis of nucleic acid sequences. *Proc. Natl Acad. Sci. USA*, **80**, 5660–5664.

Karplus,M. and Weaver,D.L. (1976) Protein folding dynamics. *Nature*, **260**, 404–406.

Katz,L. and Levinthal,C. (1966) Molecular model-building by computer. *Sci. Am.*, **214**, 42–52.

Kelly,J.M. and Meyer,E.F.J. (1983) Storage and retrieval of nucleic acid sequence data. *Comput. Chem.*, **4**, 107–111.

Kimura,M. (1968) Evolutionary rate at the molecular level. *Nature*, **217**, 624–626.

Kimura,M. (1969) The rate of molecular evolution considered from the standpoint of population genetics. *Proc. Natl Acad. Sci. USA*, **63**, 1181–1188.

Kimura,M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Kimura,M. and Ohta,T. (1972) On the stochastic model for estimation of mutational distance between homologous proteins. *J. Mol. Evol.*, **2**, 87–90.

Kimura,M. and Ohta,T. (1974) On some principles governing molecular evolution. *Proc. Natl Acad. Sci. USA*, **71**, 2848–2852.

King,J.L. and Jukes,T.H. (1969) Non-Darwinian evolution. *Science*, **164**, 788–798.

Klein,P. (1986) Prediction of protein structural class by discriminant analysis. *Biochim. Biophys. Acta*, **874**, 205–215.

Klein,P. and DeLisi,C. (1986) Prediction of protein structural class from the amino acid sequence. *Biopolymers*, **25**, 1659–1672.

Klotz,L.C., Komar,N., Blanken,R.L. and Mitchell,R.M. (1979) Calculation of evolutionary trees from sequence data. *Proc. Natl Acad. Sci. USA*, **76**, 4516–4520.

Klug,A. and Rhodes,D. (1987) 'Zinc fingers': a novel protein motif for nucleic acid recognition. *Trends Biochem. Sci.*, **12**, 464–469.

Koch,R.E. (1971) The influence of neighboring base pairs upon base-pair substitution mutation rates. *Protein Natl Acad. Sci. USA*, **68**, 773–776.

Korn,L.J. and Queen,C.L. (1984) Analysis of biological sequences on small computers. *DNA*, **3**, 421–436.

Kristofferson,D. (1987) The BIONET electronic network. *Nature*, **325**, 555–556.

Kruskal,J.B. (1983) An overview of sequence comparison. In: Sankoff, D. and Kruskal, J.B. (eds) *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, Reading, MA, pp. 1–44.

Kruskal,J.B. and Sankoff,D. (1983) An anthology of algorithms and concepts for sequence comparison. In: Sankoff, D. and Kruskal, J.B. (eds) *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, Reading, MA, pp. 265–310.

Krzywicki,A. and Slonimski,P.P. (1967) Formal analysis of protein sequences: I. Specific long-range constraints in pair associations of amino acids. *J. Theor. Biol.*, **17**, 136–158.

Kuntz,I.D. (1975) An approach to the tertiary structure of globular proteins. *J. Am. Chem. Soc.*, **97**, 4362–4366.

Kuntz,I.D., Crippen,G.M., Kollman,P.A. and Kimelman,D. (1976) Calculation of protein tertiary structure. *J. Mol. Biol.*, **106**, 983–994.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Landschulz,W.H., Johnson,P.F. and McKnight,S.L. (1988) The leucine zipper: a hypothetical structure common to a new class of DNA-binding proteins. *Science*, **240**, 1759–1764.

Lasters,I., Wodak,S.J., Alard,P. and Cutsem,E.v. (1988) Structural principles of parallel β-barrels in proteins. *Proc. Natl Acad. Sci. USA*, **85**, 3338–3342.

Lathrop,R.H., Webster,T.A. and Smith,T.F. (1987) ARIADNE: pattern-directed inference and hierarchical abstraction in protein structure recognition. *Commun. ACM*, **30**, 909–921.

Lawrence,C.B., Goldman,D.A. and Hood,R.T. (1986) Optimized homology searches of the gene and protein sequence data banks. *Bull. Math. Biol.*, **48**, 569–583.

Lawton,J.R., Martinez,F.A. and Burks,C. (1989) Overview of the LiMB database. *Nucleic Acids Res.*, **17**, 5885–5899.

