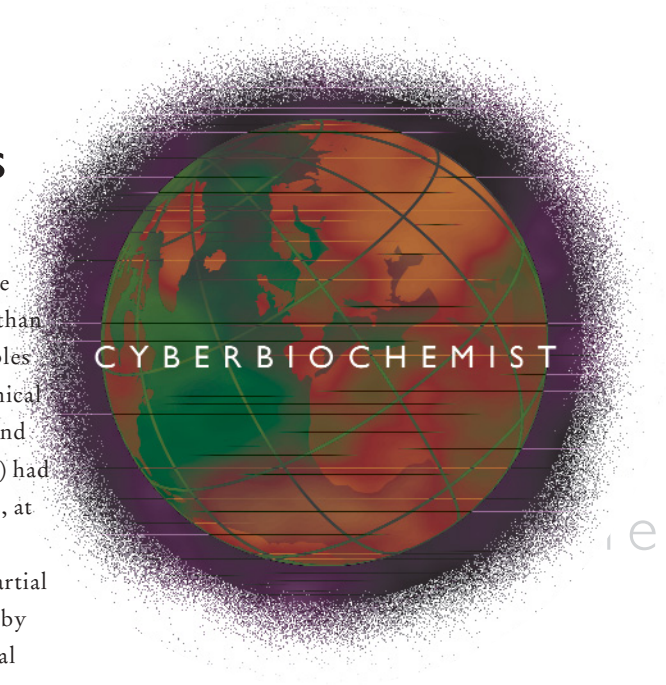
The beginnings of bioinformatics



by Clare Sansom
(Birkbeck College, London, UK)

When did bioinformatics begin? Searching PubMed with the keyword 'bioinformatics' currently results in over 13 000 records, the first being from 1990. However, bioinformatics, if defined as 'that field of science in which biology, computer science and information technology merge into a single discipline', to quote the National Center for Biotechnology Information (NCBI), pre-dates any use of that term by at least a few decades. But it seems that no-one can decide when it started. The NCBI's 'bioinformatics timeline' begins in 1962, with Pauling's theory of molecular evolution derived from the genetic code. Some commentators take the story back much further, even to the Austrian monk, Gregory Mendel (1823–1884), who described the theory of inheritance decades before the discovery of the gene. I, however, tend to agree with David Eisenberg from the University of California, Los Angeles, who, in a keynote address at a Royal Society meeting in April 2005, proposed that the key discovery was the sequencing of insulin. It is particularly appropriate to discuss this here, in the *Biochemical Journal*'s centenary year, as several of the papers describing this crucial work were published in that journal.

Bovine insulin was the first protein to be completely sequenced and this sequence was obtained by Fred Sanger and his colleagues at the University of Cambridge. The protein was first split into two chains by oxidization to remove the disulfide bonds, and sequences of the two chains were published in the *Biochemical Journal*[1–3]. The first chain was sequenced less than a decade after the basic principles of protein chemistry (the chemical structures of the amino acids and the nature of the peptide bond) had been deduced. Sequencing was, at the time, a very complicated process involving a series of partial hydrolysis reactions followed by identification of the N-terminal residue of each fragment using Sanger's own method, described only as 'the dinitrophenyl method'. This method was not explained or even referenced in the papers. This work led to Sanger being awarded the first of his two Nobel Prizes for Chemistry (in 1958) for determining the sequence of the 31 amino acids in one [bovine insulin] chain and the 20 amino acids in the other.

