1

Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective

Attwood, T.K.[1], Gisel, A.[2], Eriksson, N-E.[3] and Bongcam-Rudloff, E.[4]
[1]*Faculty of Life Sciences & School of Computer Science, University of Manchester*
[2]*Institute for Biomedical Technologies, CNR*
[3]*Uppsala Biomedical Centre (BMC), University of Uppsala*
[4]*Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences*
[1]*UK*
[2]*Italy*
[3,4]*Sweden*

1. Introduction

The origins of bioinformatics, both as a term and as a discipline, are difficult to pinpoint. The expression was used as early as 1977 by Dutch theoretical biologist Paulien Hogeweg when she described her main field of research as bioinformatics, and established a bioinformatics group at the University of Utrecht (Hogeweg, 1978; Hogeweg & Hesper, 1978). Nevertheless, the term had little traction in the community for at least another decade. In Europe, the turning point seems to have been *circa* 1990, with the planning of the "*Bioinformatics in the 90s*" conference, which was held in Maastricht in 1991. At this time, the National Center for Biotechnology Information (NCBI) had been newly established in the United States of America (USA) (Benson *et al*., 1990). Despite this, there was still a sense that the nation lacked a "*long-term biology 'informatics' strategy*", particularly regarding postdoctoral interdisciplinary training in computer science and molecular biology (Smith, 1990). Interestingly, Smith spoke here of 'biology informatics', not bioinformatics; and the NCBI was a 'center for biotechnology information', not a bioinformatics centre.

The discipline itself ultimately grew organically from the needs of researchers to access and analyse (primarily biomedical) data, which appeared to be accumulating at alarming rates simultaneously in different parts of the world. The rapid collection of data was a direct consequence of a series of enormous technological leaps that yielded what was considered, at the time, unprecedented quantities of biological *sequence* information. Hot on the heels of these developments was the concomitant wide-scale blossoming of algorithms and computational resources necessary to analyse, manipulate and store these growing quantities of data. Together, these advances gave birth to the field we now refer to as bioinformatics.

When we look back, it's clear that certain concepts and historical milestones were crucial to the evolution of this new field. Those we think most important, and consequently

remember, depend largely on the perspective from which we view the emerging bioinformatics landscape. This chapter takes a largely European standpoint, while recognising that the development of bioinformatics in Europe was intimately coupled with parallel advances elsewhere in the world, and especially in the USA. The history is intricate. Here, we endeavour to recount the story as it unfolded along a number of tightly interwoven paths, including the rise and spread of some of the technological developments that spawned the data deluge and facilitated its world-wide propagation; of some of the databases that developed in order to store the rapidly accumulating data; and of some of the organisations and infrastructural initiatives that emerged to try to put some of those pivotal databases on a more solid financial footing.

2. The seeds of bioinformatics

It is hard to pinpoint where and when the seeds of bioinformatics were originally sown. Does the story start with Franklin and Gosling's foundational work towards the elucidation of the structure of DNA (Franklin & Gosling, 1953a, b, c), or with the opportunistic interpretation of their data by Watson and Crick (Watson & Crick, 1953)? Do we fast-forward to the ground-breaking work of Kendrew *et al.* (1958) and of Muirhead & Perutz (1963) in determining the first three-dimensional (3D) structures of proteins? Or do we step back, and focus on the painstaking work of Sanger, who, in 1955, determined the amino acid sequence of the first peptide hormone? Or again, do we jump ahead to the progenitors of the first databases of macromolecular structures and sequences in the mid-1960s and early '70s? This era clearly heralded some of the most significant advances in molecular biology, as witnessed by a string of Nobel Prizes at the time: *e.g.*, Sanger's Prize in Chemistry in 1958; Watson, Crick and Wilkins' shared Prize in Physiology or Medicine in 1962, following Franklin's death; and Perutz and Kendrew's Prize in Chemistry, also in 1962. Clearly, in its own way, each of these advances played an important part in the emergence of the vibrant new field that we recognise today as 'bioinformatics'.

As a humbling reference point, we have chosen to begin our story in the mid 1940s, with Fred Sanger's pioneering work on insulin. Sanger used a range of chemical and enzymatic techniques to elucidate, for the first time, the order of amino acids in the primary structure of a protein. Back then, this was a tremendously complex puzzle to tackle, and its completion required the successful resolution of many different challenges over several years. That this was a difficult incremental process is illustrated by the fact that, between 1945 and 1955, each step was published in a separate, stand-alone article. All in all, something like 10 papers detail the series of experiments that led to the eventual determination of the sequences of bovine insulin (*e.g.*, Sanger, 1945; Sanger & Tuppy, 1951a, b; Sanger & Thompson, 1953a,b; Sanger *et al.*, 1955; Ryle *et al.*, 1955) and of ovine and porcine insulins (Brown *et al.*, 1955). This was ground-breaking work, and had taken 10 years to complete. Incredibly, the 3D structure would not be known for another 14 years (Adams *et al.*, 1969). The primary and tertiary structures of this historical protein are illustrated in Figure 1.

Such was the enormity of manual sequencing projects that it was many years before the sequence of the first enzyme (ribonuclease) was determined. Work on this protein began in 1955. After preliminary studies in 1957 and 1958, the first full 'draft sequence' was published in 1960 (Hirs *et al.*, 1960). During the months that followed, the draft was meticulously refined, and a final version was published 3 years later (Smyth *et al.*, 1963). Crucially, this 8-

year project paved the way for the elucidation of the protein's 3D structure – indeed, without the sequence information, the electron density maps could not have been meaningfully interpreted (Wyckoff *et al*., 1967). Knowledge of the primary structure of this small protein thus provided a vital piece of a 3D jigsaw puzzle that was to take a further 4

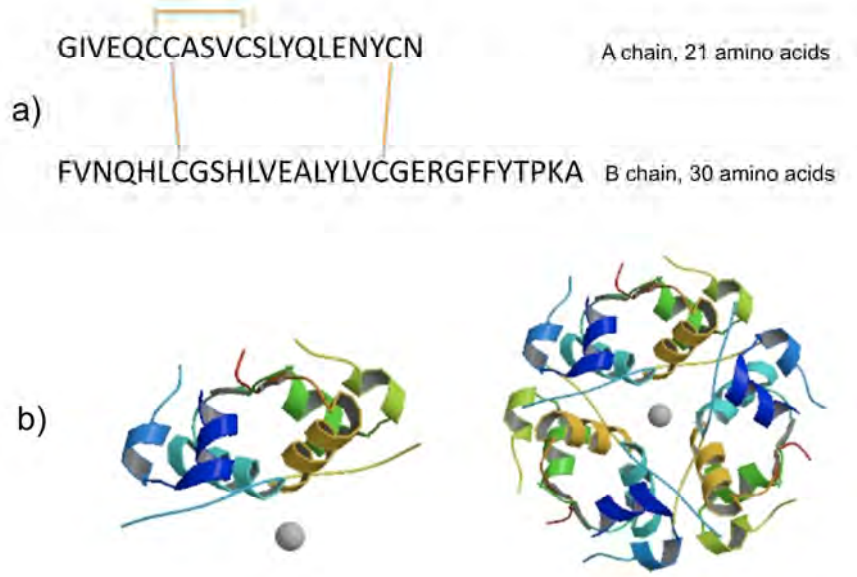

Fig. 1. Illustration of a) the primary structure of bovine insulin, showing intra- and interchain disulphide bonds connecting the a and b chains; and b) its zinc-coordinated tertiary structure (2INS), revealing two molecules in the asymmetric unit, and a hexameric biological assembly.

years to solve. Viewed in the light of the high-throughput sequence and structure determinations of today, these prolonged time-scales now seem almost inconceivable.
Notwithstanding the challenges, however, the potential of peptide sequencing technology to aid our understanding of the biochemical functions and evolutionary histories of particular proteins, and to facilitate their structural analysis, was compelling. Consequently, the sequences of many other proteins were soon deduced. In the early '60s, amongst the first to appreciate the value of biological sequences, and particularly the ability to deduce evolutionary relationships from them, was Margaret Dayhoff. To facilitate her research and the work of others in the field, she began to collect all protein sequences then available, ultimately publishing them in book form – this was the first *Atlas of Protein Sequence and Structure* (Dayhoff *et al*., 1965), often simply referred to as the *Atlas*. It may seem amusing to us now, but in a letter she wrote in 1967, she observed, "*There is a tremendous amount of information regarding the evolutionary history and biochemical function implicit in each sequence and* ***the number of known sequences is growing explosively*** [our emphasis]. *We feel it is important to collect this significant information, correlate it into a unified* whole *and interpret it*" (Dayhoff, 1967; Strasser, 2008). With the publication of the first *Atlas*, that 'explosive growth' amounted to 65 sequences!

In the decade that followed, time-consuming manual processes were gradually superseded with the advent of automated peptide sequencers, which increased the rate of sequence determination considerably. Meanwhile, another revolution was taking place, heralded by the elucidation of the 3D structures of the first proteins, those of myoglobin and haemoglobin, respectively (Kendrew *et al*., 1958; Muirhead and Perutz, 1963). Building on the ongoing sequencing work, this advance set the scene for an exciting new era in which structure determination took centre stage in our quest to understand the biophysical mechanisms that underpin biochemical and evolutionary processes. In fact, so seductive was this approach that many more structural studies were initiated, and the numbers of deduced protein structures grew accordingly.

3. The development and spread of databases, organisations and infrastructures

Key to handling this burgeoning information was the recruitment of computers to help systematically analyse and store the accumulating sequence and structure data. At this time, the idea that molecular information could be collected within, and distributed from, electronic repositories was not only very new but also posed significant challenges. Just consider, for a moment, that concepts we take for granted today (email, the Internet, the World Wide Web) had not yet emerged; there was therefore no easy way to distribute data from a central database, other than by posting computer tapes and disks to individual users, at their request. This model of data distribution was clearly rather cumbersome and slow; it was also relatively costly, and led some of the first database pioneers to adopt pricing and/or data-sharing policies that threatened to drive away many of their potential users.

3.1 The Protein Data Bank (PDB)

One of the earliest, and hence now oldest, of scientific databases was established in 1965 at the Cambridge Crystallographic Data Centre (CCDC), under the direction of Olga Kennard (Kennard *et al*., 1972; Allen *et al*., 1991) – this was a repository of small-molecule crystal structures termed the Cambridge Structural Database, or CSD. The CSD, which originated as a traditional printed dissemination, ultimately assumed an electronic form so that Kennard could fulfill a dream, which she shared with J.D.Bernal, to be able to use data collections to discover new knowledge, above and beyond the results yielded by individual experiments (Kennard, 1997).

In 1971, a few years after the creation of the CSD, at a Cold Spring Harbor Symposium on the "*Structure and Function of Proteins at the Three Dimensional Level*", Walter Hamilton and colleagues discussed the possibility of creating a similar kind of 'bank' for protein coordinate data. Key to their proposal was that this archive should be mirrored at sites in the UK and the USA (Berman, 2008). Consequently, Hamilton volunteered to set up the 'master copy' of the American bank at the Brookhaven National Laboratory (BNL), while Kennard subsequently agreed to host the European copy and to extend the CCDC small molecule format to accommodate protein structural data (Kennard *et al*., 1972; Meyer, 1997). Thus was born the Protein Data Bank (PDB); this was to be operated jointly by the CCDC and BNL, and where possible, distributed on magnetic tape in machine-readable form. News of its establishment was announced in a short bulletin in October that year (Protein Data Bank, 1971); its first release held 7 structures (Berman *et al*., 2000). Interestingly, Kennard viewed the PDB as a prototype for the EMBL data library, which was to materialise a decade later (Smith, 1990).

By 1973, the PDB was fully operational (Protein Data Bank, 1973). In August that year, the body of data it had been established to store amounted to 9 structures (see Table 1). Kennard and co-workers knew that the success of the resource was ultimately dependent on the support of the crystallography community in providing their data; but gaining sufficient community momentum to back the initiative was clearly a long, drawn-out process: note, for example, that the structure of ribonuclease, which had been determined 6 years earlier, was not yet listed amongst its holdings.

	Protein structures
1	Cyanide methaemoglobin V from sea lamprey
2	Cytochrome b_5
3	Basic pancreatic trypsin inhibitor
4	Subtilisin BPN (Novo)
5	Tosyl α-chymotrypsin
6	Bovine carboxypeptidase Aα
7	L-Lactate dehydrogenase
8	Myoglobin
9	Rubredoxin

Table 1. PDB holdings, August 1973.

Over the next 4 years, the number of structures acquired by the PDB grew slowly. By 1977, the archive also included the structure of a transfer RNA (tRNA), and hence the name *Protein Data Bank* was thought something of a misnomer (Bernstein *et al.*, 1977). Nevertheless, despite this reservation, the name stuck, and the resource (which today includes more than 5,000 nucleic acid and protein-nucleic acid complexes) is still referred to as the PDB. Interestingly, at that time, the database contained 77 sets of atomic coordinates relating to 47 macromolecules, highlighting a significant level of redundancy. Coupled with their ongoing concerns about the pace of growth of the archive, perhaps this explains why the Berstein *et al.* paper was published verbatim in May and November of 1977, and again in January 1978, in three different journals (Bernstein *et al.*, 1977a, b; 1978)? Whatever the real reasons, growth of the PDB compared to the CSD (~6,000 vs. ~150,000 structures in 1996) was slow (Kennard, 1997), and the number of unique structures remained relatively small – by 1992, the level of redundancy in the resource had been calculated to be ~7-fold (Berman, 2008; Hobohm *et al.*, 1992).