Lee,B. and Richards,F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, **55**, 379–400.

Lee,Y.M., Friedman,D.J. and Ayala,F.J. (1985) Superoxide dismutase: an evolutionary puzzle. *Proc. Natl Acad. Sci. USA*, **82**, 824–828.

Lesk,A.M. (1985) Coordination of sequence data. *Nature*, **314**, 318–319.

Lesk,A.M. (1987) The Biocomputing program at EMBL. *Trends Biotech.*, **5**, 317–318.

Lesk,A.M. (1988) The EMBL data library. In: Lesk, A.M. (ed.) *Computational Molecular Biology. Sources and Methods for Sequence Analysis*. Oxford University Press, Oxford, pp. 55–65.

Lesk,A.M. and Chothia,C. (1980) How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.*, **136**, 225–270.

Lesk,A.M. and Chothia,C. (1982) Evolution of proteins formed by β-sheets. II. The core of the immunoglobulin domains. *J. Mol. Biol.*, **160**, 325–342.

Lesk,A.M. and Hardman,K.D. (1982) Computer-generated schematic diagrams of protein structures. *Science*, **216**, 539–540.

Leszczynski,J.F. and Rose,G.D. (1986) Loops in globular proteins: a novel category of secondary structure. *Science*, **234**, 849–855.

Levin,L.A. (1973) On the notion of a random sequence. *Soviet Math. Dokl.*, **14**, 1413–1416.

Levitt,M. (1976) A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.*, **104**, 59–107.

Levitt,M. (1978) Conformational preferences of amino acids in globular proteins. *Biochemistry*, **17**, 4277–4285.

Levitt,M. (1983) Protein folding by restrained energy minimization and molecular dynamics. *J. Mol. Biol.*, **170**, 723–764.

Levitt,M. (1992) Accurate modeling of protein conformations by automatic segment matching. *J. Mol. Biol.*, **226**, 507–533.

Levitt,M. and Chothia,C. (1976) Structural patterns in globular proteins. *Nature*, **261**, 552–558.

Levitt,M. and Warshel,A. (1975) Computer simulation of protein folding. *Nature*, **253**, 694–698.

Lewin,R. (1984) National networks for molecular biologists. *Science*, **223**, 1379–1380.

Lifson,S. and Sander,C. (1979) Antiparallel and parallel beta-strands differ in amino acid residue preferences. *Nature*, **282**, 109–111.

Lifson,S. and Sander,C. (1980) Specific recognition in the tertiary structure of beta-sheets of proteins. *J. Mol. Biol.*, **139**, 627–639.

Liljas,A. and Rossman,M.G. (1974) Recognition of structural domains in globular proteins. *J. Mol. Biol.*, **85**, 177–181.

Lim,V.I. (1974) Algorithms for prediction of α-helical and β-structural regions in globular proteins. *J. Mol. Biol.*, **88**, 873–894.

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

Loomis,N.F. and Gilpin,M.E. (1986) Multigene families and vestigial sequences. *Proc. Natl Acad. Sci. USA*, **83**, 2143–2147.

Lopresti,D. (1987) P-NAC: a systolic array for comparing nucleic acid sequences. *Computer*, **20**, 98–99.

Lowrance,R. and Wagner,R.A. (1975) An extension of the string-to-string correction problem. *J. ACM*, **22**, 177–183.

Lubsen,N.H., Aarts,H.J.M. and Schoenmakers,J.G.G. (1988) The evolution of lenticular proteins: the β- and γ-crystallin super gene family. *Prog. Biophys. Mol. Biol.*, **51**, 47–76.

Lyall,A., Hammond,P., Brough,D. and Glover,D. (1984) BIOLOG—a DNA sequence analysis system in Prolog. *Nucleic Acids Res.*, **12**, 633–642.

Maizel,J.V.J. and Lenk,R.P. (1981) Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proc. Natl Acad. Sci. USA*, **78**, 7665–7669.

Margoliash,E. (1963) Primary structure and evolution of cytochrome *c*. *Proc. Natl Acad. Sci. USA*, **50**, 672–679.

Martin-Löf,P. (1966) The definition of random sequences. *Inform. Control*, **9**, 602–619.

