THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Mapping and sequencing information: the social context for the genomics revolution

OPEN ACCESS

THIS IS AN ADVANCED DRAFT OF A PUBLISHED

PAPER. REFERENCES AND QUOTATIONS SHOULD

<u>ALWAYS</u> BE MADE TO THE PUBLISHED VERSION,

WHICH CAN BE FOUND AT:

Mapping and sequencing information: the social context for the genomics revolution

Miguel García-Sancho

Centre for the History of Science

Imperial College, London

Current email: *Miguel.gsancho@ed.ac.uk*

In 1983, after devoting some eight years of his life to the description of how a nematode worm develops from an embryo into an adult, molecular biologist John Sulston embarked on a remarkably different project: he decided to map the worm's genome. Sulston's impulsive desire to characterise this creature's DNA from start to finish offers only a partial explanation for this transition. Instead, a close examination of the wider social context for this "moment" in molecular biology gives a more rewarding explanation of Sulston's intellectual leap. This reveals a world in which biotechnology gradually adapted to and integrated into an "information society" increasingly dependent on the creation, distribution and manipulation of information. The application of computing to DNA during the first half of the 1980s was crucial for this integration, fostering the emergence of genomics and ultimately the Human Genome Project.

Information processing and biological technologies

The mid 1980s witnessed an explosion of interest in mapping and sequencing DNA (see box). Organisms like yeast, *Escherichia coli* and the nematode worm *Caenorhabditis*

elegans all became the focus for this effort to locate genes to a particular part of a chromosome and determine their nucleotide structure [1]. The biographies of the scientists involved and several popular accounts often cite the emergence of new technologies as an explanation for this new direction in molecular biology, failing to address the social conditions in which those technologies arose [2,3]. In fact, it is helpful to take a look at how these new DNA mapping and sequencing tools were a product of both biotechnology and "information society", two distinct social, technical and economic forms that converged throughout the 1980s.

The decade before, the desire of individual professionals and laymen –rather than companies- to become computer customers had coincided with the development of the microprocessor and digital networks to miniaturise and interconnect the machines. This had resulted in the personal computer – Altair in 1975, Apple II in 1977 and IBM's PC in 1981 – and user-friendly software manufactured, in particular, by Xerox and Microsoft. With these key developments, information processing began to gain significant economic and cultural currency.

Sociologist Manuel Castells argues that this tendency derived in "information society", one in which "information generation, processing and transmission" became "the fundamental sources of productivity and power". One of the key features of this new order, claims Castells, is "the networking logic of its basic structure". The countries, markets and social groups in an information society tend to be electronically interconnected [4].

Also in the 1970s, biologists were looking for ways to explore and understand DNA in greater detail. They developed instruments to cut, clone and analyse genetic material, among them the physical mapping and sequencing techniques. At the end of the decade and during the 1980s, several companies – mainly university start-ups – began

commercialising products based on these technologies and, in 1986, the US firm Applied Biosystems released the first automatic DNA sequencing machine onto the market.

Scholars have coined this landscape the "biotechnology era", one characterized by the high expectations of scientists in the power of these new techniques but also by public concern over the ethical implications of developments like genetic modification and cloning [5].

The information society and biotechnology era both emerged in a similar social context. During the 1970s, the developed world experienced a growing economic crisis owing to the escalating price of oil and the increasing instability of the welfare state. Companies and governments sought new areas of investment and found them in the new information and biotechnology markets.

But even at the end of the decade, these two domains were still clearly distinct from one another. In 1979, a report by the Commission of the European Communities cited "the technology of information" and "biotechnology" as two of just three key fields "in which the potential for technological innovation and the consequent changes in the economic, social and political organisation of our societies" would be "most significant" in the next 20 years. Yet the text dealt with information society and biotechnology separately, not only treating them as genuinely independent phenomena in space but also in time. Whereas the information revolution was likely to concern "the next ten years", "bio-industry" would take place "the latest at the beginning of the twenty-first century" [6].