In 1996, shortly after the establishment of the European Bioinformatics Institute (EBI) near Cambridge, UK, a new database of macromolecular structures was created – this was the E-MSD (Boutselakis *et al.*, 2003). Building directly on PDB data, E-MSD was originally conceived as a pilot study to explore the feasibility of exploiting relational database technologies to manage structural data more effectively. In the end, the pilot project led to the creation of a database that was successful in its own right, and the E-MSD thereby became established as a major EBI resource.

During this period, a concerted effort was made to hasten the pace of knowledge acquisition from structural studies. Part of the motivation was to build on the still-limited number of structures available in the PDB, and partly also to address its growing level of redundancy. The idea was to establish a program of high-throughput X-ray crystallography – the so-called Structural Genomics Initiative (SGI) (Burley *et al.*, 1999). Several feasibility studies had

already been launched and, in light of the broad-sweeping vision of the SGI, it had become clear that coping with high-throughput structure-determination pipelines would require new ways of gathering, storing, distributing and 'serving' the data to end users. One of the PDB's responses to this, and to the many challenges that lay ahead, was the formation of a new management structure. This was to be embodied in a 3-membered Research Collaboratory for Structural Bioinformatics (RCSB): the consortium included Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California; and the Center for Advanced Research in Biotechnology of the National Institute of Standards and Technology (Berman *et al*., 2000; Berman *et al*., 2003). Once the consortium was established, the BNL PDB ceased operations and the RCSB formally took the helm on 1 July, 1999.

With the RCSB PDB in the USA, the E-MSD established in Europe, and a sister resource (PDBj) subsequently announced in Japan (Nakamura *et al*., 2002), structure collection efforts had clearly taken on an international dimension. In consequence, in 2003, the 3 repositories were brought together beneath an umbrella organisation known as the worldwide Protein Data Bank (wwPDB), to streamline their activities and maintain a single, global, publicly available archive of macromolecular structural data (Berman *et al*., 2003). By 2009, perhaps to align its nomenclature in a more obvious way with its consortium partners, E-MSD was renamed PDBe (Velankar *et al*., 2009). Today, the RCSB remains the 'archive keeper', with sole write-access to the PDB, controlling its contents, and distributing new PDB identifiers to all deposition sites. In February 2011, the archive housed 71,415 structures.

3.2 The EMBL nucleotide sequence data library

Despite the advances in protein sequence- and structure-determination technologies between the mid-1940s and -'70s, sequencing nucleic acids had remained problematic. The key issues related to size and ease of molecular purification. It had proved possible to sequence tRNAs, largely because they're short (typically less than 100 nucleotides long) and individual molecules could, with some effort, be purified; but chromosomal DNA molecules are in a different league, containing many millions of nucleotides. Even if such molecules could be broken down into smaller chunks, purification was a major challenge. The longest fragment that could then be sequenced in a single experiment was ~500bp; and yields of potentially around half a million fragments per chromosome were simply beyond the technology of the day to handle.

During the mid '70s, however, Sanger had developed a technology (to become known as the 'Sanger method') that made it possible to work with much longer nucleotide fragments: this allowed completion of the sequencing of the 5,386 bases of the single-stranded bacteriophage ϕX174 (Sanger *et al*., 1978), subsequently permitting rapid and accurate sequencing of even longer sequences – an achievement of sufficient magnitude to earn him his second Nobel Prize in Chemistry, in 1980. With this technique, he went on to sequence human mitochondrial DNA (Anderson *et al*., 1981) and bacteriophage λ (Sanger *et al*., 1982). These were landmark achievements (see Table 2), providing the first direct evidence of the phenomenon of overlapping gene sequences and of the non-universality of the genetic code (Sanger, 1988; Dodson, 2005). But it was automation of these techniques from the mid-'80s that significantly increased productivity, and began to make the human genome a realistic target.

Together, these advances prepared the way for a new revolution, one that would rock the foundations of molecular biology and make the gathered fruits of all sequencing efforts

before it appear utterly inconsequential. Here, then, was a dramatic turning point: for the first time, it dawned on scientists that the new sequencing machines were shunting the bottlenecks away from data production *per se* and onto the requirements of data management: "*the rate limiting step in the process of nucleic acid sequencing is now shifting from data acquisition towards the organization and analysis of that data*" (Gingeras & Roberts, 1980). This realisation had profound consequences in both Europe and the USA, as a centralised data bank now seemed inescapable as a tool for managing nucleic acid sequence information efficiently.

Year	Protein	RNA	DNA	No. of residues
1935	Insulin			1
1945	Insulin			2
1947	Gramicidin S			5
1949	Insulin			9
1955	Insulin			51
1960	Ribonuclease			120
1965		$tRNA_{Ala}$		75
1967		5S RNA		120
1968			Bacteriophage λ	12
1977			Bacteriophage ϕX 174	5,375
1978			Bacteriophage ϕX 174	5,386
1981			Mitochondria	16,569
1982			Bacteriophage λ	48,502
1984			Epstein-Barr virus	172,282
2004			*Homo sapiens*	2.85 billion

Table 2. Sequencing landmarks.

So, the race was on to establish the first nucleotide sequence database. First past the post, in 1980, was the European Molecular Biology Laboratory (EMBL) in Heidelberg, who set up the EMBL data library. After an initial pilot period, the first release of 568 sequences was made in June 1982. The aim of this new resource was not only to make nucleic acid sequence data publicly available and encourage standardisation and free exchange of data, but also to provide a European focus for computational and biological data services (Hamm & Cameron, 1986).

From the outset, it was recognised that maintenance of such a centralised repository, and of its attendant services, would require international collaboration. In the UK, a copy of the EMBL library was being maintained at Cambridge University, together with its manual, indices and associated sequence analysis, and search and retrieval software. This integrated system also provided access to the library of sequences then being developed at Los Alamos, GenBank (Kanehisa *et al.*, 1984). It makes fascinating reading to learn that, "*this system is presently being used by over 30 researchers in eight departments in the University and in local research institutes. These users can keep in touch with each other via the MAIL command*"! With the support of the Medical Research Council (MRC), the Cambridge services were extended to the wider UK community on the Joint Academic network (JANET) (Kneale & Kennard, 1984). As with the PDB before it, it was important not only to push the data out to researchers, but also to pull their data in. Hence, a further planned development was to

centralise collection of nucleic acid data from UK research groups, and to periodically transfer the information to the EMBL library. It was hoped that this would minimise both data-entry errors and the workload of EMBL staff at a time when the number of sequence determinations was predicted to "*increase greatly*" (Kneale & Kennard, 1984). Of course, the size of this 'great increase' could hardly have been predicted; in December 2010, the database contained 199,720,869 entries.

3.3 GenBank

The birth of GenBank, in December 1982, brought 606 sequences into the public domain. A consensus had emerged on the necessity of creating an international nucleic acid sequence repository at a scientific meeting at Rockefeller University in New York, in March 1979. At that time, several groups had expressed a desire to be a part of this endeavour, including those led by Dayhoff at the National Biomedical Research Foundation (NBRF); Walter Goad at Los Alamos National Laboratories; Doug Brutlag at Stanford; Olga Kennard and Fred Sanger at the MRC Laboratory in Cambridge; and Ken Murray and Hans Lehrach at the EMBL (Smith, 1990), all of whom had begun to create their own nucleotide sequence collections. However, it took the best part of 3 years for an appropriate funding model to emerge from the US National Institutes of Health (NIH), by which time the EMBL data library had already been publicly available for 6 months under the direction of Greg Hamm. By then, 3 proposals remained on the table for NIH support: 2 of these were from Los Alamos (one with Bolt, Beranek and Newman (BBN), the other with IntelliGenetics), and the third from NBRF. To the surprise of many, the decision was made in June 1982 to establish the new GenBank resource at Los Alamos (in collaboration with BBN, Inc.) rather than at the NBRF (Smith, 1990; Strasser, 2008).

Although there was a general sense of relief that a decision had finally been made, some members of the community (and doubtless Dayhoff herself) felt that the NBRF would have been a more appropriate home for GenBank, particularly given Dayhoff's successful track record as a curator of protein sequence data (Smith, 1990). Los Alamos, by contrast, although undoubtedly offering excellent computer facilities, was probably best known for its role in the creation of atomic weapons – this was not an obvious environment in which to establish the nation's first public nucleotide sequence database. The crux of the matter seemed to rest with the different philosophical approaches embodied in the NBRF and Los Alamos proposals, particularly as they related to scientific priority, data sharing/privacy and intellectual property policies. Dayhoff had intended to continue gathering sequences directly from literature sources and from bench scientists, and wasn't interested in matters of history or priority (Eck & Dayhoff, 1966); the Los Alamos team, on the other hand, advocated the collaboration of journal editors in making the publication of articles contingent on authors yielding their sequence data to the database. This latter approach was particularly compelling, as it would allow scientists to assert priority, and to keep their research results private until formally published and their provenance established; perhaps more importantly, it was unencumbered by proprietary interest in the data. Unfortunately, the fact that Dayhoff had prevented redistribution of NBRF's protein sequence library and sought revenues from its sales (albeit only to cover costs) worked against her – allowing the data to become the private hunting grounds of any one group of researchers was considered antithetical to the spirit of open access (Strasser, 2008). That the data and associated software tools should be free and open was thus paramount; it is perhaps ironic, then, that the site chosen for the database was within the secured area of what many in the community may have darkly perceived as 'The Atomic City' (en.wikipedia.org/wiki/The_Atomic_City).

As an aside, it's interesting that the vision of free data and programs was advocated so strongly at this time, not least because there was no funding model to support it! And precisely the same arguments are still being vehemently propounded today with regard to free databases, free software and free literature (*e.g.*, Lathrop *et al*., 2011). But even now, database funding remains an unsolved and controversial issue: as Olga Kennard put it almost 15 years ago, "*Free access to validated and enhanced data worldwide is a beautiful dream. The reality, however, is more complex*" (Kennard, 1997).

Returning to our theme, perhaps the final nail in the coffin of Dayhoff's proposal was that the NBRF had only limited means of data distribution (via modems), whereas the Los Alamos outfit had the enormous benefit of being able to distribute their data via ARPANET, the computer network of the US Department of Defense. Together, these advantages were sufficient to swing the pendulum in favour of the Los Alamos team.

But the new GenBank did not, indeed could not, function in isolation. From its inception, it evolved in close collaboration with the EMBL data library and, from 1986 onwards, also with the DNA Data Bank of Japan. Although the databases were not identical (each with its own format, naming convention, and so on), the teams adopted common data-entry standards and data-exchange protocols in order to improve data quality and to manage both the growth of the resource and the annotation of its entries more effectively. Of this collaborative process, Temple Smith commented in 1990, "*By working out a division of labor with the EMBL and newer Japanese database efforts, and by involving the authors and journal editors, GenBank and the EMBL databases are currently keeping pace with the literature*." Today, the boot seems to be very much on the other foot, as the literature can no longer keep up with the data: by February 2011, GenBank contained 132,015,054 entries, presenting insurmountable annotation hurdles! (Note that this appears smaller than the size of the EMBL data library because GenBank doesn't report sequences from Whole Genome Shotgun projects in its total). Perhaps not surprisingly, the initial funding for GenBank was insufficient to adequately maintain this growing mass of data; hence, responsibility for its maintenance, with increased funding under a new contract, passed to IntelliGenetics in 1987; then, in 1992, it became the responsibility of the NCBI, where it remains today (Benson *et al*., 1993; Smith, 1990).

3.4 The PIR-PSD

To some extent, the gathering momentum of nucleic acid sequence-collection efforts had begun to overshadow the steady progress being made in the world of protein sequences, most notably with the *Atlas*. By October 1981, this had run into its fifth volume, a large book with three supplements, listing more than 1,660 proteins. This information, as with all data collections, required constant updating and revision in the light both of new knowledge and of new data appearing in the literature. Moreover, as the community had become increasingly keen to harness the efficiency gains of central data repositories, and more databases were appearing on the horizon, making and maintaining cross-references to database entries, of necessity, had to become part of data-annotation and update processes if scientists were to be able to exploit new and existing sequence data fully. Under the circumstances, continued publication of the *Atlas* in paper form simply became untenable: the time was ripe to exploit the advances in computer technology that had given rise to the CSD, the PDB, the EMBL data library and GenBank. In 1984, the *Atlas* was consequently made available on computer tape as the Protein Sequence Database (PSD).