Martinez,H.M. (1983) An efficient method for finding repeats in molecular sequences. *Nucleic Acids Res.*, **11**, 4629–4634.

Mathews,F.S. (1985) The structure, function and evolution of cytochromes. *Prog. Biophys. Mol. Biol.*, **45**, 1–56.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

McLachlan,A.D. (1982) Rapid comparison of protein structures. *Acta Cryst.* A, **38**, 871–873.

Metzler,W.J., Hare,D.R. and Pardi,A. (1989) Limited sampling of conformational space by the distance geometry algorithm: implications for structures generated from NMR data. *Biochemistry*, **28**, 7045–7052.

Miyazawa,S. and Jernigan,R.L. (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**, 534–552.

Mural,R.J. and Uberbacher,E.C. (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl Acad. Sci. USA*, **88**, 11261–11265.

Murata,M., Richardson,J.S. and Sussman,J.L. (1985) Simultaneous comparison of three protein sequences. *Proc. Natl Acad. Sci. USA*, **82**, 3073–3077.

Nagano,K. and Hasegawa,K. (1975) Logical analysis of the mechanism of protein folding. *J. Mol. Biol.*, **94**, 257–281.

Naora,H. and Deacon,N.J. (1982) Relationship between the total size of exons and introns in protein-coding genes of higher eukaryotes. *Proc. Natl Acad. Sci. USA*, **79**, 6196–6200.

Naora,H., Miyahara,K. and Curnow,R.N. (1987) Origin of non-coding DNA sequences: molecular fossils of genome evolution. *Proc. Natl Acad. Sci. USA*, **84**, 6195–6199.

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.

Nei,M. (1969) Gene duplication and nucleotide substitution in evolution. *Nature*, **221**, 40–42.

Neumann,J.v. (1966) *Theory of Self-Reproducing Automata*. University of Illinois Press, Urbana, IL.

Neumann,J.v. and Morgenstern,O. (1953) *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, USA.

Neurath,H. (1984) Evolution of proteolytic enzymes. *Science*, **224**, 350–357.

Nishikawa,K., Kubota,Y. and Ooi,T. (1983a) Classification of proteins into groups based on amino acid composition and other characters. I. *J. Biochem.*, **94**, 981–995.

Nishikawa,K., Kubota,Y. and Ooi,T. (1983b) Classification of proteins into groups based on amino acids composition and other characters. II. *J. Biochem.*, **94**, 997–1007.

Nolan,C. and Margoliash,E. (1968) Comparative aspects of primary structures of proteins. *Ann. Rev. Biochem.*, **37**, 727–791.

Novotny,J. (1973) Genealogy of immunoglobulin polypeptide chains: a consequence of amino acid interactions, conserved in their tertiary structure. *J. Theor. Biol.*, **41**, 171–180.

Novotny,J. (1982) Matrix program to analyze primary structure homology. *Nucleic Acids Res.*, **10**, 127–131.

Novotny,J., Bruccoleri,R.E. and Karplus,M. (1984) An analysis of incorrectly folded models. Implications for structure prediction. *J. Mol. Biol.*, **177**, 787–818.

Nussinov,R. (1983) Efficient algorithms for searching for exact repetition of nucleotide sequences. *J. Mol. Evol.*, **19**, 283–285.

Nussinov,R. and Jacobson,A.B. (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl Acad. Sci. USA*, **77**, 6309–6313.

Ochman,H. and Wilson,A.C. (1987) Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. *J. Mol. Evol.*, **26**, 74–86.

Ohno,S. (1970) *Evolution by Gene Duplication*. Springer-Verlag, New York.

Ohta,T. (1987) Simulating evolution by gene duplication. *Genetics*, **115**, 207–213.

Ohta,T. and Kimura,M. (1971) Functional organization of genetic material as a product of molecular evolution. *Nature*, **233**, 118–119.

Okuda,T., Tanaka,E. and Kasai,T. (1976) A method for correction of garbled words based on the Levenshtein metric. *IEEE Trans. Comput. C*, **25**, 172–177.

Orcutt,B.C. and Barker,W.C. (1984) Searching the protein sequence database. *Bull. Math. Biol.*, **46**, 545–552.

Orcutt,B.C., George,D.G. and Dayhoff,M.O. (1983) Protein and nucleic acid sequence database systems. *Ann. Rev. Biophys. Bioeng.*, **12**, 419–441.

