LECTURES FOR THE XXIst CENTURY

edited by
Bart Raymaekers

LEUVEN UNIVERSITY PRESS
2008

D/2008/1869/ 16
ISBN 978 90 5867 648 1
NUR: 74

FROM BIOINFORMATICS TO SYSTEMS BIOLOGY

BART DE MOOR

INTRODUCTION

In a famous one-page article in *Nature* in 1953, Francis Crick and James Watson described the chemical structure of DNA for the first time[1]. Since that day, scientific research in molecular biology and biotechnology has exploded. Our knowledge about the genetic and biochemical processes in the cell is increasing exponentially. We also know that the impact of applications with respect to men, animals and plants will be enormous. Here, we will describe some of the basic ingredients that characterise this explosion of knowledge on biological and biomedical systems.

We are currently also witnessing an exponential evolution of applications of information and communication technology. What today we call 'hardware' originated in the laws of electromagnetism, discovered in the late 19th century by Maxwell and others. Current day applications include our power-generating system (electricity) and wireless communication technology. The fundamentals of quantum mechanics were laid down by physicists like Einstein, Bohr, Schrödinger and Heisenberg in the first half of the 20th century. Their insights led to the invention of the transistor in 1948, the basic building block of all our computers and electronic devices today, such as lap tops, iPods, PDAs (Personal Digital Assistants), mobile phones and many more.

The spectacular growth of information technology applications is driven by 'Moore's law'[2], which says that the number of transistors on a siliconchip of one square millimetre doubles every eighteen months. This implies that our computers can contain more

[1] This important discovery, for which Crick and Watson received the Nobel Prize, was commemorated in a special issue of *Nature* on 23 January 2003 (Vol. 421).

[2] Moore's law is named after Gordon Moore, the person that started Intel (www.intel.com). In the beginning of the 1960s, at the dawn of the era of microelectronics, he predicted that the number of transistors per unit of chip surface would double every eighteen months. In the financial world, a comparable growth rate would correspond to an interest rate of 59 percent. If you would have invested €1 in 1968 at an interest rate of 59 percent, you would be €100 million richer by now (2008)!

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

and more data in their memory (think of the incredible growth rate of the World Wide Web), and that their computing power, i.e., the number of calculations they can perform per second, also doubles about every eighteen months.

What we call 'software' is based on the many mathematical discoveries and developments between 1850 and 1950, the formulation of information theory by Shannon about 50 years ago and the research in the computer sciences that started around 1950. 'Software' includes numerical algorithms, databases, transmission and computer security protocols, computer languages, etc.

The combination of hardware and software has led to the development of the World Wide Web, which in less than twenty years has grown into an incredible repository of information and databases. Our environment has literally evolved into what is called 'a small world': it only takes an average of four to six clicks with your mouse to reach any arbitrary website from anywhere in the world. This 'small world phenomenon' is one of the major drivers in what we call 'globalisation'. It also plays a major role in the advancement of science in general, and that of the biological and biomedical sciences in particular.

This article is about the synergy between molecular biology and information technology: two sciences that at first sight have nothing in common. Their symbiosis is called 'bioinformatics'. It has drastically modified the way in which we perform research in biology and biomedicine today. "Biology has become an information science", says Leroy Hood of the Institute for Systems Biology in Seattle (cf. http://www.systemsbiology.org/). In what follows we will describe the relevant ingredients of bioinformatics. We will also discuss several applications and also have a glimpse at the near future of systems biology.

WHAT IS BIOINFORMATICS?

Before we elaborate on bioinformatics, we need to highlight some basics of biology, technology and mathematics. Don't be afraid: we will keep it simple!

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

> *It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.*
> *Last sentence from the Crick & Watson article in* Nature*, 1953.*

Biology as a science has witnessed an incredible evolution over the last 50 years or so, characterised by an overwhelming amount of scientific breakthroughs and discoveries involving viruses, bacteria, plants, animals and *homo sapiens*. Before we proceed, it is necessary to make the reader familiar with some basics of molecular biology[3].

1. DNA and genomes

The human body consists of billions of cells. In the nucleus of every cell, we find chromosomes. They can be considered to be the chapters of a book, which is written in the alphabet of the DNA. The acronym DNA[4] stands for 'Deoxyribonucleic Acid'. Every DNA molecule is a linear concatenation of four genetic basic components, called nucleotides, which are indicated by the letters A (Adenine), C (Cytosine), T (Thymine) and G (Guanine). What Crick and Watson described in their 1953 paper in *Nature* was the molecular geometric structure of the DNA molecule, the famous 'double helix' model. The global molecule looks like a spiral staircase, with winding stairs that consist of pairs of molecules, with molecules A and T as one type of stair, and molecules G and C as another type. These pairs are called base-pairs. The double spiral around these stairs consists of chains of phosphate-deoxyribose-sugar polymers. The whole structure is called double-stranded DNA.

The fact that A always pairs with T (or the other way around), and C always with G is called 'complementarity'. This property is the basis of the fundamental principle by which DNA is copied during cell division, in which the steps of the staircases are split up in the middle, separating each A from its pairing T, and

[3] If you want to learn more, you can consult some splendid books, such as Griffiths (1996); Kreuzer (1996), Griffiths (1999), Brown (2002) and Karp (2002).

[4] In the recent past, already nine Nobel prizes have been awarded for discoveries related to DNA. Yet, DNA is still the subject of a lot of ongoing research (see the special issue of the *New Scientist* of 15 March 2003, entitled 'DNA, the next fifty years').