Later, in 1986, in order to facilitate protein sequence analysis more broadly, the NBRF established the Protein Identification Resource (PIR) (George *et al*., 1986). This new online system included the PSD, several bespoke query and analysis tools (*e.g.*, the Protein Sequence Query (PSQ), SEARCH and ALIGN programs), and a new, efficient search program, FASTP. The latter was a modification of an earlier algorithm for searching protein and nucleic acid sequences (Wilbur & Lipman, 1983). Interestingly, given that the number of deduced sequences had, by that time, grown into the thousands, the great advantage of Wilbur and Lipman's method was considered to be its speed. Indeed, their paper reported a "*substantial reduction in the time required to search a data bank*". Improving on this even further, the new FASTP algorithm was able to compare a 200-amino-acid sequence to the 2,677 sequences of the PSD in "*less than 2 minutes on a minicomputer, and less than 10 minutes on a microcomputer (IBM PC)*" (Lipman & Pearson, 1985). Looking back, such search times on such small numbers of sequences seem incredibly slow; at the time (when a contemporary algorithm required 8 hours for the same search), they were revolutionary.

As the PIR was built on NBRF's existing resources, it also made available its DNA databank (Dayhoff *et al*., 1981a) and associated software tools, together with copies of GenBank and the EMBL data library; it also retained the NBRF's cost-recovery model, levying a charge for copies of its databases on magnetic tape and an annual subscription fee for use of its online services – in 1988, these amounted to $200 per tape release and $350 per annum respectively (Dayhoff *et al*., 1981b; Sidman *et al*., 1988). By 1992, the PSD had shown steady growth, with increasing contributions from European and Asian protein sequence centres – most notably, from MIPS (Martinsried, Germany) and from JIPID (Tokyo, Japan). Accordingly, a tripartite collaboration was established, termed PIR-International, to formalise these relationships and establish and disseminate a comprehensive set of protein sequences (Barker *et al*., 1992). By this time, charging for access to the resource was no longer mentioned, possibly both as a consequence of this more formal distribution arrangement and the advent of browsers like Mosaic, which had suddenly and dramatically changed the way that information could be broadcast and received over the World Wide Web (or, simply, the Web). In 1997 PIR changed its name to the Protein Information Resource (George *et al*., 1997) and, by 2003, with 283,000 sequences (Wu *et al*., 2003), the PSD was the most comprehensive protein sequence database in the world.

3.5 Swiss-prot

While these events were taking place, a newly qualified Swiss student (who, as a teenager, had been interested in space exploration and the search for extraterrestrial life) attempted to embark on a Masters project involving both 'wet' and 'dry' work – this was Amos Bairoch. The experimental side of his project immediately hit problems when it was discovered that the new mass spectrometer he was to have used didn't work properly. He therefore set to work instead developing protein sequence analysis programs on the computer system running the spectrometer. These were the first steps towards creating the software system that was later to be known as PC/Gene, and was to become the most widely used PC-based sequence analysis package of its day (Bairoch, 2000).

Part of what made this software suite unique was its focus on proteins at a time when the analysis of nucleotide sequences was very much in vogue. In creating these tools, Bairoch entered >1,000 protein sequences into his computer by hand: some of these he gleaned from

the literature; most were taken from the *Atlas*, which had not yet been released in electronic form. Of course, this was an immensely tedious process, and was also highly error-prone. Realising this, and anxious to avoid such problems for others in future, he wrote a letter to the *Biochemical Journal* recommending that researchers publishing protein and peptide sequences should compute checksums to "*facilitate the detection of typographical and keyboard errors*" (Bairoch, 1982). As part of the letter, he illustrated the computation of such a 'checking number' for an imaginary peptide, as shown in Figure 2. Although this recommendation was never widely adopted in publishing circles, Bairoch was at least able to ensure that it was implemented in his own database.

Peptide: H E L P I H A T E M A T H

$$CN \text{ computation: } CN = 1\cdot9+2\cdot7+3\cdot11+4\cdot15+5\cdot10+6\cdot9+7\cdot1+8\cdot17+9\cdot7+10\cdot13+11\cdot1+12\cdot17+13\cdot9 = 788$$

$$COMP = A_2R_0N_0D_0C_0Q_0E_2G_0H_3I_1L_1K_0M_1F_0P_1S_0T_2W_0Y_0V_0$$

$$NR = 13 \qquad MMP = 1186.66 \qquad CN = 788$$

Fig. 2. Computation of a 'checking number' (CN) for an imaginary peptide, as published in a letter to the *Biochemical Journal* in 1982. The journal editors either didn't notice, or chose to ignore, the hidden message in the peptide. Reproduced with permission, from Bairoch, A. (1982), *Biochemical Journal*, 203, 527-528. © the Biochemical Society

Several other important developments were to emerge from the work of this enthusiastic and industrious student. For the analysis software he was developing, he needed to distribute both a nucleotide and a protein sequence database. In 1983, he acquired a computer tape containing 811 sequences in version 2 of the EMBL data library; for his protein sequence database, he initially used the sequences he'd typed in for his Masters project. However, the following year, he received the first electronic copy of the *Atlas*. He was quick to appreciate the advantages and disadvantages of the PIR and EMBL formats, recognising that converting the manually annotated data of the former into something like the semi-structured format of the latter could produce a resource with the strengths of both – he called this PIR+ and released it side-by-side with his software package, PC/Gene, which by that time he'd commercialised through IntelliGenetics (Bairoch, 2000).

Use of the publicly available PIR data-set in this way was not without its problems. Amongst other, deeper, issues were the difficulty of parsing PIR files to extract specific information (*e.g.*, relating to post-translational modifications (PTMs), *etc.*); the lack of functional annotations for some of the newer entries; the lack of cross-referencing to the parent DNA of a given protein sequence; and so on. Somewhat ironically, given what he went on to achieve, Bairoch has written of this period, "*As I was not interested in building up databases I kept sending letters to PIR to ask them to remedy this situation*". But his pleas met with little success. In the summer of 1986, in the face of increasing demand for unencumbered access to his database, he decided to release PIR+ independently of PC/Gene, to make it freely available to the entire research community. The new, public version of the database was released on 21 July 1986 and contained ~3,900 sequences (the exact number is unknown as the original floppy disks have been lost!) This new resource

was called Swiss-Prot (Bairoch & Boeckmann, 1991), and was to become the foremost manually annotated database of protein sequences in the world.

3.6 The European Molecular Biology Network (EMBnet)

It is interesting that, during this era, the distribution of databases like the EMBL data library, PIR, Swiss-Prot and so on, was still largely effected by the exchange of computer tapes and disks. By this time, a variety of computer networks had begun to evolve: the first such network, ARPANET (which began life with 4 nodes in late 1969), was the progenitor of the Internet, and was superseded by it in 1983 – recall, it was partly owing to the existence of ARPANET that GenBank was established at Los Alamos. Other networks that offered gateways into the Internet later merged with it, including Usenet and BITNET; commercial and educational networks, such as Telenet (or Sprintnet), Tymnet, Compuserve and JANET, were interconnected with it in the 1980s.

In 1988, Chris Sander at the EMBL helped to establish a new network, EMBnet, to disseminate data, knowledge and services, to support and advance molecular biology and biotechnology research across Europe. A major driver for creating EMBnet was the need for local access to databases such as the EMBL data library from centralised sources. Essentially, this is because scientists were now demanding to use client workstations with Graphical User Interfaces (GUIs) that provided real-time interaction with their back-end data/analysis servers. At the time, high-speed data communication across Europe was in its infancy, and access to remote computers using ordinary command-line oriented terminals was too slow. It was clear that communication delays could be eliminated if servers held copies of data locally; the sheer amount of compute resources needed for European research in this field also pointed to a distributed solution (note that computer cluster technology only gained widespread acceptance much later). Thus, an organised way of distributing data and resources from the EMBL to its member states had to be established.

The concept of a network of national 'nodes', each serving its country with up-to-date biological databases and also providing compute resources for data analysis, was formulated. It was given the name the European Molecular Biology network, EMBnet. The first practical steps were taken in the spring of 1988 to solicit feedback from scientists around Europe; and in July 1988, the first EMBnet Workshop was organised at EMBL, with participants from EMBL, Daresbury (UK), CITI2 (France), the CAOS/CAMM Centre (the Netherlands) and Hoffmann-La Roche. In November of that year, the EMBL Director General corresponded with EMBL Council members, encouraging them to stimulate local processes to identify regional EMBnet nodes. As more countries joined the network (France, Sweden, the UK, the Netherlands, Spain, Israel, Norway, Italy and Denmark, with Switzerland, West Germany, Austria, Greece and Finland waiting in the wings), EMBnet received its first European grant under the BRIDGE framework, in 1991.

The principal project objective was to promote EMBnet as a computer network for European bioinformatics. Service provision and knowledge sharing was to be orchestrated primarily by 'National Nodes', with government mandates to support their local communities, especially by providing access to bioinformatics data synchronised with the EMBL, GenBank and DDBJ central data repositories – in time, the network also attracted a number of 'Specialist' and 'Industrial' Nodes, whose resources and know-how were seen to complement those of its National Nodes (this arrangement of cooperating Nodes is illustrated in Figure 3).

Most EMBnet Nodes had VAX computers, and the original intention was to use DECNET as the underlying transport protocol. However, after a short, but expensive, period of using

X25/DataPak, this was replaced by a TCP/IP-package called MultiNet, which was licensed for all EMBnet Nodes from SRI (Stanford Research Institute). FTP-transmissions of database updates were often interrupted by network problems, and, to overcome the need for frequent re-transmissions, the NDT (Network Data Transfer, later xNDT for extended NDT) protocol was developed at the Swedish EMBnet Node at Uppsala Biomedical Centre, by Peter Gad. It was given a so-called 'systems well-known port' (embl-ndt, 394/udp,# EMBL Nucleic Data Transfer) by the Internet authorities, and is thus in good company with, for example, Telnet (port 23) and FTP (ports 20, 21). For a few years, (x)NDT, and its accompanying suite of client-server programs, was the method par preference, used at almost all EMBnet Nodes to keep their local databases updated. NDT took care of the transmission (database) entry by entry and didn't have to re-start following network interruptions. The Greek node, situated in Crete, only had a modem connection to the mainland, and benefited hugely from using the xNDT-suite. Indeed, at the time the European Bioinformatics Institute was established (when the EMBL Data Library moved from Heidelberg to Cambridge), most of the nucleotide sequence database update traffic in Europe was routed via the Swedish node using xNDT.

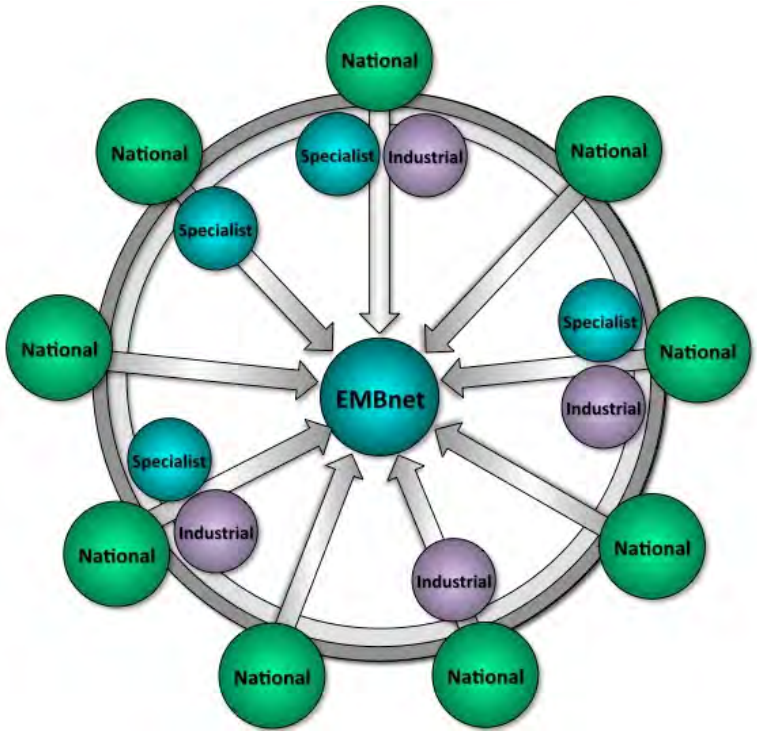

Fig. 3. Illustration of the relationship between the different Nodes of the early EMBnet: some National Nodes had either Specialist or Industrial Nodes affiliated with them; some had both; some had neither. Today, 31 National and Specialist Nodes contribute to the Network.

By the early '90s, biomolecular databases could be accessed across the Internet by means of the WAIS and Gopher network retrieval systems; and, under the auspices of EMBnet, Reinhard Doetlz had developed a new network access protocol, HASSLE, the Hierarchical Access System for Sequence Libraries in Europe (Doelz, 1994). But it was the advent of graphical Web browsers (first, Mosaic from the National Center for Supercomputing Applications in 1993, and then Netscape Navigator in 1994) that revolutionised the processes of database dissemination and information consumption – literally, at the click of a mouse button.

Of course, browsers allowed data and documents of all kinds to be instantly shared, and individuals and organisations across the globe were quick to establish their own unique 'Web presence'. EMBnet was no exception, and embraced the Web as a means of communicating more effectively with its widening community, in particular by publishing a regular newsletter, *EMBnet.news.* The newsletter was designed to provide reports and updates on its internal and international activities and achievements, together with technical and scientific papers on new developments in bioinformatics, computational biology and biocomputing. In 2000, the organisation provided an educational grant to help support the creation of the peer-reviewed journal *Briefings in Bioinformatics* (BiB) and, as a mark of its own success, *EMBnet.news* is also now in the process of transitioning to a peer-reviewed journal.