Sulston's research during this period was removed from both the information and the new biological technologies. His investigations between the mid-1970s and 1983 focused exclusively on the development of the worm *C. elegans,* including systematic microscopic work to trace the divisions and movements of every single cell from the embryo to the adult nematode [7] (Figure 1). Despite early attempts to introduce

computers to his group of researchers [8], this did not occur until the mid-1980s when the worm's genome became the practice ground for new DNA technologies.

Bioinformatics: a crucial connection

The institution in which Sulston was based – the Laboratory of Molecular Biology in Cambridge, UK (LMB) – had been using computers since the 1950s, not for work into *C. elegans* or DNA, but to automate analysis of x-ray crystallographic data and reveal protein structure [9]. It was only in the 1970s, when another group at the LMB leaded by Fred Sanger came up with ways to sequence DNA, that computers became associated with this molecule.

Sanger sought to create programs that would store and handle the DNA sequences generated by his group. For this purpose, he brought in Rodger Staden, a mathematical physicist working at the crystallography department of the LMB. The purpose of Staden's first software programmes – designed between 1977 and 1980 – was to order and store data from DNA gels, the sequence outputs derived from Sanger's techniques. Researchers could then save the sequences on a magnetic disk as text files, which may be edited, searched and compared at a later stage [10] (Figure 2).

In 1982, Staden introduced an algorithm to ease sequence assembly. Sanger's methods entailed breaking the molecule into overlapping pieces, so to reconstruct them it was necessary to match the sequences at the edges of each fragment. Staden's algorithm employed the same strategy as "hash-coding", a technique lifted directly from word-processing software, which allowed the user to locate a sequence of common letters in a mass of text. So, for example, a search for *pupp** would turn up both *puppy* and *puppet*.

Clearly then, Straden saw the DNA sequence as a text, an analogy he made explicit in the following remark:

The sequence can be considered as a series of 7 character words that overlap each of their neighbours by 6 characters. Using an alphabet of 4 [the number of DNA nucleotides] it is possible to make 16,348 (4 to the power of 7) different words of length 7. If we have a table of 16,348 positions we can record whether or not each of the possible words of length 7 are present in the consensus [assembled sequence] (…). In order to compare a gel reading [DNA fragment] with the consensus we simply look in the table to see if any of the gel's 7 character words are present. [11]

Managing texts was a key problem for those involved in the emergent computer industry. By borrowing a tool from word-processing software, Staden began to break down the barriers between information society and biotechnology.

From DNA to data

The challenge Sanger and Staden had faced of rearranging DNA fragments into a coherent sequence was similar to that which confronted Sulston in his effort to map genes –also DNA fragments- to a particular place in the worm's genome. It therefore made sense for Sulston to borrow and adapt Staden's sequencing software to meet his own mapping needs. This he did from the outset. Hence, the use of these sequencing programs with their particular text-processing features made the reduction of DNA to information the main goal of the *C. elegans* genome-mapping project.

Indeed, in the first paper on the project, published in 1986, Sulston and his colleagues described the mapping method – also known as the "fingerprinting" technique – as a

procedure "for digital characterisation and comparison of DNA fragments". They broke the DNA into fragments and treated each fragment as "data" which was stored "in a computer data base". The computer detected overlaps between the fragments and used them to construct a physical map of the genome [12] (Figure 3). Evidently, mapping was little more than an exercise in handling and processing information.

This informational shift had a huge impact on the way that researchers across the world worked together. Since the mid-1970s, the investigations on *C. elegans* had been simultaneously conducted at various institutions, mainly in the United States but coordinated by the LMB. The emergence of the mapping project resulted in the LMB producing and administering the computer database, first alone and since the late 80s in cooperation with other researchers at Washington University in Saint Louis leaded by Robert Waterston. This increased the dominance of both institutions over other centres carrying investigations on *C. elegans.*

From the mid-1980s until the publication of the complete *C. elegans* map and sequence in the late 90s, any laboratory working on a nematode gene had to send the corresponding DNA fragment to the LMB or Washington University to identify its location within the entire genome. Sulston and Waterston's groups benefited from these incoming data –i.e. DNA fragments to be located- and, in return, sent out a physical map of the region in which the new fragment was contained. The early correspondence between the LMB and the other centres working on the worm illustrates how this exchange of information was conducted (Figure 4). Sulston's team referred to it as "genomic communication" in its 1986 paper [13].

