Genomics: Halfway between Biology and Computing

Roderic Guigó

Institut Municipal d'Investigació Mèdica
Centre de Regulació Genòmica
Universitat Pompeu Fabra

Fifty years after the discovery of the structure of the DNA, this year has witnessed the completion of the sequence of the Human Genome. Fifty years of unprecedent technological advances, best exemplified in two technologies that have had profound impact in our lives: Molecular Biology and Informatics.

Molecular biology and informatics were born around the same time. In 1953 Watson and Crick published their landmark paper describing DNA structure, and Sanger deciphered, for the first time, the amino acid sequence of a protein; only a few years before the first digital computers had came into existence.

Almost a decade was needed after the discovery of the DNA structure, to start understanding the mechanisms by means of which DNA plays its function. Experiments by Koranna, Brenner and others in the early 60s made possible the decipher the so-called "genetic code", the set of instructions by means of which the nucleotide sequence of the DNA specifies the amino acid sequence of the proteins. The discovery is contemporary to the invention of FORTRAN, the first popular programming language, that facilitate enormously the use of computers. It may not be by chance, that the basic terminology of Molecular Genetics, established at that time, is full of terms which a strong computational flavor: translation, transcription, code, message, ...

In the mid 60s, molecular biologists and biochemists have deciphered the amino acid sequence of dozens of proteins, which Margaret Dayhoff and co-workers have started to compile. At the same time, transistors were substituting vacuum tubes in computers, which in turn, by the end of the decade were substituted by integrated circuits. Computers became smaller, faster and more affordable. They entered universities and research institutes. Computers proved crucial to analyze Dayhoff's sequence compilation. By comparing sequences of different by related proteins, it became clear that the amino acid sequence reflects the function and the history of the proteins. The concept of sequence alignment (see Figure 1) was formulated, and Needleman and Wunsch in the early seventies implemented the first dynamic programming algorithm to automatically obtain the optimal alignment between two sequences; an algorithm that was latter extended by Smith and Waterman.

During the 70s the number of known protein sequences did not stop growing, but it was not until the mid 70s, when Sanger and coworkers developed their sequencing method, that the first DNA sequences were obtained. It was also around these years, that American scientists created Internet. At that time, two unrelated advances, but that together made possible, twenty five years latter, the sequencing of the human genome.

DNA sequencing progressed rapidly, and by the early 80s traditional archiving methods did not suffice to cope with the amount of data being generated at molecular biology laboratories around the world. Electronic databases to store known nucleic acid sequences were then created at Los Alamos National Laboratory in the

```
>sp|P01128|TSIS_SMSAV TRANSFORMING PROTEIN P28-SIS
Length = 226 Score =  140 bits (350), Expect = 2e-33
Identities = 75/161 (46%), Positives = 100/161 (61%), Gaps = 11/161 (6%)

Query: 25  IPREVIERLARSQIHSIRDLQRLLEIDSVGSEDSLDTSLRAHGVHATKHVPEKRPLPIRR 84
           IP E+ + L+   I S  DLQRLL+ DS G ED  +  L     H+   +        R
Sbjct: 10  IPEELYKMLSGHSIRSFDDLQRLLQGDS-GKEDGAELDLNMTRSHSGGELESLA----RG 64

Query: 85  KRSI------EEAVPAVCKTRTVIYEIPRSQVDPTSANFLIWPPCVEVKRCTGCCNTSSV 138
           KRS+      E A+ A CKTRT ++EI R  +D T+ANFL+WPPCVEV+RC+GCCN  +V
Sbjct: 65  KRSLGSLSVAEPAMIAECKTRTEVFEISRRLIDRTNANFLVWPPCVEVQRCSGCCNNRNV 124

Query: 139 KCQPSRVHHRSVKVAKVEYVRKKPKLKEVQVRLEEHLECAC 179
           +C+P++V  R V+V K+E VRKKP  K+  V LE+HL C C
Sbjct: 125 QCRPTQVQLRPVQVRKIEIVRKKPIFKKATVTLEDHLACKC 165
```

Figure 1. Sequence aligment between the Platelet Derived Growth Factor ("Query") and the oncogen P28-SIS ("Subject"). Some amino acids are identical in the two sequences, indicated by the middle row, other amino acids are substituted by related amino acids (indicated by "+" in the middle row), while other amino acids are not conserved at all. In addition, some amino acids do not have the corresponding equivalent one in the other sequence (indicated by "-" in this sequence). The similarity between these two sequences, though to be functionally unrelated, was discovered by means of exhaustive sequence database searches. The sequence similarity is higly signficant according to its expected probablity (the "Expect" value in the figure), and it is strongly suggestive of related functionality.

United States (GenBank), and at the European Molecular Biology Laboratory in Heidelberg. At around the same time, IBM launched the first personal computer, the popular PC. With PC's, computers made it from centralized informatics facilities to researcher's desks.