From the outset, EMBnet has promoted the development of distributed computing services to share workload among international servers; it has contributed to the development and maintenance of advanced database systems; it has been an advocate of the deployment of Grid technologies for the life sciences through its contributions to major European Grid projects; it developed, and continues to promote the use of, an e-learning system both to support distance learning in bioinformatics and to complement face-to-face bioinformatics teaching and training; and it is committed to bringing the latest software and algorithms to users, free of charge.

The combined expertise of its Nodes has allowed EMBnet to provide services to its local European life science communities with far greater effect than could be achieved by any of its individual Nodes in isolation. Following this success, a variety of Nodes world-wide have joined EMBnet such that, today, the network is global, with many countries from Asia, Africa and America joining in recent years (including Sri Lanka, Pakistan, Kenya and Costa Rica). Currently, the network connects 31 member Nodes extending over 27 countries; together, the Nodes continue to work to disseminate data, to share compute resources and to provide training support, reaching out to many thousands of users.

3.7 PROSITE

While EMBnet was being conceived, before the Internet had truly taken off, and while bioinformatics was still in the throes of been born, the computer savvy molecular biologists of the day were still busily swapping biomolecular databases on magnetic tapes and computer disks. Perhaps an inevitable consequence of the systematic collection of protein and nucleotide sequences in this way was the need to organise and classify these molecular entities in meaningful ways. The first endeavour to categorise protein sequences into evolutionarily related families, and to provide the diagnostic means to detect potential new family members, arose once again as a derivative of the PC/Gene suite. Inspired by the sequence analysis primer, *Of URFs and ORFs* (Doolittle, 1986), Bairoch began to amass examples of short sequences, characteristic of particular binding and active sites, and

developed a program to scan his growing collection of sequence 'patterns'. This part-program, part-database chimera he named PROSITE (Bairoch, 1991). In March 1988, as part of PC/Gene, the first release of this new resource contained 58 entries.

As with Swiss-Prot before it, PROSITE swiftly gained popularity. Its growing band of users began not only to suggest additional patterns that could be included in the database, but also to pressure Bairoch into giving PROSITE an independent life of its own, outside PC/Gene. Consequently, the availability of a new public version was announced in October 1989, and formally released the following month with 202 entries (version 4.0). Diagnostically, it was clear that sequence patterns had certain limitations. In particular, matching a pattern is a binary 'match/no-match' event: even the most trivial difference (a single amino acid) results in a mis-match. As Swiss-Prot expanded and accommodated more and more divergent members of its various superfamilies, the more evident this particular weakness became. One solution to this problem emerged in the form of position-specific weight matrices, or profiles. Built from comprehensive sequence alignments, profiles are tolerant both of amino acid substitutions and of insertions/deletions; they therefore allow the relationships between families of sequences to be modelled more 'realistically'. Accordingly, with the help of Phillipp Bucher, Bairoch began to augment PROSITE with sequence profiles – the first release to include them came with version 12.0, in June 1994 (Bairoch & Bucher, 1994).

Another solution, which arose (at least methodologically) independently from PROSITE, was the development of protein family 'fingerprints'. Fingerprints are groups of conserved motifs, evident in multiple sequence alignments, whose unique inter-relationships provide distinctive signatures for particular protein families and structural/functional domains. They are diagnostically more powerful and flexible than patterns, because they can tolerate mis-matches at the level both of individual motifs and of the fingerprint as a whole. Fingerprints formed the basis of a database that began life as the Features Database, part of the SERPENT information storage and analysis resource for protein sequences established at the University of Leeds (Akrigg *et al*., 1992). Its first release, in October 1991, contained 29 entries: two thirds of these were linked to equivalent entries in PROSITE, which by then held 441 family descriptions.

Although disparate in size, the Features and PROSITE databases had various aspects in common; most notable amongst these was the principle of added-value through hand-crafted annotation of their diagnostic signatures. In March 1991, Bairoch met Terri Attwood for the first time at the British Crystallographic Association spring meeting in Sheffield. Faced with the same, relentlessly time-consuming, manual-annotation burdens, they shared their woes and discussed the wisdom of unifying the PROSITE and Features databases. Motivated by common ideals, they later formalised their ideas in the guise of their first European grant proposal to merge their databases into an integrated protein family annotation resource. This was 1992; they were not successful.

In the meantime, inspired by PROSITE, a range of other signature databases began to emerge. One of the earliest of these was Blocks, first described by Steve and Jorja Henikoff in December 1991 (Henikoff & Henikoff, 1991). Later came ProDom (Sonnhammer & Kahn, 1994), and later still Pfam (Sonnhammer *et al*., 1997). Initially linked closely to the annotation of predicted proteins from genomic sequencing of *Caenorhabditis elegans*, Pfam was to become one of the most widely used protein family databases across Europe and the USA.

3.8 The European Bioinformatics Institute (EBI)

Notwithstanding the proliferation of databases in the '80s, funding for their maintenance was becoming a significant problem. By the early '90s, supporting the EMBL data library was becoming increasingly difficult, and there was growing awareness that a more efficient European bioinformatics infrastructure would be needed to sustain it in future. In 1992, the EMBL concluded that the most robust solution would be to establish a new outstation, devoted to bioinformatics. The vision of creating a European Bioinformatics Institute (EBI) quickly took hold and, in December that year, the EMBL Governing Council published a call for proposals to host the new facility. The deadline was extremely short (February 1993); despite the interest of many countries, therefore, few were able to submit bids in time.

In a study by PA Consulting Group, commissioned by the EC's DGXII, a plan had been developed for a European Nucleotide Sequence Centre (ENSC). The EMBL Council decided to "*negotiate with the EC for the inclusion of the ENSC within the EBI*"; the EBI "*would provide bioinformatics services for European scientists, be a home for the Data Library, and include expansions in research and development necessary for long-term viability and strengthening of neglected areas such as user support*"(Philipson, 1992).

In EMBL's proposal for an EBI from October 1992, worries were expressed that Europe was lagging behind the USA: "*Over the last decade increments in US support for such resources have far outstripped those in Europe*," and the EBI was conceived "*to ensure that European research needs are satisfied in a way which is appropriate to this global competitive context*" (EMBL, 1992). The need for supportive relations between EBI and the European scientific community was emphasised, as "*It would be impossible and undesirable for the EBI to be the sole bioinformatics resource in Europe*". It was noted that support should be given to "*major European interest groups such as software developers, database hosts and other bioinformatics institutes*"; more specifically, "*In recognition of the need for strong national bioinformatics activities, the EBI will give technical and organisational support to the EMBnet Nodes, as is currently done by the EMBL Data Library*"(EMBL, 1992).

Among the bidders for the EBI were Germany, Sweden and the UK. Very favourable conditions were offered from all three. The Swedish bid for an EBI close to Uppsala Biomedical Centre, included, for example, sufficient office space, free of rent, and high-speed network connections. But Michael Ashburner led a more compelling UK bid. The proposal was to host the EBI on a park, newly purchased by the Wellcome Trust, at Hinxton, on the outskirts of Cambridge. The Trust and MRC had agreed each to fund half of the initial capital costs of creating a complete genomics infrastructure on this site, which would also include the newly established Sanger Centre (which, by then, had become embroiled in the HGP) and the Human Genome Mapping Project Resource Centre (Dickson & Abbott, 1993). With its "*clear commitment from all levels of the UK scientific community and Government*", the UK bid won over both Uppsala and the alternative location in Heidelberg, directly adjacent to the EMBL; it was accepted by Council in March 1993. Paulo Zanella (who had directed the CERN Data Handling Division) was subsequently appointed as EBI's first director (Bairoch, 2000).

The EBI became fully operational after completion of the new building in September 1995 – this will no doubt have come as a great relief to the EMBL data library group, who had been accommodated in portable cabins on the Hinxton site since the end of 1994! The new facility had 3 broad divisions: research, industry and services, the latter being mostly devoted to provision and maintenance of the EMBL data library and Swiss-Prot (Bairoch, 2000). The EBI's mission was to ensure that the growing corpus of data from molecular biology and

genome research was placed in the public domain and was freely accessible to the entire scientific community in order to promote scientific progress. Today, with its original 3-fold structure still largely in place, the Institute builds, maintains and disseminates databases and information services relevant to molecular biology, genetics, medicine and agriculture, and undertakes leading-edge research in bioinformatics and computational biology.
Despite its pivotal role as Europe's main bio-database provider, four years later, the EBI was in financial trouble. While the Wellcome Trust and MRC had financed the initial capital costs, the Institute relied on the EU for almost half its budget. In March 1999, however, the member states had advised the Commission that core funding and operational costs for infrastructure should not qualify for funding; the EBI's application for Framework funds was consequently rejected for being out of scope. Graham Cameron, by then joint Head of the Institute with Michael Ashburner, was quick to point out that without an immediate solution, "*we will have to abandon major projects like the DNA database, the draft human genome, the macromolecular structure database and the microarray expression database*" (Butler, 1999). The EBI was in a tricky situation, and Britain had shot itself in the foot: it could hardly contest the Commission's ruling against supporting the EBI because, a Commission official pointed out, "*it was among the countries most against funding infrastructure directly*" (Butler, 1999). The situation was neatly summed up in an editorial *Nature* ran at the time, "*If this Kafkaesque affair has any merit, it is that it has exposed the absence of a clear mechanism for the planning and support of research infrastructure at the European level*" (Nature Editorial, 1999). The cries for new mechanisms for infrastructural support, with stable partners, stable financing and long-term political commitment, doubtless helped to sew the seeds that in 2008 grew into the preparatory phase of ELIXIR, the European Life Science Infrastructure for Biological Information project.

3.9 Global data overload

The late '80s and early '90s were fertile years, giving rise to a flourishing number of new molecular structures and sequences, to new breeds of protein family signatures, and to new databases in which to store them. Looking back at this period of fervent activity, it's incredible to reflect that two major developments had yet to take place: together, these would not only seed an overwhelming explosion of biological data but would also spur their global dissemination – they were the advent of the Web and the arrival of high-throughput DNA sequencing. The latter made whole-genome sequencing practically feasible for the first time. Seizing this opportunity, there followed an unprecedented burst of sequencing activity, yielding, in quick succession, for example, the genomes of *Haemophilus influenzae* and *Mycoplasma genitalium* in 1995 (Fleischmann *et al*., 1995; Fraser *et al*., 1995), of *Methanococcus jannachii* and *Saccharomyces cerevisiae* in 1996 (Bult *et al*., 1996; Goffeau *et al*., 1996), of *Caenorhabditis elegans* in 1998 (*C.elegans* sequencing consortium, 1998), of *Drosophila melanogaster* in 2000 (Adams *et al*., 2000) and, the ultimate prize, of *Homo sapiens* in 2001 (Lander *et al*., 2001; Venter *et al*., 2001; IHGSC, 2004). Hundreds of genomes have been sequenced since this fruitful dawn.
Hand-in-hand with these activities came the development of numerous organism-specific databases to store the emerging genomic data: for example, FlyBase (Ashburner & Drysdale, 1994), ACeDB (Eeckman & Durbin, 1995), SGD (Cherry *et al*., 1998), TAIR (Huala *et al*., 2001), Ensembl (Hubbard *et al*., 2002), DictyBase (Kreppel *et al*., 2004) and, of course, many more. For some, the value of this genomic 'gold rush' was not entirely clear: with much of the amassed data seemingly impossible to characterise, and vast amounts of it non-coding, the hoped-for

treasure troves were beginning to look about as inspiring as large-scale collections of butterflies (Strasser, 2008), and perhaps suggested that molecular biology had entered a somewhat vacuous era of "*high-tech stamp collecting*" (Hunter, 2006). Arguments like this characterised some of the early opposition to the establishment of GenBank, and to the substantial resistance to the Human Genome Project (HGP) a few years later (Strasser, 2008).

Perhaps inevitably, then, the HGP was an extraordinarily high-profile affair. This was partly for the reasons outlined above, coupled with its considerable price-tag (estimated at $3 billion from 1990-2003), but in part also because of the public-private race between Francis Collins (who was directing the NIH National Human Genome Research Institute contributions to the HGP) and Craig Venter (then Head of Celera Genomics) to obtain the first rough draft of man's genetic blueprint. This intensely political 'drama' had been preceded by a similar struggle to be the first to sequence *Drosophila*, which served as a kind of 'warm up' battle for the human genome (Ashburner, 2006); it also had an intriguing parallel in the competition between two public-private corporations to sequence the genome of the commercially valuable *Agrobacterium tumefaciens* (Goodner *et al*., 2001; Wood *et al*., 2001; Harvey & McMeekin, 2004). The principal tension between these public and private, and public-private hybrid, enterprises arose not just from the race to be first to complete the sequencing: the struggle was as much about making the results public, on the one hand, and obtaining the property rights (for commercial exploitation, including gene patenting), on the other. Like the concerns in the early '80s surrounding NBRF's proprietary interest in protein sequences culled from the public domain, such conflicts raised serious questions about the duty of public science to ensure that genome sequences were made available for the public good; moreover, they challenged such wasteful competition, resulting in the acquisition of duplicate data-sets and, usually, back-to-back publications in high-profile journals (Harvey & McMeekin, 2004).