The LMB, thanks to its mastery in the new DNA technologies, had, therefore, achieved two of the key requirements for leadership in information society. It had, firstly, founded its power position on the control of the mapping information produced through the application of computers to DNA. It had, secondly, exerted such dominancy –first individually and then with Washington University- within a network of interconnected laboratories. With the generalisation of Internet in the late 80s, such network turned, additionally, increasingly articulated through electric wires. The web considerably eased the access of researchers to *C. elegans* mapping information, but did not erode the dominancy of Sulston and Waterston's groups as the sources of the data.

The emergence of genomics and the human genome

This cross-over between information society and biotechnology was not exclusive to Cambridge. Elsewhere, large-scale mapping and sequencing projects were under way, most notably the Human Genome Project (HGP). These initiatives led to the configuration of a new discipline – "genomics" – engaged with the application of DNA mapping and sequencing to different organisms. The term was coined in 1986 by geneticist Thomas H. Roderick and his colleagues Victor McKusick and Frank Ruddle. One year later, McKusick and Ruddle went on to found the journal *Genomics*. In the first issue, they defined the discipline as resulting from "a marriage of molecular and cell biology with classical genetics", being also "fostered by computational science". Researchers in this emergent field should be "competent in constructing and interpreting" various types of "maps" and interested in "learning their biological significance" for "development and disease". In the journal's editorial, McKusick and Ruddle made it abundantly clear they saw DNA as the data of this new discipline:

Mapping all expressed genes...regardless of whether their function is known, sequencing these genes together with their introns [non-coding regions], and sequencing out from these is seen by many as 'the way to go'. The ultimate map, the sequence, is seen as a rosetta stone from which the complexities of gene expression in development can be translated and the genetic mechanisms of disease interpreted. For the newly developing discipline of mapping / sequencing (including analysis of the information) we have adopted the term GENOMICS" [14].

Genomics embodied the idea of reducing the genes to data that would yield scientific knowledge and power. Genetic information had acquired a scientific value and the new discipline was ready to exploit it. Indeed, two of the advertisements included in *Genomics'* first issue – announcing a new bioinformatics company and a journal devoted to publishing DNA sequences – reflect how researchers then perceived the significance of DNA's digital output (Figure 5).

There was another reason to get excited about genomics: it began to attract serious funding from public and private institutions alike. Indeed, as soon as genomics emerged as a formal discipline, so scientists in the US began to discuss the feasibility of the HGP and seek money for it [15]. Other countries were quick to follow. In Britain in 1986, Sydney Brenner – the first person to propose *C. elegans* as a model organism– applied for funding from the UK's Medical Research Council for a physical map of the human genome.

Brenner's map would require scaling-up the techniques used on the nematode by a factor of 50 – the difference in size between the *C. elegans* and human genomes. His proposal also called for a centralised laboratory with a group of dedicated technicians to perform

the task. According to the grant application, "the most important product of the work" would be the "information about the human genes", which, Brenner claimed, would bring benefits "to medical research and practice". The same mapping technologies may also be applied to "industrial microorganisms and plants" and the resulting "information could become proprietary", he suggested [16].

The UK government approved Brenner's application, giving rise to the British Human Genome Mapping Project, which began in 1989. This coincided with a number of public human genome mapping and sequencing enterprises in the United States, Germany and France, among other countries. The national initiatives progressively converged and gave rise to a concerted international effort during the 90s, which is what we familiarly refer now as the Human Genome Project (HGP). Celera Genomics, a company leaded by Craig Venter, launched a private enterprise in 1998, forcing the HGP to increase its already multi-millionaire budget. Both initiatives derived in published human DNA sequences in 2001 [17].