Ouzounis,C., Sander,C., Scharf,M. and Schneider,R. (1993) Prediction of protein structure by evaluation of sequence-structure fitness: aligning sequences to contact profiles derived from three-dimensional structures. *J. Mol. Biol.*, **232**, 805–825.

Pace,N.R., Olsen,G.J. and Woese,C.R. (1986) Ribosomal RNA phylogeny and the primary lines of evolutionary descent. *Cell*, **45**, 325–326.

Pain,R.H. and Robson,B. (1970) Analysis of the code relating sequence to secondary structure in proteins. *Nature*, **227**, 62–63.

Pauling,L. and Corey,R.B. (1953) Two pleated-sheet configurations of polypeptide chains involving both cis and trans amide groups. *Proc. Natl Acad. Sci. USA*, **39**, 247–252.

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

Perlman,P.S. and Butow,R.A. (1989) Mobile introns and intron-encoded proteins. *Science*, **246**, 1106–1109.

Philipson,L. (1988) The DNA data libraries. *Nature*, **332**, 676–676.

Piatigorsky,J. and Wistow,G.J. (1989) Enzyme/crystallins: gene sharing as an evolutionary strategy. *Cell*, **57**, 197–199.

Ponder,J.W. and Richards,F.M. (1987) Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.*, **193**, 775–791.

Priestle,J.P. (1988) RIBBON: a stereo cartoon drawing program for proteins. *J. Appl. Crystallogr.*, **21**, 572–576.

Ptitsyn,O.B. (1969) Statistical analysis of the distribution of amino acid residues among helical and non-helical regions in globular proteins. *J. Mol. Biol.*, **42**, 501–510.

Ptitsyn,O.B. and Finkelstein,A.V. (1980) Similarities of protein topologies: evolutionary divergence, functional convergence or principles of folding? *Qu. Rev. Biophys.*, **13**, 339–386.

Pustell,J. and Kafatos,F.C. (1984) A convenient and adaptable package of computer programs for DNA and protein sequence management, analysis and homology determination. *Nucleic Acids Res.*, **12**, 643–655.

Rackovsky,S. and Goldstein,D.A. (1988) Protein comparison and classification: a differential geometric approach. *Proc. Natl Acad. Sci. USA*, **85**, 777–781.

Ramachandran,G.N., Ramakrishnan,C. and Sasisekharan,V. (1963) Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, **7**, 95–99.

Rashin,A.A. (1981) Locations of domains in globular proteins. *Nature*, **291**, 85–86.

Rawlings,C.J. (1986) *Software Directory for Molecular Biologists*. McMillan, New York.

Rawlings,C.J. (1988) Designing databases for molecular biology. *Nature*, **334**, 477–477.

Reanney,D.C. (1986) Genetic error and genome design. *Trends Genet.*, **2**, 41–46.

Reeck,G.R., Haën,C.d., Teller,D.C., Doolittle,R.F., Witch,W.M., Dickerson,R.E., Chambon,P., McLachlan,A.D., Margoliash,E., Jukes,T.H. *et al*. (1987) "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell*, **50**, 667–667.

Reggia,J.A., Armentrout,S.L., Chou,H.-H. and Peng,Y. (1993) Simple systems that exhibit self-directed replication. *Science*, **259**, 1282–1287.

Richards,F.M. (1974) The interpretation of protein structures: total volume, group volume distribution and packing density. *J. Mol. Biol.*, **82**, 1–14.

Richards,F.M. (1977) Areas, volumes, packing and protein structures. *Ann. Rev. Biophys. Bioeng.*, **6**, 151–176.

Richardson,D.C. and Richardson,J.S. (1992) The kinemage: a tool for scientific communication. *Protein Sci.*, **1**, 3–9.

Richardson,J. (1981a) Anatomy and taxonomy of protein structure. *Adv. Protein Chem.*, **34**, 168–339.

Richardson,J.S. (1977) β-Sheet topology and the relatedness of proteins. *Nature*, **268**, 495–500.

Richardson,J.S. (1981b) The anatomy and taxonomy of protein structure. *Adv. Protein Chem.*, **34**, 167–339.