Another, more tangible, consequence of this intense orgy of genomic sequencing was the generation of more data than could realistically be managed and annotated by hand – and this was just the tip of an enormous future iceberg. As illustrated in Figure 4, with each

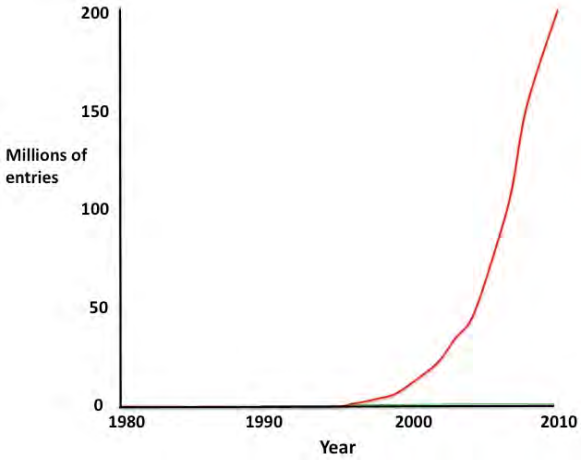

Fig. 4. Growth of the EMBL data library (millions of entries) since its inception (red curve). Also shown are the corresponding growth of the manually-annotated Swiss-Prot (green line), and of structures deposited in the PDB (this line is too small to be visible!).

passing year from the mid '90s, there was a widening gulf both between the volume of accumulating uncharacterised genomic sequence data and the fraction of this that it was possible to annotate, and between the quantities of deposited biomolecular sequence and structure data. Against this backdrop, Bairoch announced the development of a separate, automatically generated counterpart to augment Swiss-Prot, to help disseminate the fruits of the increasingly abundant genome projects more efficiently, without compromising the quality of Swiss-Prot by including within it substantial quantities of uncharacterised data.

3.10 TrEMBL

By 1996, the first shock-waves from the impact of whole genome sequencing were beginning to be felt. The aftermath was greatest for databases whose maintenance involved significant amounts of manual annotation. Some did not recover. Swiss-Prot did survive the quake, but to do so, new processes had to be put in place.

At the time, Swiss-Prot had the highest standard of annotation of any publicly available protein sequence database: from the outset, one of its leading goals was to provide critical analyses for all of its constituent sequences. To this end, each entry was accompanied by a significant amount of annotation, derived primarily from original publications and review articles by an expanding group of curators, with occasional input from an international panel of experts. This high degree of meticulous manual annotation had always been the rate-limiting step for each release of the resource; however, faced with the increased data flow from the growing number of genome projects, this hugely labour-intensive process simply became untenable.

To keep up, it was clear that a new approach was needed. The products of genomic sequences had to be made available more swiftly; but how could this be achieved without compromising the high quality of the existing Swiss-Prot data, or eroding the editorial standards of the database in future? The answer was to prepare a computer-generated supplement, with entries in a Swiss-Prot-like format, derived by translation of coding sequences in the EMBL library – this was TrEMBL, first released in October 1996 (Bairoch & Apweiler, 1996). TrEMBL 1.0 contained almost 105,000 entries, not far off twice the size of Swiss-Prot 34.0 (59,000 entries), with which it was released in parallel.

Initially, TrEMBL was an unannotated supplement to Swiss-Prot. Over the years, however, to accelerate the process of upgrading TrEMBL entries to the Swiss-Prot standard, automatic protocols have been established to annotate sequences with information about their potential functions, metabolic pathways, active sites, cofactors, binding sites, domains, subcellular location, and so on. Such information was derived from similarity and motif searches, initially using patterns, profiles, fingerprints and so on from databases like PROSITE, PRINTS and Pfam, and later using the amalgamated protein family resource, InterPro. By February 2011, with many millions of entries, TrEMBL was almost 26 times larger than Swiss-Prot, illustrating the vast disparity between manual and computer-assisted annotation strategies.

3.11 InterPro

Rolf Apweiler was to spearhead the development of TrEMBL at the EBI in collaboration with Bairoch at the Swiss Institute of Bioinformatics (SIB). In 1997, Michael Ashburner (then Director of the EBI) awarded Attwood an EBI Visiting Fellowship. This entailed weekly visits from London, and led to frequent discussions between Apweiler, Attwood and Bairoch about sequence annotation. The feasibility of uniting PROSITE and PRINTS again

reared its head, but this time primarily as an instrument to help analyse and functionally annotate the growing numbers of uncharacterised genomic sequences. Compared to the original proposal in 1992, the case was much stronger, especially as there were now other related databases to bring into the picture: Daniel Kahn had released ProDom in 1994, and Richard Durbin had just announced Pfam. A new proposal was therefore submitted to the European Commission, and the vision of an integrated protein family database was finally funded.

In October 1999, a beta release of the unified resource was made with 2,423 entries (representing 615 domains, 1776 families, 27 repeats and 8 sites of PTM), based on Swiss-Prot 38.0 and TrEMBL 11.0 – this was InterPro (Apweiler *et al*., 2001). By that time, PROSITE and the Features Database had both undergone significant changes: PROSITE had seen 3-fold growth to 1,370 entries (release 16.0); meanwhile, the Features Database had grown 40-fold to 1,157 entries (release 23.1) and had been renamed 'PRINTS' (Attwood *et al*., 1994). The first release of InterPro therefore combined the contents of PROSITE 16.0 and PRINTS 23.1; it also incorporated descriptors from 241 profiles, together with 1,465 hidden Markov models from Pfam 4.0.

Fig. 5. Stylised illustration of the relationship between the InterPro integrating hub, its founding databases and its later additional partners, all of which contribute diagnostic signatures and, in some cases, protein family and domain annotation. The arrows indicate that information is shared both between satellite databases and between satellites and the central hub. See Table 3 for further details.

ProDom, although part of the original consortium (see Figure 5), was not included in the first release, initially because there was no obvious way of doing so. ProDom is built from automatically generated sequence clusters: it isn't a true signature database, in the sense that it doesn't exploit diagnostic discriminators; moreover, its sequence clusters need not have precise biological correlations, so can change between database releases. Assigning stable accession numbers to its entries was therefore impossible; this issue had to be addressed before it could be meaningfully included in InterPro. Other factors rendered a step-wise approach to the development of InterPro desirable. The scale of amalgamating just PROSITE, PRINTS and Pfam was immense. Trying to sensibly merge apparently equivalent database entries that, in fact, defined specific families, domains within those families, or even repeats within those domains, presented enormous challenges. In the beginning, InterPro therefore focused on amalgamating databases that offered some level of annotation, to facilitate the integration process.

Over the years, further partners joined the InterPro consortium, as illustrated in Figure 5. Today, with 12 primary sources, the integration challenges are legion (some of the complexity can be understood from the list of partners, and the numbers of their signatures that InterPro has incorporated, shown in Table 3)! With 21,185 entries in February 2011 (release 31.0), it is the most comprehensive integrated protein family database in the world (Hunter *et al.*, 2009).

Signature Database	Version	Signatures	Integrated Signatures
GENE3D	3.3.0	2,386	1,377
HAMAP	021210	1,675	1,429
PANTHER	7.0	80,933	1,777
PIRSF	2.74	3,248	2,791
PRINTS	41.1	2,050	2,009
PROSITE patterns	20.66	1,308	1,292
PROSITE profiles	20.66	901	877
Pfam	24.0	11,912	11,465
ProDom	2006.1	1,894	1,008
SMART	6.1	895	882
SUPERFAMILY	1.73	1,774	1,154
TIGRFAMs	9.0	3,808	3,796

Table 3. InterPro release 31.0, February 2011.

3.12 UniProt

The year 2004 marked a turning point for the way in which protein sequence data were to be collected and disseminated globally. The PIR-PSD, which had evolved from Dayhoff's *Atlas*, had been available online since 1986; Swiss-Prot, which originally built on PIR data, also became available in 1986; and TrEMBL had been released in 1996. The ongoing maintenance of these disparate resources over so many years had posed major funding headaches. For PIR, some of the difficulties were mitigated, at least in the early years, by charging for copies of their databases and for online access to their software; later, the international collaboration with MIPS and JIPID, supported by NSF and European grants, no doubt helped to sustain the resource.

Swiss-Prot, meanwhile, had had a rocky ride and had had to be rescued from the brink of closure, following a procedural 'catch-22' catastrophe: viewing Swiss-Prot as an international resource, the Swiss government declined to provide further support unless the database also gained a financial injection from a European Union (EU) grant; a joint proposal with the EBI for an EU infrastructure grant, however, was declined because Swiss-Prot was not being supported by the Swiss government! In May 1996, with only 2 months of salary remaining for the Swiss-Prot entourage, an Internet appeal was launched announcing the forthcoming closure, on 30 June, of Swiss-Prot and its associated databases and software tools, owing to lack of funding. This appeal stimulated a storm of protest on the Internet, in high-profile academic journals, and in the media. Such was the barrage that the Swiss government stepped in, offering interim funding until the end of the year. In the negotiations that followed, the need to create a stable vehicle for long-term funding both of Swiss-Prot and of the Swiss EMBnet Node was discussed, and resulted in the drafting of outline plans to establish a Swiss Institute of Bioinformatics (Bairoch, 2000).

Against this background, in 2002, with multinational funding from NIH, the NSF, the Swiss federal government and the EU, Swiss-Prot, TrEMBL and the PIR-PSD joined forces as the UniProt consortium. In forming the consortium, the idea was to build on the partners' many years of foundational work, by providing a stable, high-quality, unified database. This would serve as the world's most comprehensive protein sequence knowledgebase, replete with accurate annotations and extensive cross-references, and accompanied by freely-available, easy-to-use querying interfaces.

Under its hood, UniProt initially consisted of 3 separate database layers: the UniProt Archive (UniParc), to provide a complete, non-redundant collection of all publicly available protein sequence data; the UniProt Knowledgebase (UniProt), consisting of Swiss-Prot and TrEMBL, to act as the central database of protein sequences, with accurate, consistent and rich sequence and functional annotation; and the UniProt NREF databases (UniRef), to provide non-redundant subsets of the UniProt Knowledgebase, for efficient database searching (Apweiler *et al.*, 2004). By 2011, UniProt also included a Metagenomic and Environmental Sequence component, termed UniMES (The UniProt Consortium, 2011); by this time, UniProtKB:Swiss-Prot contained 525,207 entries, accompanied by UniProtKB:TrEMBL, with a staggering 13,499,622 entries.

3.13 The Swiss Institute of Bioinformatics (SIB)

Like the EBI, the need for which largely grew out of high-level negotiations to try to put the EMBL data library on a more stable financial footing, the Swiss Institute of Bioinformatics (SIB) grew out of similar high-level negotiations to establish long-term financial support for Swiss-Prot. At the time of the Swiss-Prot funding crisis, Bairoch was aware that the Swiss scientific authorities had been emphasising the need to establish centres of excellence in economically important, interdisciplinary areas that would be crucial for 'tomorrow's society'. Seizing upon this, together with Ron Appel, Philipp Bucher, Victor Jongeneel and Manuel Peitsch, he submitted a proposal to create a Swiss bioinformatics institute, whose goals were to:

- promote the development of bioinformatics software tools and databases;
- sustain high-quality bioinformatics research;
- collaborate with academic partners to provide a curriculum to train research scientists in the field of bioinformatics; and
- offer services to its user community through the Swiss Node of EMBnet.

After a lengthy period of consultation, the SIB was finally created as a non-profit foundation in March 1998, with Victor Jongeneel as the first director. The founders then went on to win funds for some of the SIB's activities from the Swiss Federal government: by law, only 50% of the Institute's work could be funded in this way – the rest had to come from other sources, preferably by commercial exploitation of its research.
Partly in response to this stipulation, but partly also because it had become clear that Swiss-Prot could not be reliably sustained solely with public funding, the decision was made to ask commercial users of the database to pay a licence fee. Various models for achieving this were tested; in the end, in 1997, Bairoch, Appel and Denis Hochstrasser decided that the best way forward was to set up a new company – this was Geneva Bioinformatics SA (GeneBio). Up to three quarters of the revenues now generated by GeneBio from sales of annual database and software licences are returned to SIB, thereby helping to bolster the work of the Swiss-Prot groups (Bairoch, 2000).
Today, the SIB leads and coordinates the field of bioinformatics in Switzerland: its vision, to help shape the future of the life sciences through excellence in bioinformatics services, research and education; its mission, to provide world-class core bioinformatics resources to both national and international research communities in fields spanning genomics, proteomics and systems biology. Many of its core activities, including maintenance of databases such as UniProt and InterPro, are carried out in close collaboration with the EBI.