Limitations of informational thinking

The emergent large-scale mapping and sequencing projects were not without their critics. In 1992, shortly after the HGP's launch, philosophers of biology Alfred Tauber and Sahotra Sarkar weighed into the debate. Investing such vast resources to analyse whole genomes was unjustifiable, they argued, since there was no "practicable way to begin characterising biological function…from sequence information alone" [18]. Such criticism questioned the informational understanding of DNA on which genomics was

based. According to Tauber and Sarkar, the nucleotide text of this molecule, which the HGP sought to obtain, would be insufficient to deduce how genes worked.

As the HGP progressed, so this concern began to spread. Researchers increasingly revealed the complexities of gene regulation and protein folding, consequently eroding the belief that DNA sequence was they key to understanding disease and other phenomena. When the draft sequence of the human genome was published in 2001, it was clear that the HGP's expectations of knowledge and health were more than a little optimistic. A new concensus with an emphasis on biological factors beyond sequence information began to emerge in the scientific community.

Interestingly, this shift in scientific opinion over the last ten years has been accompanied by criticism of information society. Historian of technology David Edgerton considers the term a "fashionable, and misleading, way of saying little more than service industries now account for very large proportions of GDP and employment". And, he claims, these industries are not necessarily based on information or other immaterial entities. This is evident from "the sheer bulk of the things associated with them, the unprecedented weight of stuff in the shops, the piles of paper in any office, not to mention the proliferation of computers, fax machines, and Xeroxes" [19]. Edgerton's critique suggests that there is a material counterpart, which should be taken into account when analysing the power of information.

The challenges to so-called information society have coincided with new explanatory models of genetic action. Systems biology, a new way of thinking that has emerged since the late 90s, proposes that DNA is in continuous interaction with other cell components

and external environmental inputs. This suggests that "emergent properties" not deducible from isolated parts can only be understood by taking into account all the elements in a system [20]. Systems biology offers an alternative vision in which the DNA sequence does not offer a direct route to biological meaning, but is part of a more complex coordinated process.

The simultaneous emergence of systems biology and criticism of information society suggests a new social and scientific paradigm for the current post-genomics era. Natural scientists, on the one hand, are beginning to think of cellular function in terms of complex interactions between biological molecules rather than just a product of the DNA sequence. Social scientists, on the other, consider that the material basis of society is as important as information (or any other abstraction) for explaining relationships between knowledge and power. Both tendencies, if converging, are likely to define a new research space in which the genetic text will at least be placed into its material context.

The two engines of the DNA race

The new line of biological research emerging in the mid 80s implied two types of activities over DNA: mapping and sequencing. Both were directed toward the structure of this molecule, but derived in different sorts of descriptions. Mapping yielded an ordered set of DNA fragments normally prepared for sequencing, i.e. determination of the nucleotides characterising each fragment (figure 6). The connection between these techniques derived in controversies among the actors involved in the Human Genome Project.

Mapping was normally applied to large genomes as that of *C. elegans.* Various samples of the DNA to be mapped were cut into thousands of fragments and those fragments, instead of sequenced, ordered in a set covering the genome from beginning to end –what is called a physical map. The ordering was achieved through finding overlaps between the fragments of the different samples of the DNA (figure 4).

The ordered fragments of the map may, afterwards, be sequenced with a similar technique. The outcome, in this case, was not a set of DNA pieces, but the different nucleotides –adenine, cytosine, guanine and thymine- integrating each piece (figure 2). Apart of sequencing, the physical map could be applied to other biomedical endeavours, as finding the genes responsible for certain diseases.

The necessity of mapping large genomes before sequencing was questioned in 1998, when the company Celera Genomics proposed a sequencing technique not requiring a previous ordering of the fragments. Such method was aimed to achieve the human genome sequence faster than the competing consortium of public laboratories (see main text). The release of Celera's method triggered a debate on which was the most reliable technique to tackle DNA [21].