Riley,M. and Anilionis,A. (1978) Evolution of the bacterial genome. *Ann. Rev. Microbiol.*, **32**, 519–560.

Robson,B. (1974) Analysis of the code relating sequence to conformation in globular proteins—theory and application of expected information. *Biochem. J.*, **141**, 853–867.

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

Rose,G.D. (1979) Hierarchic organization of domains in globular proteins. *J. Mol. Biol.*, **134**, 447–470.

Rossmann,M.G. and Argos,P. (1976) Exploring structural homology of proteins. *J. Mol. Biol.*, **105**, 75–95.

Rossmann,M.G. and Argos,P. (1980) Three-dimensional coordinates from stereo diagrams of molecular structures. *Acta Crystallogr.*, **36**, 819–823.

Rost,B. and Sander,C. (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl Acad. Sci. USA*, **90**, 7558–7562.

Rothschild,L.J., Ragan,M.A., Coleman,A.W., Heywood,P. and Gerbi,S.A. (1986) Are rRNA sequence comparisons the Rosetta Stone of phylogenetics? *Cell*, **47**, 640–640.

Sackin,M.J. (1971) Crossassociation: a method of comparing protein sequences. *Biochem. Genet.*, **5**, 287–313.

Sankoff,D. (1972) Matching sequences under deletion/insertion constraints. *Proc. Natl Acad. Sci. USA*, **69**, 4–6.

Sankoff,D. and Cedergren,R.J. (1973) A test for nucleotide sequence homology. *J. Mol. Biol.*, **77**, 159–164.

Sankoff,D. and Cedergren,R.J. (1983) Simultaneous comparison of three or more sequences related by a tree. In: Sankoff, D. and Kruskal, J.B. (eds) *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, Reading, MA, pp. 253–263.

Sankoff,D. and Goldstein,M. (1989) Probabilistic models of genome shuffling. *Bull. Math. Biol.*, **51**, 117–124.

Sankoff,D. and Sellers,P.H. (1973) Shortcuts, diversions and maximal chains in partially ordered sets. *Discr. Math.*, **4**, 287–293.

Sattath,S. and Tvertsky,A. (1977) Additive similarity trees. *Psychometrika*, **42**, 319–345.

Savageau,M.A. (1979a) Allometric morphogenesis of complex systems: derivation of the basic equations from first principles. *Proc. Natl Acad. Sci. USA*, **76**, 6023–6025.

Savageau,M.A. (1979b) Growth of complex systems can be related to the properties of their underlying determinants. *Proc. Natl Acad. Sci. USA*, **76**, 5413–5417.

Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374–374.

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

Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415–431.

Schulz,G.E. (1977) Recognition of phylogenetic relationships from polypeptide chain fold similarities. *J. Mol. Evol.*, **9**, 339–342.

Schulz,G.E., Baryy,C.D., Friedman,J., Chou,P.Y., Fasman,G.D., Finkelstein,A.V., Lim,V.I., Pititsyn,O.B., Kabat,E.A., Wu,T.T. *et al*. (1974) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Nature*, **250**, 140–142.

Schulz,G.E. and Schirmer,R.H. (1979) Prediction of secondary structure from the amino acid sequence. In: *Principles of Protein Structure*. Springer-Verlag, Berlin, pp. 108–130.

Schwartz,R.M. and Dayhoff,M.O. (1978) Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. *Science*, **199**, 395–403.

Sellers,P.H. (1974a) An algorithm for the distance between two finite sequences. *J. Combin. Theor. A*, **16**, 253–258.

Sellers,P.H. (1974b) On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.*, **26**, 787–793.

Sellers,P.H. (1980) The theory and computation of evolutionary distances: pattern recognition. *J. Algorithms*, **1**, 359–373.

Sellers,P.H. (1984) Pattern recognition in genetic sequences by mismatch density. *Bull. Math. Biol.*, **46**, 501–514.

Senapathy,P. (1986) Origin of eukaryotic introns: a hypothesis, based on codon distribution statistics in genes, and its implications. *Proc. Natl Acad. Sci. USA*, **83**, 2133–2137.

Shannon,C.E. and Weaver,W. (1962) *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL.

Sharp,P. (1985) On the origin of RNA splicing and introns. *Cell*, **42**, 397–400.

Shepard,R.N. (1980) Multidimensional scaling, tree-fitting and clustering. *Science*, **210**, 390–398.