This first determination of a complete protein sequence sparked a growing interest in the relationship between protein structure and function, a topic that is, of course, still of immense interest to 21st Century biochemists and bioinformaticians. However, a slightly later paper by Sanger and his colleagues, also published in the *Biochemical Journal*[4], may have done even more to stimulate the idea that information theory could be used to study biological problems. This paper described the determination of two more sequences — those of pig and sheep insulin. Sanger was among those researchers who had previously shown that all mammalian insulins showed the same biochemical and immunological behaviour, and crystallized in the same form, but that there were slight differences between them in chemical (i.e. amino acid) composition. The methods employed to locate these differences were the same as those used with the bovine sequence: oxidation of the disulfide bonds to divide the insulin molecules into two chains, followed by partial hydrolysis and identification of the N-terminal residue of each fragment. They found that the longer chain was identical in each of the three sequences, and so concentrated on the shorter chain (referred to as the 'glycyl chain' as its N-terminal residue is glycine). The only differences between the three sequences were found to be in the three residues in positions 8–10 in that shorter chain: Ala-Ser-Val in bovine insulin, Thr-Ser-Ile in the pig sequence and Ala-Gly-Val in the sheep sequence.

This discovery showed for the first time that there were significant similarities in amino-acid sequence between proteins that could be assumed to be homologous (evolutionarily related). During the early 1960s, researchers including Emile Zuckerkandl and the Nobel Laureate Linus Pauling began exploring the idea that analysis of

these small differences at a molecular level could help explain evolutionary relationships. This insight led directly to the development of sequence alignment methods, which are still the most widely used bioinformatics tools.

Another scientist whose name is synonymous with the early days of bioinformatics is Margaret Dayhoff. Working in the late 1970s, she studied the sequence alignments of the handful of protein families then available and encoded the amino-acid substitution patterns she observed into the PAM (percent accepted mutations) matrices, which are still sometimes used in protein sequence alignment today. She also invented the single-letter code for amino acids, without which sequence alignments would be almost impossible to read. David Lipman, director of the NCBI since 1989, has called Dayhoff "the mother and father of bioinformatics"[5]. If this is that case, I would like to propose Sanger as its grandfather.

References

1. Sanger, F. and Tuppy, H. (1951). *Biochem. J.* **49**, 463–481
2. Sanger, F. and Thompson, E.O.P. (1953) *Biochem. J.* **53**, 353–366
3. Sanger, F. and Thompson, E.O.P. (1953) *Biochem. J.* **53**, 366–374
4. Brown, H., Sanger, F. and Kitai, R. (1955) *Biochem. J.* **60**, 556–565
5. http://en.wikipedia.org/wiki/Margaret_Oakley_Dayhoff

Clare Sansom

c.sansom@mail.cryst.bbk.ac.uk

Best of the Web

LabLit.com is a website for the discussion of science in fiction and in fact. Founded as a corrective to the misleading images of science and scientists in novels, drama and film, it has grown to become a valued resource for scientists and non-scientists.

Edited by Jennifer Rohn, the site has essays about how laboratories work and how they are portrayed in fiction (emphatically not the same thing), original science-related short stories, serialized novels, poetry, jokes, comment, reviews and artwork. It also has interviews with scientists, journalists, novelists, playwrights and artists (e.g. Nicholas Harberd, Ben Goldacre, Ann Lackie, Sidney Perkowitz and Lizzie Burns).

The site gives short summaries of books in the 'LabLit' field, e.g. *Brazzaville Beach* by William Boyd 'mathematics meets malign chimps — Jane Goodall with a twist' and *Mendel's Dwarf* by Simon Mawer 'a megalomaniac achondroplasiac geneticist studies his own disease'. There are also interactive parts: competitions (Best laboratory Haiku), surveys "What's the biggest plus about being a scientist?" and over a dozen forums (the one on Science and Art asks: "Intriguing juxtaposition or pretentious clap-trap?"). This makes it sound like a crowded site, but it is well-designed and easy to navigate.

LabLit welcomes 'quality material' from its visitors, although they should contact the editor with their proposals before they begin typing.

LabLit.com goes about its business with a light touch and some panache. It is not as cross and didactic as Insultingly Stupid Movie Physics (www.intuitor.com/moviephysics/), for instance, but it does have its share of the earnest and the solipsist. There is a recent trend towards the, to put it politely, fringes of science. However, there is always something interesting to read here. My current favourite is "What can nuking zombies tell us about infectious disease control?".

Mark Burgess (Executive Editor)