3.14 The European Nucleotide Archive (ENA)

Meanwhile, with the advent of large-scale sequencing projects and the dawn of Next Generation Sequencing (NGS) technologies, a mounting tsunami of nucleotide sequence data was growing force across the globe; a number of important developments were to take place in its wake. By 2003, it was clear that there was a need to provide access not only to the most recent versions of sequences, but also to their historical artifacts – following the rush to patent genetic information, issues of priority became increasingly important, and it was vital to be able to see sequence entries exactly as they appeared in the past. Accordingly, the EBI established a Sequence Version Archive (Leinonen *et al*., 2003), to store both current and earlier versions of entries in the EMBL data library (which, by then, had been dubbed EMBL-Bank).
By September 2004, EMBL-Bank had grown prodigiously, with more than 42 million entries (Kanz *et al*., 2005) and, by 2007, was accompanied by the Ensembl Trace Archive (ETA) – the ETA was set up to provide a permanent archive for single-pass DNA sequencing reads (from whole-genome shotgun, EST and other large-scale sequencing projects) and associated traces and quality values. Together, EMBL-Bank and the ETA became known as ENA, the European Nucleotide Archive, Europe's primary nucleotide-sequence repository (Cochrane *et al*., 2008). Throughout 2007, ENA continued to grow in terms both of its volume and of the nature of data it contained such that, by October of that year, it included more than 1.7 billion records (comprising ~1.7 trillion ($1.7x10^{12}$) base pairs of sequence) (Cochrane *et al*., 2008). By 2010, ENA had embraced a third component – the Sequence Read Archive (SRA) – and now contained ~500 billion raw and assembled sequences, comprising $50x10^{12}$ base pairs; this is a phenomenal growth in just 3 years! During this period, NGS reads held in the SRA had become the largest and fastest growing source of new data, and accounted for ~95% of all base pairs made available by ENA (Leinonen *et al*., 2011). Contributing to this

mass of data were the completed genomes of more than 1,400 cellular organisms, and 3,000 viruses and phages.

But such enormous progress comes at a cost, challenging current IT infrastructures to the limit. Some of the oldest data in ENA date back to the early '80s, with the inception of the EMBL data library. As an aside, it is somewhat ironic that, even in those days, there were distribution headaches. Bairoch, for example, relates how difficult it was to transfer version 2 of the EMBL data library from computer tape to a mainframe computer and thence to his microcomputer, because the mainframe had no communication protocol to talk to a microcomputer – he therefore had to spend the night transferring the data, screen by screen, using a 300 baud acoustic modem (Bairoch, 2000). To put this in perspective, this version of EMBL-Bank contained 811 nucleotide sequences (with more than 1 million base pairs) – this is about the same amount of data that currently enters ENA every 2 seconds.

Today, ENA holds more than 20 terabases of nucleotide sequence data, which, combined with its annotation information, and so on, occupies more than 230 terabytes of disk space. The infrastructure required to store, maintain and service such a vast archive, and the cost of doing so, is beyond anything that either the originators of the first databases, or the developers of the new sequencing technologies could have conceived. Interestingly, in February 2011, the NCBI announced that it would be discontinuing its Sequence Read and Trace Archives for high-throughput sequence data, owing to budget constraints. The closure of the databases is to be phased, and completed within 12 months. The NCBI is still committed to supporting and developing information resources for biological data derived from NGS technologies (genotypes, variations, assemblies, gene expression data, and so on), but will need to find new funding strategies for access to and storage of the existing data.

3.15 ELIXIR

The opportunities NGS technologies present for advancing life science research (especially in areas such as healthcare, food security, energy diversification and environmental protection) are incredibly exciting; but these opportunities will be lost if they are not underpinned by a robust, effective and sustainable information infrastructure. The best estimates today suggest that, by 2020, NGS technologies will be producing data at up to a million times the current rate. Development of an appropriate infrastructure to manage the data deluge is therefore paramount.

The ELIXIR project is the realisation of this urgent need. Recognising that the task is of such magnitude that it cannot be tackled by a single organisation, it is a call to arms for international cooperation in building a pan-European infrastructure to help extract the maximum value from the investments that have already been made, and from those that will be made in future, in this area. The plan is for the ELIXIR infrastructure to be distributed across a variety of 'Nodes' hosted by centres of excellence across Europe, and for each of these to be connected to the EBI central 'Hub'. It is expected that some of the Nodes will act as national coordination centres to expedite interactions both with the Hub and with local funders; Nodes that perform similar functions will be expected to collaborate to form ELIXIR service networks, providing data or compute resources, or training, according to their speciality, as depicted in Figure 6.

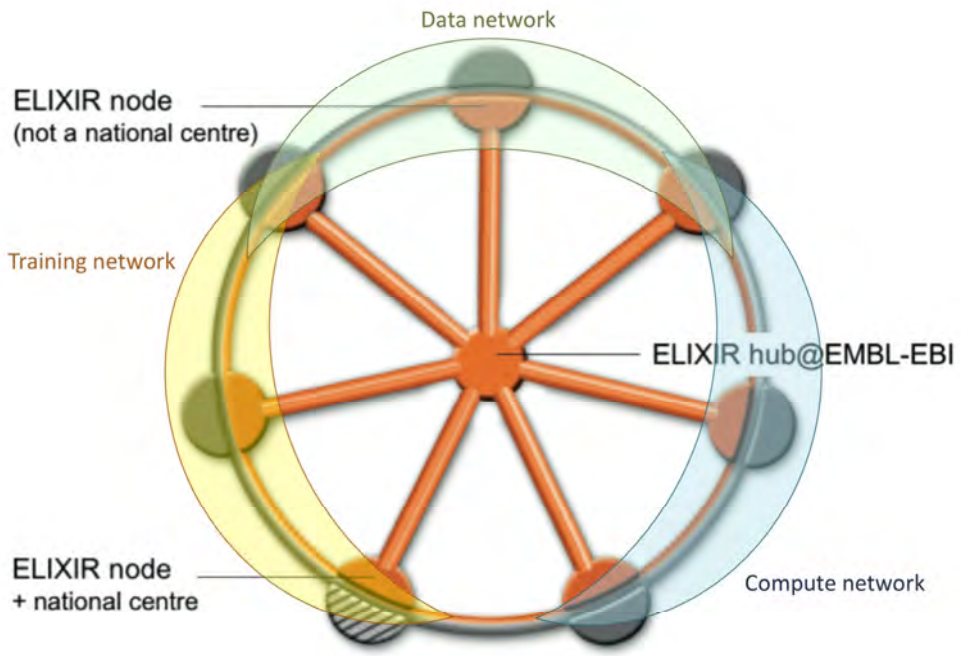

Fig. 6. Proposed topology of the ELIXIR Hub and Nodes. In an arrangement reminiscent of EMBnet 23 years before it, some of the Nodes are expected to serve as national bioinformatics centres; others, with similar functions, will collaborate as service networks, for example to provide data or compute resources, or training.

Initially, the numbers of Nodes is expected to be small, growing to ~20 during the first 5 years of the initiative (during the preparatory phase, more than 50 institutions submitted expressions of interest in becoming ELIXIR Nodes), at a cost of several hundred million euro. To garner support for the business case, governments of the European Member States have been invited to sign a non-binding Memorandum of Understanding (MoU) in order to initiate negotiations to construct ELIXIR; the MoU will become effective once 5 countries and the EMBL have signed. Europe's databases (estimated to number around 500), especially those hosted by the EBI, will become the foundation of the new ELIXIR infrastructure as part of its mission, *"to construct and operate a sustainable infrastructure for biological information in Europe to support life science research and its translation to medicine and the environment, the bio-industries and society"* (Thornton, 2011).

4. The development and spread of tools to keep pace with the new technologies

With the sequencing of biopolymers and subsequent organisation of the growing mass of biosequences in databases, visual comparison techniques became tedious, not least because

"*the determination of the significance of a given result usually is left to intuitive rationalization*" (Needleman & Wunsch 1971). To reduce reliance on manual (often subjective) interpretation and put sequence analysis on a more systematic footing, algorithms to analyse and compare sequences began to emerge. As early as 1966, Fitch proposed computational analysis to study evolutionary homology, using mutation values to indicate how many nucleotides in the genomic code must change in order to introduce change (mutation) at the amino acid level. In 1970, Needleman and Wunsch described the first algorithm to quantify the similarity between two protein sequences (so-called global alignment) – today, this algorithm is still used to identify similarities between two sequences and infer likely ancestry. Years later, Smith and Waterman (1981) presented an algorithm to find local similarities: "*to find a pair of segments, one from each of two long sequences, such that there is no other pair of segments with greater similarity*". In time, more efficient methods were required to compare newly sequenced proteins against the rapidly expanding databases. FASTP was the first 'fast' algorithm (Lipman & Pearson 1985).

Search algorithms like this afforded many of the earliest and most exciting discoveries attributable to 'bioinformatics'. For example, one of the first observations that gave a clue to the molecular mechanism of neoplastic transformation was provided by the finding of a near identity in amino acid sequence between the platelet-derived growth factor (PDGF) B-chain and a region in the transforming protein, p28sis, of simian sarcoma virus (SSV), an agent that causes sarcomas and gliomas in experimental animals (Waterfield *et al*., 1983). This finding arose from computer searches using the Wilbur and Lipman algorithm on the, at the time (1983) available, NEWAT protein database created by Doolittle *et al*. This first success story, where simple sequence comparison led to the completely new concept of gene-oncogene, showed the medical community the enormous potential of computer techniques for sequence comparison and analysis.

In a similar way, DNA sequencing having been revolutionised by Sanger and by subsequent improvements of his technique, and having given rise to the growing number of nucleotide sequences being collected in data repositories like the EMBL data library and GenBank, so too algorithms to search these databases became a necessity. FASTA was a more sensitive modification of FASTP, and had the advantage of being able to search nucleotide sequence databases with either a nucleic acid or protein sequence by translating the DNA database during the search (Pearson & Lipman 1988). Later, somewhat overshadowing these developments, came the Basic Local Alignment Search Tool, BLAST (Altschul *et al*., 1990); this offered an extended tool-set to apply any kind of sequence database search, and is still the most widely used tool in bioinformatics. The success of BLAST spawned a number of more specialised sequence search methods, such as PSI-BLAST, PHI-BLAST, BLAT, and so on, and is itself still in continuous development (Camacho *et al*., 2009).

Aside from these very popular database search tools, many other sequence, annotation and expression analysis tools were developed for a broad range of applications: *e.g*., for pattern recognition, for protein and RNA secondary structure prediction, for microarray data analysis, for proteome and genome annotation, and so on. In the early '90s, building on the existing University of Wisconsin Genetics Computer Group (UWGCG, or simply GCG) package, several such algorithms were collected at the EMBL and packaged as 'GCGEMBL Utilities', later known as 'Extended GCG'. However, GCG was then commercialised and its distribution policy changed. Reacting against the new policies, in 1998 several software developers founded EMBOSS, the European Molecular Biology Open Software Suite. Their

aim was to develop new sequence analysis tools, by "*replacing popular but obsolete EGCG applications*," and integrating with SRS, ACEDB, and a range of other publicly available software interfaces and tools. The idea was to encourage other developers to use the EMBOSS software libraries, and especially to harness the expertise and potential additional manpower at EMBnet Nodes (*e.g.*, in Germany, Italy, France, The Netherlands, Austria, Russia, Switzerland, Israel, Spain, Norway, and so on). Target users of the resource included those at the Sanger Centre, those served by EMBnet, and those in academic and pharmaceutical settings. Funded by the Wellcome Trust for 3 years, the project was a collaborative effort of the Sanger Centre, EMBnet UK (SEQNET), the EBI and CNRS Montpellier.

With the pivotal support of EMBnet, EMBOSS quickly became a comprehensive bioinformatics resource (Rice *et al.*, 2000). There are now several incarnations of the suite with different GUIs, including the EMBOSS team's Java-based interface, jEMBOSS; the Belgian and Argentinian EMBnet Nodes' wEMBOSS; and the EMBOSS GUI from the National Research Council of Canada. Today, EMBOSS is still being developed, adopting new specific file formats and algorithms in order to embrace the world of NGS data analysis.

Another important development driven by the EMBL was the Sequence Retrieval System (SRS), an information indexing system applied to flat-file databases, such as the EMBL data library, Swiss-Prot and PROSITE (Etzold and Argos, 1993). SRS became the most widely used data-retrieval system for flat-file systems, with an extended GUI to extract not only sequences but all related information, via an exhaustive sequence query and export system (Zdobnov *et al.*, 2002).

Europe-wide, there are vast numbers of other specialised biological data-analysis, data-visualisation and data-retrieval tools available: many of these are provided by the EBI; others by the SIB's ExPASy Proteomics Server; some are offered via the National and Specialist Nodes of EMBnet; others are available as Web services collected in the BioCatalogue (Bhagat *et al.*, 2010). The BioCatalogue evolved from the EMBRACE registry (Pettifer *et al.*, 2010), one of the end products of the EMBRACE project (European Model for Bioinformatics Research and Community Education) – this was a 5-year FP7 Network of Excellence, whose main goal was to orchestrate highly integrated access to a broad range of bio-molecular data and software packages. Achieving this required standardised access to tools and databases; to this end, the decision was to use Web services. In consequence, many of the project partners adapted their tools and database-access protocols, and logged their Web services in a common registry. At the end of EMBRACE, in 2010, the registry was handed over to the BioCatalogue, which is now being maintained in collaboration with myExperiment, myGrid, seekda and BioMoby, and hosts 2,053 services from 147 service providers and 505 members.