Shepherd,J.C.W. (1981) Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proc. Natl Acad. Sci. USA*, **78**, 1596–1600.

Sibanda,B.L. and Thornton,J.L. (1985) β-Hairpin families in globular proteins. *Nature*, **316**, 170–174.

Sippl,M.J. (1980) On the problem of comparing protein structures. *J. Mol. Biol.*, **156**, 359–388.

Sippl,M.J. and Scheraga,H.A. (1985) Solution of the embedding problem and decomposition of symmetric matrices. *Proc. Natl Acad. Sci. USA*, **82**, 2197–2201.

Smith,D.H., Brutlag,D.L., Friedland,P. and Kedes,L.H. (1986) BIONET: a national computer resource for molecular biology. *Nucleic Acids Res.*, **14**, 17–20.

Smith,T.F. and Waterman,M.S. (1981a) Comparison of biosequences. *Adv. Appl. Math.*, **2**, 482–489.

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

Sogin,M.L., Elwood,H.J. and Gunderson,J.H. (1986) Evolutionary diversity of eukaryotic small-subunit rRNA genes. *Proc. Natl Acad. Sci. USA*, **83**, 1383–1387.

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

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.

States,D.J. and Botstein,D. (1991) Molecular sequence accuracy and the analysis of protein coding regions. *Proc. Natl Acad. Sci. USA*, **88**, 5518–5522.

Steele,J.M. (1982) Long common subsequences and the proximity of two random strings. *SIAM J. Appl. Math.*, **42**, 731–737.

Sternberg,M.J.E. and Thornton,J.M. (1978) Prediction of protein structure from amino acid sequence. *Biochem. Soc. Trans.*, **6**, 1119–1123.

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.

Swanson,R. (1984) A vector representation for amino acid sequences. *Bull. Math. Biol.*, **46**, 623–639.

Sweet,R.M. and Eisenberg,D. (1983) Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J. Mol. Biol.*, **171**, 479–488.

Szent-Györgyi,A.G. and Cohen,C. (1957) Role of proline in polypeptide chain configuration of proteins. *Science*, **126**, 697.

Tanaka,S. and Scheraga,H.A. (1975) Model of protein folding: inclusion of short-, medium-, and long-range interactions. *Proc. Natl Acad. Sci. USA*, **72**, 3802–3806.

Tavaré,S. (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. In: Miura, R.M. (ed.) *Some Mathematical Questions in Biology—DNA Sequence Analysis*, Vol. 17. American Mathematical Society, Providence, RI, pp. 57–86.

Taylor,W.R. (1986) The classification of amino acid conservation. *J. Theor. Biol.*, **119**, 205–218.

Taylor,W.R. and Orengo,C.A. (1989) Protein structure alignment. *J. Mol. Biol.*, **208**, 1–22.

Thornton,J.M. (1981) Disulfide bridges in globular proteins. *J. Mol. Biol.*, **151**, 261–287.

Thornton,J.M., Flores,T.P., Jones,D.T. and Swindells,M.B. (1992) Prediction of progress at last. *Nature*, **354**, 105–106.

Thornton,J.M. and Gardner,S.P. (1989) Protein motifs and data-base searching. *Trends Biochem. Sci.*, **14**, 300–304.

Tinoco,I., Uhlenbeck,O.C. and Levine,M.D. (1971) Estimation of secondary structure in ribonucleic acids. *Nature*, **230**, 362–367.

Trifonov,E.N. and Sussman,J.L. (1980) The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc. Natl Acad. Sci. USA*, **77**, 3816–3820.

Turing,A.M. (1952) The chemical basis for morphogenesis. *Phil. Trans. R. Soc. London B*, **237**, 37–72.

Turner,D.H., Sugimoto,N. and Freier,S.M. (1988) RNA structure prediction. *Ann. Rev. Biophys. Biophys. Chem.*, **17**, 167–192.

Ukkonen,E. (1985) Algorithms for approximate string matching. *Inform. Control*, **64**, 100–118.

Unger,R., Harel,D., Wherland,S. and Sussman,J.L. (1989) A 3D building blocks approach to analyzing and predicting structure in proteins. *Proteins*, **5**, 355–373.