As DNA sequences accumulated at an accelerated pace, comparisons of newly obtained sequences with known sequences already stored in the electronic databases, became the most common method to infer the function of the new sequences. As databases grew larger, however, dynamic programming based comparison methods became too slow, and a new generation of algorithms, based on hash tables, was developed. As the folk history of bioinformatics tells, Dolittle was running one of these algorithms in his PC to perform exhaustive sequence comparisons in the recently created electronic databases, when he came across the strong similarity between two apparently unrelated proteins: an oncogen and a growth factor (Figure 1). This similarity contributed substantially to our understanding of the molecular basis of abnormal cell growth during cancer progression.

Hash table based programs like FASTA and BLAST have become since then, one of the most widely used tool in Molecular Biology laboratories around the world. Indeed, the paper describing BLAST, for instance, is the most cited paper in biological sciences during the decade of the 90s.

During the 80s, the number of known nucleic acid sequences grew exponentially, and the traditional model of the relation between sequence providers, sequence databases, and sequence users had entered into a crisis. For instance, by the end of the decade, almost two years on average lapsed between the obtention of a sequence and the sequence being accessible in the database. In the early 90s, however, scientists at the CERN (European Laboratory for Particle Physics) invented the World Wide Web (WWW) over the Internet. WWW soon became the platform to efficiently address the problems of sequence data submission and access in Molecular Biology.

It was also in the early nineties that the Human Genome project started officially. By that time, the intimate relation between Biology and Computation had already became evident. Indeed, genome projects are essentially informatics projects. For instance, more than half of the 224 scientists that authored the paper describing the recently sequenced, mouse genome, contributed only to the computational analysis of the sequence. Internet has become the virtual laboratory in which genome research takes place. Sites like ENSEMBL (*http://www.ensembl.org*, figure 2) in Europe, and the GENOME BROWSER (*htpp://genome.ucsd.edu*) in the United States store, create, and distribute genome information.

In summary, with genomics biology has become a discipline highly dependent on computers. Because of the complexity of their complexity and the vast amount of data that they generated, biological problems are the most computationally challenging. But the relation between biology and computation goes beyond the quantity, and relates to the "quality" of data. Indeed, the sequencing of the human genome is the confirmation of Schroedinger's intuition, who in the 1942 speculated that the chromosome had to be an aperiodic crystal made of the repetition of a small number of isomeric elements. It is the specific order (sequence) of these elements, rather than the underlying physico-chemical properties, the responsible for the functionality of the chromosomes, he claimed. In short, life is the computation of the nucleotide sequence of the genome. It is because of this basic computational nature of the biological phenoma, and not only because of the vast amounts of data that research in Biology is producing nowadays, that Biology and Computation are, and will remain, inextricably linked. ■

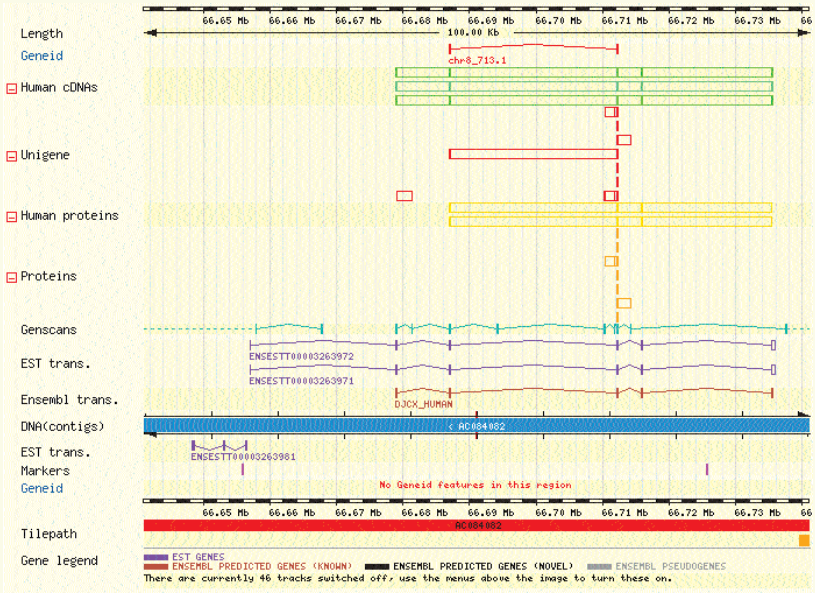

Figure 2. Screenshot of the ensembl (*www.ensembl.org*) genome browser. It displays information about the features encoded in a region of about 1 milion nucleotides from the human chromosome 8. Genome information is very dynamic, and it is provided through laboratories world wide. The ensembl browser allows, through the DAS (Distributed Annotation Server) protocol, third party annotations of the human genome over the Internet. For instance, gene prediction in the track labelled geneid are not stored at ensembl, but at our server at the IMIM. When instructed, the ensembl browser submites a request to our DAS server, which provides the requested information that together with the information provided by other DAS servers around the world is visualized through a common interface.