5. The central place of bioinformatics in modern biology

Clearly, we have travelled a very long way since Jensen and Evans positioned a single amino acid (a terminal phenylalanine) in insulin (Jensen & Evans, 1935; Sanger, 1945; Sanger, 1988) and Sanger elucidated its complete sequence, the first of any protein (recall Table 2). In a story spanning something like 70 years, bioinformatics has given us the first 'complete' catalogues of DNA and protein sequences, including the genomes and proteomes

of organisms across the entire Tree of Life; it has furnished the requisite software to help analyse biological data on an unprecedented scale; it has hence yielded the possibilities to understand more about evolutionary processes in general, our place in the Tree of Life in particular, and ultimately, a great deal more about health, disease and disease processes. Figure 7 offers a summary of some of the most important landmarks that have charted the development of bioinformatics in Europe and helped to place it at the heart of 21st century biology.

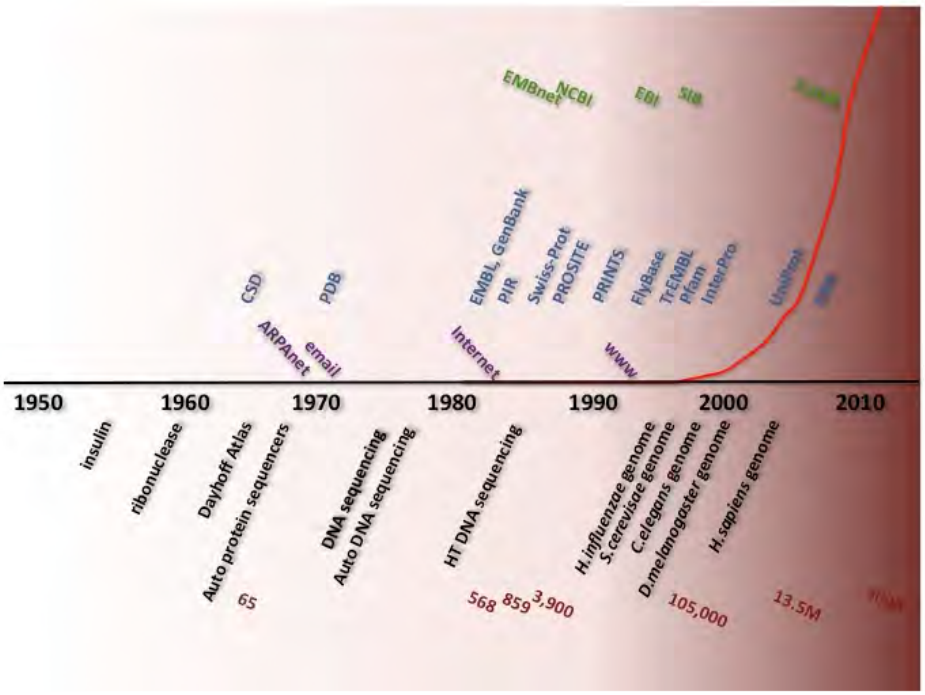

Fig. 7. Historical milestones that have placed bioinformatics at the heart of 21st century biology, from the determination of the first amino acid sequence, to the development of an archive of 500 billion nucleotide sequences. Some major milestones are denoted in black; key computing innovations are indicated in purple; example databases are indicated in blue; organisations and institutions in green; numbers of sequences in red, the growing mass of which is highlighted both in the red curve and the background gradient – the impact of genomic sequencing in the mid '90s is clear.

6. Conclusion – European bioinformatics goes global

The history of bioinformatics has clearly been a convoluted interplay between events in Europe, the USA, Japan and across the globe. Here, we have attempted to recount the story primarily from a European perspective as it unfolded largely from the point of view of sequence data: in terms of the technological innovations that spawned their extraordinary

growth and dissemination, of the databases that grew up to manage and analyse them, and of the institutions and infrastructural initiatives that arose to try to give those databases some measure of financial stability. In so doing, we accept that we've only scratched the surface, and we regret any shortcomings that may have arisen from the necessary omission of so many of the other important details and perspectives.

Clearly, the evolution and impact of bioinformatics reaches far beyond Europe, and there are now many organisations world-wide with missions to bring life science data to their local communities, to make freely available easy-to-use software tools with which to analyse the data, and to provide training, both to users of bioinformatics databases and software, and to new generations of bioinformatics trainers (Schneider *et al.*, 2010). In this context, EMBnet, for example, which began life as the European Molecular Biology Network, is now a global bioinformatics network, maintaining fruitful cooperations with the Iberoamerican (SoIBio) and Asia Pacific (APBioNet) bioinformatics networks, as well as with the USA-based International Society for Computational Biology (ISCB); it has also established close ties with the African Society for Bioinformatics and Computational Biology (ASBCB), and synergies with other relevant groups in northern Africa are now developing. Interestingly, 33 years ago, Joshua Lederberg observed that, "*the claim of science to universal validity is supportable only by virtue of a strenuous commitment to global communication*" (Lederberg, 1978). Today, this is a commitment that EMBnet vigorously pursues; in a similar spirit, we can be quite sure that the contribution of Europe to the future evolution of bioinformatics will continue in a global arena.

7. Acknowledgement

We would like to thank Vicky Schneider for providing the inspiration (and the title) for this chapter.

8. References

Adams, M.J.; Blundell, T.L., Dodson, E.J., Dodson, G.G., Vijayan, M., Baker, E.N., Harding, M.M., Hodgkin, D.C., Rimmer, B. & Sheat, S. (1969) Structure of Rhombohedral 2 Zinc Insulin Crystals. *Nature,* 224, 49-495.

Adams, M.D.; Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D. *et al*. (2000) The genome sequence of *Drosophila melanogaster*. *Science*, 287, 2185-2195.

Akrigg, D.A.; Attwood, T.K., Bleasby, A.J., Findlay, J.B.C., North, A.C.T., Parry-Smith, D.J., Perkins, D.N. & Wootton, J.C. (1992) SERPENT - An information storage and analysis resource for protein sequences. *CABIOS*, 8(3), 295-296.

Allen, F. H.; Davies, J.E., Galloy, J.J., Johnson, O., Kennard, O., Macrae, C.F., Mitchell, E.M., Mitchell, G.F., Smith, J.M. & Watson, D.G. (1991) The Development of Versions 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.*, 31, 187-204.

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

Anderson, S, Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., Schreier, P.H., Smith, A.J., Staden, R. and Young, I.G. (1981) Sequence and organization of the human mitochondrial genome. *Nature*, 290, 457-465.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M., Servant, F., Sigrist, C.J. & Zdobnov, E.M. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, 29(1), 37-40.

Apweiler, R.; Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N. & Yeh, L.S. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, 32(Database issue), D115-119.

Ashburner, M. (1996) Won for all: how the Drosophila genome was sequenced. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA.

Ashburner, M. & Drysdale, R. (1994) FlyBase – the *Drosophila* genetic database. *Development*, 120(7), 2077-2079.

Bairoch, A. (1982) Suggestion to research groups working on protein and peptide sequence. *Biochem.J.*, 203(2), 527-528.

Bhagat, J., Tanoh, F., Nzuobontane, E., Laurent, T., Orlowski, J., Roos, M., Wolstencroft, K., Aleksejevs, S., Stevens, R., Pettifer, S., Lopez, R. & Goble, C.A. (2010) BioCatalogue: a universal catalogue of web services for the life sciences. Nucleic Acids Res., 38, W689-694

Bairoch, A. (2000) Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times! *Bioinformatics*, 16(1), 48-64.

Bairoch, A. & Boeckmann, B. (1991) The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.*, 19 Suppl., 2247-2249.

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

Bairoch, A. & Apweiler, R. (1996) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res.*, 24(1), 21-25.

Bairoch, A. & Bucher, P. (1994) PROSITE: recent developments. *Nucleic Acids Res.*, 22(17), 3583-3589.

Barker, W.C.; George, D.G., Mewes, H.W. & Tsugita, A. (1992) The PIR-International Protein Sequence Database. *Nucleic Acids Res.*, 20 Suppl., 2023-206

Benson, D.; Boguski, M., Lipman, D.J. & Ostell, J. (1990) The National Center for Biotechnology Information. *Genomics*, 6, 389-391.

Benson, D.; Lipman, D.J. & Ostell, J. (1993) GenBank. *Nucleic Acids Res.*, 21(13), 2963-2965.

Berman, H.M.; Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, 28(1), 235-242.

Berman, H.; Henrick, K. & Nakamura, H. (2003) Announcing the worldwide Protein Data Bank. *Nature Structural Biology,* 10, 980.

Berman, H. (2008) The Protein Data Bank: A historical perspective. *Foundations of Crystallography*, 64(1), 88-95.

Bernstein, F.C.; Koetzle, T.F., Williams, G.J., Meyer, E.F. Jr, Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) The Protein Data Bank. A computer-based archival file for macromolecular structures. *J.Mol.Biol.*, 112(3), 535-

542. Reprinted in *Eur. J. Biochem*., 80(2), 319-24 (1977); and *Archives of Biochemistry and Biophysics*, 185(2), 584-591 (1978).

Boutselakis, H.; Dimitropoulos, D., Fillon, J., Golovin, A., Henrick, K., Hussain, A., Ionides, J., John, M., Keller, P. A., Krissinel, E., McNeil, P., Naim, A., Newman, R., Oldfield, T., Pineda, J., Rachedi, A., Copeland, J., Sitnov, A., Sobhany, S., Suarez-Uruena, A., Swaminathan, J., Tagari, M., Tate, J., Tromm, S., Velankar, S. & Vranken, W. (2003) E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. *Nucleic Acids Res*., 31(1), 458-462.

Brown, H.; Sanger, F. & Kitai, R. (1955), The structure of pig and sheep insulins. *Biochemical Journal,* 60(4), 556–565.

Burley, S.K.; Almo, S.C., Bonanno, J.B., Capel, M., Chance, M.R., Gaasterland, T., Lin, D., Sali, A., Studier, F.W. & Swaminathan, S. (1999) Structural genomics: beyond the human genome project. *Nat. Genet*., 23(2), 151-157.

Butler, D. (1999) Life science facilities in crisis as Brussels switches off funding. *Nature*, 402, 3-4. *C. elegans* Sequencing Consortium. (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, 282, 2012-2018.

Camacho, C.; Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T.L. (2009) BLAST+: architecture and applications. BMC Bioinformatics, 10, 421.

Cherry, J.M.; Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S. & Botstein, D. (1998) SGD: Saccharomyces Genome Database. *Nucleic Acids Res*., 26(1), 73-79.

Cochrane, G.; Akhtar, R., Aldebert, P., Althorpe, N., Baldwin, A., Bates, K., Bhattacharyya, S., Bonfield, J., Bower, L., Browne, P., Castro, M., Cox, T., Demiralp, F., Eberhardt, R., Faruque, N., Hoad, G., Jang, M., Kulikova, T., Labarga, A., Leinonen, R., Leonard, S., Lin, Q., Lopez, R., Lorenc, D., McWilliam, H., Mukherjee, G., Nardone, F., Plaister, S., Robinson, S., Sobhany, S., Vaughan, R., Wu, D., Zhu, W., Apweiler, R., Hubbard, T. & Birney, E. (2008) Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Res*., 36(Database issue), D5-D12.

Dayhoff, M.O.; Eck, R.V., Chang, M.A. & Sochard, M.R. (Eds.) (1965) *Atlas of protein sequence and structure*. National Biomedical Research Foundation, Silver Spring, Maryland, USA.

Dayhoff, M.O. to Berkley, C. (1967) Margaret O.Dayhoff Papers, Archives of the National Biomedical Research Foundation, Washington, D.C., USA.

Dayhoff, M.O.; Schwartz, R.M., Chen, H.R., Barker, W.C., Hunt, L.T. & Orcutt, B.C. (1981) Nucleic Acid Sequence Database. *DNA*, 1, 51-58; b) Dayhoff, M.O., Schwartz, R.M., Chen, H.R., Hunt, L.T., Barker, W.C. & Orcutt, B.C. (1981) Data Bank. *Nature*, 290, 8.

Dickson, D. & Abbott, A. (1993) Cambridge and Heidelberg compete for new European gene database. *Nature*, 361, 383.

Dodson, G. (2005) Fred Sanger: sequencing pioneer. *Biochem. J*., doi:10.1042/BJ2005c013.

Doelz, R. (1994) Biocomputing on a Server Network. *EMBnet.news*, 1(2), 6-8.

EMBL (1992) The European Bioinformatics Institute (EBI): A Proposal

Eck, R.V. & Dayhoff, M.O. (1966) *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Spring, Maryland, USA.

Doolittle, R.F. (1986) Of Urfs and Orfs: a primer on how to analyze derived amino acid sequences. University Science Books, 20 Edgehill Road, Mill Valley, CA 94941, USA. ISBN 0-935702-54-7

Etzold, T. & Argos, P. (1993) SRS – an indexing and retrieval tool for flat file data libraries. Comput.Appl. Biosci., 9, 49-57.