-Figures:

Figure 1. Pictures of *C. elegans* from which John Sulston deduced the identity and movements of cells from a nematode embryo to a nematode adult. *Developmental Biology* (1977).

Figure 2. The output of one of the programmes of Rodger Staden. *Nucleic Acids Research* (1980).

Figure 3. The "fingerprinting" mapping technique used at the beginning of *C. elegans* project: (1) the DNA molecule is broken with enzymes; (2) the fragments are cut again and separated according to size; (3) the position of each sub-fragment is entered into a computer, which detects where sequences overlap; and (4) the computer constructs a physical map of the genome.

Figure 4. Correspondence between the Massachusetts Institute of Technology and the Laboratory of Molecular Biology in Cambridge (UK) between 1985 and 1986. *Laboratory of Molecular Biology Storage*.

Figure 5. One of the advertisements in the first issue of the journal *Genomics* (1987).

Figure 6: Difference between mapping and sequencing. Whereas mapping (left) implies reducing the genome to an ordered set of unsequenced DNA fragments, sequencing (right) yields the nucleotide structure of the molecule.

-Acknowledgements:

-References:

[1] Olson, M. et al (1986) Random-clone strategy for genomic restriction mapping in yeast. *Proceedings of the National Academy of Sciences*, 83, Kohara, Y. et al (1987) The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell,* 50 and Coulson A., Sulston J. et al (1986) Toward a physical map of the genome of the nematode *C. Elegans*. *Proceedings of the National Academy of Sciences*, 83.

[2] In the scientific side see Sulston J. (2002) *The Common Thread: A Story of Science, Politics, Ethics and the Humane Genome* (London: Bantam) and Brenner, S. (2001) *My Life in Science* (London: BioMed), especially ch. 8. In the popular science one, Cook-Deegan R. (1994) *The Gene Wars: Science, Politics and the Human Genome* (London: W.W. Norton and Company), chs. 2-4 and Wills C. (1991) *Exons, Introns and Talking Genes: The Science behind the Human Genome, Project* (Basic Books).

[3] Two remarkable exceptions to the above referred exclusively technological accounts are de Chadarevian S. (2004) Mapping the worm's genome: tools, networks, patronage. H.J. Rheinberger and J.P. Gaudillière (eds.) (2004) *From Molecular Genetics to Genomics: The Mapping Cultures of Twentieth Century Biology* (Routledge) and Ankeny R. (2001) The natural history of *C. Elegans* research. *Nature Review Genetics*, 2.

[4] Quotes from Castells, M. (1996) *The Information Age: Economy, Society and Culture* (Blackwell), vol. 1, p. 21. On Information Society see also Parker, E.B. (1973) Implications of new information technology. *Public Opinion Quarterly,* 37(4) and Quah

D.T. Increasingly weightless economies. *Bank of England Quarterly Bulletin,* February 1997. On the history of computing see Ceruzzi P. (1998) *A History of Modern Computing* (MIT), chs. 6-9.

[5] Bud, R. (1993) *The Uses of Life: A History of Biotechnology* (Cambridge), chs. 7-9; Rabinow P. (1996) *Making PCR: A Story of Biotechnology* (Chicago), chs. 1-2; Kenney M. (1986) *Biotechnology: The University-Industrial Complex* (Yale).

[6] Godet M. and Ruyssen O. (1979) *The Old World and the New Technologies* (Commission of the European Communities), ch. 5, quotes from pp. 115 and 125.

[7] Sulston J. and Horvitz, R. (1977) Post-embryonic cell lineages of the nematode *C. Elegans*. *Developmental Biology*, 56 and Sulston J. et al (1983) The embryonic cell lineage of the nematode *C. Elegans*. *Developmental Biology*, 100.

[8] Sydney Brenner and John White, members of Sulston's group, attempted without success to apply the computer to their research line, aimed to characterise every single neuron in *C. elegans* nervous system and to describe their connections. Sulston op. cit. 2002, pp. 28-29; White, J. (1974) *Computer Aided Reconstruction of the Nervous System of C. elegans* (PhD dissertation, University of Cambridge, UK).