Wagner,R.A. and Fischer,M.J. (1974) The string to string correction problem. *J. ACM*, **21**, 168–173.

Wako,H. and Scheraga,H. (1982) Distance-constraint approach to protein folding. II. Prediction of the three-dimensional structure of bovine pancreatic trypsin inhibitor. *J. Protein Chem.*, **1**, 85–117.

Wako,H. and Scheraga,H.A. (1981) On the use of distance constraints to fold a protein. *Macromolecules*, **14**, 961–969.

Walker,E.J., Saraste,M., Runwick,M.J. and Gay,N.J. (1982) Distantly related sequences in the α- and β-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.*, **1**, 945–951.

Walls,P.H. and Sternberg,M.J. (1992) New algorithm to model protein–protein recognition based on surface complementarity. Applications to antibody–antigen docking. *J. Mol. Biol.*, **228**, 277–297.

Warme,P.K. and Morgan,R.S. (1978) A survey of amino acid side-chain interactions in 21 proteins. *J. Mol. Biol.*, **118**, 289–304.

Waterman,M.S. (1983) Sequence alignments in the neighborhood of the optimum with general application to dynamic programming. *Proc. Natl Acad. Sci. USA*, **80**, 3123–3124.

Waterman,M.S., Arratia,R. and Galas,D.J. (1984) Pattern recognition in several sequences: consensus and alignment. *Bull. Math. Biol.*, **46**, 515–527.

Waterman,M.S. and Smith,T.F. (1978a) On the similarity of dendrograms. *J. Theor. Biol.*, **73**, 789–800.

Waterman,M.S. and Smith,T.F. (1978b) RNA secondary structure: a complete mathematical analysis. *Math. Biosci.*, **42**, 257–266.

Waterman,M.S., Smith,T.F. and Beyer,W.A. (1976) Some biological sequence metrics. *Adv. Math.*, **20**, 367–387.

Waterman,M.S., Smith,T.F., Singh,M. and Beyer,W.A. (1977) Additive evolutionary trees. *J. Theor. Biol.*, **64**, 199–213.

Watson,J.D. and Crick,F.H.C. (1953) Genetic implications of the structure of deoxyribonucleic acid. *Nature*, **171**, 964–967.

Weber,P.C. and Salemme,F.R. (1980) Structural and functional diversity in four-α-helical proteins. *Nature*, **287**, 82–84.

West,M.W. and Ponnamperuma,C. (1970) Chemical evolution and the origin of life. *Space Life Sci.*, **2**, 225–295.

Wetlaufer,D.B. (1973) Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl Acad. Sci. USA*, **70**, 697–701.

Wilbur,W.J. (1985) On the PAM matrix model of protein evolution. *Mol. Biol. Evol.*, **2**, 434–447.

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

Wilbur,W.J. and Lipman,D.J. (1984) The context dependent comparison of biological sequences. *SIAM J. Appl. Math.*, **44**, 557–567.

Woese,C.R. (1970) The problem of evolving a genetic code. *BioScience*, **20**, 471–485.

Woese,C.R. (1987) Bacterial evolution. *Microbiol. Rev.*, **51**, 221–271.

Wolfram,S. (1984) Cellular automata as models of complexity. *Nature*, **311**, 419–424.

Wu,T.T., Fitch,W.M. and Margoliash,E. (1974) The information content of protein amino acid sequences. *Ann. Rev. Biochem.*, **43**, 539–566.

Wüthrich,K. (1989) Protein structure determination in solution by nuclear magnetic resonance spectroscopy. *Science*, **243**, 45–50.

Yamamoto,K. and Yoshikura,H. (1986) A new representation of protein structure: vector diagram. *Comput. Appl. Biosci.*, **2**, 83–88.

Ycas,M., Goel,N.S. and Jacobsen,J.W. (1978) On the computation of the tertiary structure of globular proteins. *J. Theor. Biol.*, **72**, 443–457.

Zuckerkandl,E. and Pauling,L. (1965a) Evolutionary divergence and convergence in proteins. In: Bryson, V. and Vogel, H.J. (eds) *Evolving Genes and Proteins*. Academic Press, New York, pp. 97–166.

Zuckerkandl,E. and Pauling,L. (1965b) Molecules as documents of evolutionary history. *J. Theor. Biol.*, **8**, 357–366.