Eeckman, F.H. & Durbin, R. (1995) ACeDB and macace. *Methods Cell Biol*., 48, 583-605.

Fitch, W.M. (1966) An improved method of testing for evolutionary homology. J.Mol.Biol., 16, 9-16.

Fleischmann, R.D.; Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F. *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*. *Science*, 269, 496-512.

Franklin, R.E. & Gosling, R.G. (1953) a) The structure of sodium thymonucleate fibres. I. The influence of water content. *Acta Cryst*., 6, 673-677; b) The structure of sodium thymonucleate fibres. II. The cylindrically symmetrical Patterson function. *Ibid*., 678-685; c) Molecular configuration in sodium thymonucleate. *Nature*, 171, 740-741.

Fraser, C.M.; Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A. *et al*. (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*, 270, 397-403.

George, D.G.; Barker, W.C. & Hunt, L.T. (1986) The protein identification resource (PIR). *Nucleic Acids Res*., 14(1), 11-15.

George, D.G.; Dodson, R.J., Garavelli, J.S., Haft, D.H., Hunt, L.T., Marzec, C.R., Orcutt, B.C., Sidman, K.E., Srinivasarao, G.Y., Yeh, L.S., Arminski, L.M., Ledley, R.S., Tsugita, A. & Barker, W.C. (1997) The Protein Information Resource (PIR) and the PIR-International Protein Sequence Database. *Nucleic Acids Res*., 25(1), 24-28.

Gingeras, T.R. & Roberts, R.J. (1980) Steps towards computer analysis of nucleotide sequences. *Science*, 209, 1322-1328.

Goffeau, A.; Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B. *et al*. (1996) Life with 6000 genes. *Science*, 274, 546-567.

Goodner, B.; Hinkle, G., Gattung, S., Miller, N., Blanchard, M. *et al*. (2001). Genome Sequence of the Plant Pathogen and Biotechnology Agent *Agrobacterium tumefaciens* C58. *Science*, 294, 2323-2328.

Hamm, G.H. & Cameron, G.N. (1986) The EMBL data library. *Nucleic Acids Res*., 14(1), 5–9.

Harvey, M. & McMeekin, A. (2004) Public-private collaborations and the race to sequence *Agrobacterium tumefaciens*. *Nat. Biotechnol*., 22(7), 807-810.

Henikoff, S. & Henikoff, J.G. (1991) Automated assembly of protein blocks for database searching. *Nucleic Acids Res*., 19(23), 6565-6572.

Hirs, C.H.W.; Moore, S. & Stein, W.H. (1960) The Sequence of the Amino Acid Residues in Performic Acid-oxidized Ribonuclease. *J.Biol.Chem*., 235, 633-647.

Hobohm, U.; Scharf, M., Schneider, R. & Sander, C. (1992) Selection of representative protein data sets. *Protein Sci*., 1(3), 409-417.

Hogeweg, P. (1978) Simulating the growth of cellular forms. *Simulation*, 31, 90-96.

Hogeweg, P. & Hesper, B. (1978) Interactive instruction on population interactions. *Comput. Biol. Med*., 8, 319-327.

Huala, E.; Dickerman, A.W., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, M., Huang, W., Mueller, L.A., Bhattacharyya, D., Bhaya, D., Sobral, B.W., Beavis, W., Meinke, D.W., Town, C.D., Somerville, C., Rhee & S.Y. (2001) The Arabidopsis Information Resource (TAIR): a comprehensive

database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res*., 29(1), 102-105.

Hubbard, T.; Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I. & Clamp, M. (2002) The Ensembl genome database project. *Nucleic Acids Res*., 30(1), 38-41.

Hunter, D.J. (2006) Genomics and proteomics in epidemiology: treasure trove or "high-tech stamp collecting"? *Epidemiology*, 17(5), 487-489.

Hunter, S.; Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res*., 37 (Database Issue), D211-D215.

International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, 431, 931-945.

Jensen, H. & Evans Jr., E.A. (1935) Studies on crystalline insulin. XVIII. The nature of the free amino groups in insulin and the isolation of phenylalanine and proline from cyrstalline insulin. *J.Biol.Chem*., 108, 1-12.

Kanehisa, M.; Fickett, J.W. & Goad, W.B. (1984) A relational database system for the maintenance and verification of the Los Alamos sequence library. *Nucleic Acids Res*., 12(1), 149-158.

Kanz, C.; Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., Diez, F.G., Harte, N., Kulikova, T., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Sobhany, S., Stoehr, P., Tuli, M.A., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W. & Apweiler, R. (2005) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res*., 33(Database issue), D29-D33.

Kendrew, J.C.; Bodo, G., Dintzis, H.M., Parrish, R. G., Wyckoff, H. & Phillips, D. C. (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181, 662-666.

Kennard, O.; Watson, D. G. & Town, W. G. (1972) Cambridge Crystallographic Data Centre. I. Bibliographic File. *J. Chem. Doc*., 12(1), 14-19.

Kennard, O. (1997) From private data to public knowledge. In *The Impact of Electronic Publishing on the Academic Community*, an International Workshop organised by the Academia Europaea and the Wenner-Gren Foundation, Wenner-Gren Center, Stockholm, 16-20 April, 1997. Ian Butterworth, Ed. Published by Portland Press Ltd., London, UK. ISBN 1 85578 122 0

Kneale, G.G. & Kennard, O. (1984) The EMBL nucleotide sequence data library. *Biochem. Soc. Trans*., 12, 1011–1014.

Kreppel, L.; Fey, P., Gaudet, P., Just, E., Kibbe, W.A., Chisholm, R.L. & Kimmel, A.R. (2004) dictyBase: a new Dictyostelium discoideum genome database. *Nucleic Acids Res*., 32(Database issue), D332-D333.

Lander, E.S.; Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, 409, 860-921.

Lederberg, J. (1978) Digital Communications and the Conduct of Science; the New Literacy. *Proceedings of the IEEE*, 66, 1314-1319.

Leinonen, R.; Nardone, F., Oyewole, O., Redaschi, N. & Stoehr, P. (2003) The EMBL sequence version archive. *Bioinformatics*, 19(14), 1861-1862.

Leinonen, R.; Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Ten Hoopen, P., Vaughan, R., Zalunin, V. & Cochrane, G. (2011) The European Nucleotide Archive. *Nucleic Acids Res.*, 39(Database issue), D28-31.

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

Meyer, E.F. (1997) The first years of the Protein Data Bank. *Protein Science,* 6, 1591-1597.

Muirhead, H. & Perutz, M. (1963) Structure of hemoglobin. A three-dimensional fourier synthesis of reduced human hemoglobin at 5.5 Å resolution. *Nature,* 199, 633–38.

Nakamura, H.; Ito. N. & Kusunoki, M. (2002) Development of PDBj: Advanced database for protein structures. *Tanpakushitsu Kakusan Koso.,* 47(8 Suppl), 1097-1101.

Nature Editorial. (1999) Vacuum at the heart of Europe. *Nature*, 402, 1.

Needleman, S.B. & Wunsch, C.D. (1971) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J.Mol.Biol., 48, 443-453.

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

Pettifer, S., Ison, J., Kalas, M., Thorne, D., McDermott, P., Jonassen, I., Liaquat, A., Fernandez, J.M., Rodriguez, J.M., Partners, I., Pisano, D.G., Blanchet, C., Uludag, M., Rice, P., Bartaseviciute, E., Rapacki, K., Hekkelman, M., Sand, O., Stockinger, H., Clegg, A.B., Bongcam-Rudloff, E., Salzemann, J., Breton, V., Attwood, T.K., Cameron, G. & Vriend, G. (2010) The EMBRACE web service collection. Nucleic Acids Res., 38, Suppl. W683-688

Philipson, L. (1992) Letter to EMBL Council Delegates, with annexes

Protein Data Bank (1971) *Nature New Biology*, 233, 223.

Protein Data Bank (1973) *Acta Crystallogr. sect. B,* 29, 1746.

Rice, P., Longden, I. & Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet., 16, 276-277

Ryle, A.P.; Sanger, F., Smith, L.F. & Kitai, R. (1955) The disulphide bonds of insulin. *Biochem. J.*, 60(4), 541–556.

Sanger, F. (1945) The free amino groups of insulin. *Biochem. J.*, 39, 507-515.

Sanger, F. & Tuppy, H. (1951) a) The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochem. J.*, 49, 463–481; b) The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Ibid.*, 481–490.

Sanger, F. & Thompson, E.O.P. (1953) a) The amino-acid sequence in the glycyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochem. J.*, 53, 353–366; b) The amino-acid sequence in the glycyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Ibid.*, 366–374.

Sanger, F.; Thompson, E.O.P. & Kitai, R. (1955) The amide groups of insulin. *Biochem. J.*, 59(3), 509–518.

Sanger, F.; Coulson, A.R., Friedmann, T., Air, G.M., Barrell, B.G., Brown, N.L., Fiddes, J.C., Hutchison, C.A. 3rd, Slocombe, P.M. & Smith, M. (1978) The nucleotide sequence of bacteriophage phiX174. *J. Mol. Biol*., 125(2), 225-246.

Sanger, F.; Coulson, A.R., Hong, G.F., Hill, D.F. & Petersen, G.B. (1982) Nucleotide sequence of bacteriophage lambda DNA. *J.Mol.Biol*., 162(4), 729-773.

Sanger, F. (1988) Sequences, sequences, and sequences. *Ann.Rev.Biochem*., 57, 1-28.

Sidman, K.E.; George, D.G., Barker, W.C. & Hunt, L.T. (1988) The protein identification resource (PIR). *Nucleic Acids Res*., 16(5), 1869-1871.

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

Smith, T.F. (1990) The history of the genetic sequence databases. *Genomics*, 6, 701-707.

Smyth, D.G.; Stein, W.H. & Moore, S. (1963) The Sequence of Amino Acid Residues in Bovine Pancreatic Ribonuclease: Revisions and Confirmations. *J.Biol.Chem*., 238, 227-234.

Schneider, M.V.; Watson, J., Attwood, T., Rother, K., Budd, A., McDowall, J., Via, A., Fernandes, P., Nyronen, T., Blicher, T., Jones, P., Blatter, M.C., De Las Rivas, J., Judge, D.P., van der Gool, W. & Brooksbank, C. (2010) Bioinformatics training: a review of challenges, actions and support requirements. *Brief Bioinform*., 11(6), 544-551

Sonnhammer, E.L. & Kahn, D. (1994) Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci*., 3(3), 482-492.

Sonnhammer, E.L.; Eddy, S.R. & Durbin, R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28(3), 405-420.

Strasser, B. (2008) *GenBank – Natural history in the 21st century?* Science, 322, 537-538.

Thornton, J. (2011) European Life Sciences Infrastructure for Biological Information, ELIXIR Business Case. European Bioinformatics Institute, Hinxton, Cambridge, UK.

UniProt Consortium. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res*., 39(Database issue), D214-D219.

Velankar, S.; Best, C., Beuth, B., Boutselakis, C. H., Cobley, N., Sousa Da Silva, A.W., Dimitropoulos, D., Golovin, A., Hirshberg, M., John, M., Krissinel, E.B., Newman, R., Oldfield, T., Pajon, A. , Penkett, C. J., Pineda-Castillo, J., Sahni, G., Sen, S., Slowley, R., Suarez-Uruena, A., Swaminathan, J., van Ginkel, G., Vranken, W. F., Henrick, K. & Kleywegt, G. J. (2010) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res*., 38, D308–D317.

Venter, J.C.; Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J. *et al.* (2001) The sequence of the human genome. *Science*, 291, 1304-1351.

Waterfield MD, Scrace GT, Whittle N, Stroobant P, Johnsson A, Wasteson A, Westermark B, Heldin CH, Huang JS, Deuel TF. Platelet-derived growth factor is structurally related to the putative transforming protein p28sis of simian sarcoma virus. Nature. 1983 Jul 7-13;304(5921):35-9.

Watson, J.D. & Crick, F.H.C. (1953) Molecular structure of nucleic acids. *Nature*, 171, 737-738.

Wilbur, W.J. & Lipman, D.J. (1983) Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA*., 80(3), 726-730.

Wood, D.W.; Setubal, J.C., Kaul, R., Monks, D.E., Kitajima, J.P. *et al.* (2001). The Genome of the Natural Genetic Engineer *Agrobacterium tumefaciens* C58. *Science*, 294, 2317-2323.

Wu, C.H.; Yeh, L.S., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z., Kourtesis, P., Ledley, R.S., Suzek, B.E., Vinayaka, C.R., Zhang, J. & Barker, W.C. (2003) The Protein Information Resource. *Nucleic Acids Res*., 31(1), 345-347.

Wyckoff, H.W.; Hardman, K.D., Allewell, N.M., Inagami, T., Johnson, L.N. & Richards, F.M. (1967). The structure of ribonuclease-S at 3.5 Å resolution. *J. Biol. Chem*., 242, 3984–3988.

Zdobnov, E.M., Lopez, R., Apweiler, R. & Etzold, T. (2002) The EBI SRS server – new features. Bioinformatics, 18, 1149-1150.