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 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
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 Słomiński, 1967), derivation of preferences for amino acid residues in secondary structures (Pain and Robson, 1970; Pitsyn, 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; Wetlauffer, 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 1. Ten institutions that pioneered and fostered computation in biology**

<table>
<thead>
<tr>
<th>Institutions</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birkbeck College, University of London</td>
<td>UK</td>
</tr>
<tr>
<td>Boston University</td>
<td>USA</td>
</tr>
<tr>
<td>European Molecular Biology Laboratory (EMBL)</td>
<td>DE and EMBL</td>
</tr>
<tr>
<td>Institute of Protein Research, Academy of Sciences, Pushchino</td>
<td>Former USSR</td>
</tr>
<tr>
<td>Laboratory of Molecular Biology (LMB), MRC</td>
<td>UK</td>
</tr>
<tr>
<td>Los Alamos National Laboratory (LANL)</td>
<td>USA</td>
</tr>
<tr>
<td>National Biomedical Research Foundation (NBRF), Georgetown U</td>
<td>USA</td>
</tr>
<tr>
<td>Stanford University</td>
<td>USA</td>
</tr>
<tr>
<td>University of California San Francisco (UCSF)</td>
<td>USA</td>
</tr>
<tr>
<td>University College, University of London (UCL)</td>
<td>UK</td>
</tr>
</tbody>
</table>

---

**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, 1928), formal studies of protein structures (Pain and Robson, 1970; Ptitsyn, 1969), the invention of preferences for amino acid residues in secondary structure (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; Wetlauffer, 1973).
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; Rossman 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 (Grandham 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; Peng et al., 1985; Hegeweg 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 (Gribkov 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 (Gribkov and Burgess, 1986) and the nature of signal sequences (Heijne, 1981, 1985). Other studies included optimality in sequence alignment (Altshul et al., 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 (Gotow, 1987; Korn and Queen, 1984). Algorithms for the prediction of antigenic determinants (Hopp and Woods, 1981), the detection of open reading frames (Fickett, 1982, 1987; 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 LiM 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...
for more efficient analysis of biological sequences emerged experimentation with various dedicated hardware platforms
forly devoted to computational biology (Lesk, 1987). Finally, such as EMBL formed the very first departments exclus-
ively of age (Heijne, 1988). Entire programs in various institutes distributed computing for the wider community was coming

(ii) 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 (Roosman and Kodak, 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; McClellan, 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].

(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,
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 ground-breaking 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


Havel, T.F., Crippen, G.M., Kuntz, I.D. and Blaney, J.M. (1983a) The


Hadwiger, M.A. and Fox, G.E. (1989) Distances as degrees of

Gribskov, M. and Burgess, R. (1986) Sigma factors from E.coli,

Greer, J. (1981) Comparative model-building of the mammalian


C.A.Ouzounis and A.Valencia