[9] de Chadarevian, S. (2002) *Designs for Life: Molecular Biology after World War II* (Cambridge), ch. 4; Bennet, J.M. and Kendrew, J. (1952) The computation of Fourier syntheses with a digital electronic calculating machine. *Acta Crystallographica*, 5.

[10] Staden R. (1977) Sequence data handling by computer. *Nucleic Acids Research*, 4 (11); id. (1979) A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, 6 (7); id. (1980) A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Research*, 8 (16).

[11] Staden R. (1982) Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. *Nucleic Acids Research*, 10 (15). The text-processing algorithm had previously been introduced in sequencing by the French researchers Jean Pierre Dumas and Jacques Ninio: Dumas J.P. and Ninio J. (1982) Efficient algorithms for folding and comparing nucleic acid sequences. *Nucleic Acids Research*, 10 (1).

[12] Coulson A., Sulston J. et al op. cit., quotes in p. 7821 and 25. See also Sulston J., Coulson A. et al (1988) Software for genome mapping by fingerprinting techniques. *Computer Applications in the Biological Sciences*, 4 (1). On the digitalisation of the DNA data see Staden R. (1984) A computer program to enter gel reading data into a computer. *Nucleic Acids Research*, 12 (1).

[13] Coulson A. and Sulston J. op. cit. For more details about the international exchanges among the worm's community see Sulston J. op. cit. and de Chadarevian S. op. cit.

[14] McKusick, V. and Riddle, F. (1987) Editorial: a new discipline, a new name, a new journal. *Genomics*, 1, quote p. 1. On the coinage of the term *genomics* see Kuska B. (1998) Beer, Bethesda and biology: how genomics came into being. *Journal of the National Cancer Institute*, 90 (2).

[15] Sinsheimer R. (1989) The Santa Cruz Workshop, May 1985. *Genomics* 5 (4).

[16] Brenner S. (1986) A physical map of the human genome, quote p. 1; id. (1987) The human genome, quote p. 5. Both held at the UK National Archives, folder nº FD 23/3441.

[17] For more details about the transition of the human genome initiatives see Cook-Deegan op. cit., especially Part Two. The public and private sequences are available at International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, 409, pp. 860-921 and Venter C. et al (2001) The sequence of the human genome. *Science*, 291, pp. 1304-51.

[18] Tauber A. and Sarkar S. (1992) The Human Genome Project: has blind reductionism gone so far? *Perspectives in Biology and Medicine*, 35 (2), quote p. 223.

[19] Edgerton (upcoming) *The Shock of the Old. Technology and Global History since 1900* (Oxford), quote pp. 96-97. For other critiques of Information Society see Winston B. (1998) *Media Technology and Society: A History from the Telegraph to the Internet* (Routledge) and Webster F. (1997) Is this the Information Age? Towards a critique of Manuel Castells. *City,* 8. Reprinted in id. And Dimitriou B. (eds.) *Manuel Castells* (SAGE Publications), vol. III.

[20] One of the main advocates of Systems Biology –and of the former informational understanding of DNA- is Leroy Hood, father of the Human Genome Project and now heading the Institute for Systems Biology in Seattle. See Hood L. (2004) The networks of

life. Schrödinger Lecture, delivered at Imperial College, London. Hood L. et al (upcoming) *Biological Information and the Emergence of Systems Biology* (Roberts & Co). A group at the University of Pittsburgh is also currently researching on a new definition of the concept of gene in the light of the increasing limitations of DNA sequences to account for its properties http://www.pitt.edu/~kstotz/genes/genes.html

[21] For more details about this debate see Bostanci A. (2004) Sequencing human genomes. J.P. Gaudillière and H.J. Rheinberger (eds.) *From Molecular Genetics to Genomics: The Mapping Cultures of Twentieth Century Biology* (London: Routledge).