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

each G from its pairing C. Next, each of the two separated, but complementary strands are completed again: to each molecule A binds a new molecule T, to each T a new A, to each G a new C and to each C a new G. Of course, our mechanistic description here is a gross simplification of reality, but it is a starter! The complementarity in our DNA allows nature to 'double' the amount of genetic information in each division step, just like we do with a copy machine. Not only is this complementarity the basis of genetic inheritance, but it is also the basis of a lot of new technology, such as microarrays, which we will describe below.

Another recent and extremely important breakthrough is the availability of the complete DNA sequence – the genome – of an increasing number of organisms. At the start of this millennium, we witnessed the unravelling of the human genome, which was published in two path-breaking papers in *Nature* and *Science* [5] (cf. Lander, 2001; Venter, 2001)[6]. Besides the human genome, which counts about 3 billion nucleotides, the genome of many other organisms has been sequenced. We now know the genome for viruses[7], bacteria[8] (e.g., *Haemophilus influenzae*), organisms such a yeast (*Saccharomyces cerevisae*), plants such as *Arabidopsis*

[5] The fact that the genome of one person cannot be published on paper in one article is trivial: the human genome counts approximately 3 billion letters. The four letters A, C, T and G can be encoded with binary numbers of two bits (e.g., A=00, T=01, C=10, G=11). This means that for 3 billion letters, we need 6 billion bits, corresponding to a memory requirement of 750 MB.

[6] With the technology developed by Venter, the genome is shot into pieces. Using computer algorithms, these pieces are then put back together in silico. For the human genome, a supercomputer was used, counting 800 processors with 70 TB memory (a Terabyte is '2 to the power of 40' bytes, or about 1000 billion bytes, or 1000 GB (gigabyte)). One byte is eight bits.

[7] One can debate whether viruses can be considered to be 'alive', or whether they are just lifeless pockets of molecules. They can cause diseases like the flu, or AIDS. They are much smaller and simpler than bacteria and consist of genetic material covered in a mantle of proteins.
The genetic code of the SARS virus (Severe Acute Respiratory Syndrome) was unravelled in a record time of three weeks in April 2003 (cf. http://www.bcgsc.ca/bioinfo/SARS, and a slightly different sample can be found on www.cdc.gov). It consists of about 29,700 building blocks. Knowing the genome of the virus, might lead to better diagnostic tests and therapies.

[8] Bacteria cause many infections, like lung infections. Contrary to viruses, bacteria can survive sometimes during years without a host. Recently, the genome of several bacteria was unravelled. Many bacteria play an important role in food (*Streptococcus thermophilus* in the production of cheese, yoghurt, etc.). This type of bacteria has about 1,900 genes.

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

thaliana (*Nature*, 14 December 2000), rice[9], the marine diatom *Thlassiosira pseudonana*[10], the nematode worm *Caenorhabditis elegans*[11], the fruit fly *Drosophila melanogaster* (*Science*, 24 March 2000) and the mouse *Mus musculus* (*Nature*, 5 December 2002) in addition to many, many others, including several mammals.

2. Genes, amino acids and proteins, and so-called 'junk-DNA'

We have described the genome of living organisms as a linear cascade of nucleotides. In every genome, there are certain functional segments, intertwined with other ones that, at first sight, do not seem to have any function. Of the functional segments, genes are those best known. There are many possible definitions of a gene[12], but here we provide a simplified information-theoretic description. The 'words' that compose a gene are written in the DNA alphabet

[9] The genome maps of two different subspecies of rice were published in *Science* on 5 April 2002. Based on this genome information, the search for better versions of rice can begin. Rice shares a lot of common features with wheat, sorghum, etc., because they all share common ancestors. These correspondences can be found in databases like www.gramene.org. Unravelling the rice genome took 74 days with the same 'whole-genome shotgun' method that was used in the human genome project. In this technique, the genome of an organism is shot into overlapping pieces, and then pasted together using numerically intensive computer algorithms. It is estimated that rice possesses between 32,000 to 50,000 genes, and that, remarkably enough, each rice gene only codes for one single protein (which is not the case in humans).

[10] These are small but relatively important plants that serve as food for fish and absorb almost as much carbon dioxide as all tropical rain forests together. *Thalassiosira pseudonana* has 24 pairs of chromosomes, numbering 11,500 genes. Remarkably they have a skeleton made of glass (silicone dioxide, the same material from which we build our ICT chips today).

[11] This little worm only has 959 cells. It is an example of what in biotechnology is called a 'model-organism'. Many of its genes can be used as a 'model' for comparable genes in humans. The Nobel Prize of 2002 was given to three scientists who helped to unravel the genetic processes in this little worm.

[12] A gene is defined as the 'complete DNA sequence required for the syntheses of a functional polypeptide or RNA molecule'. The biochemical process that transforms the information encoded in a gene into a protein is very complicated. The process starts when an activated transcription factor enters the nucleus and then binds to the DNA. The presence of an activated transcription factor will attract RNA polymerase to start the transcription of a gene. The RNA polymerase reads the DNA sequence and then generates a single-stranded RNA molecule that is complementary to the gene that was read. In the next step messenger RNA (mRNA) is formed by splitting out so-called 'introns'.
The mRNA is then compacted and transformed into the cytoplasm where finally proteins are formed by
ribosomes (see Figure 1). This step is called 'translation'.

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

(A, C, G and T). Each 'word', called a 'codon', consists of three consecutive nucleotides. Each codon 'encodes' (describes) a different amino acid, of which there are twenty different ones in nature. Every gene starts with a start codon (typically ATG) and ends with a stop codon (which can be either TAA, TAG or TGA)[13]. Now imagine that a gene can be 'read' by some reading mechanism, which starts with the start codon, and then proceeds by reading each codon, and ends with a stop codon. Every time a codon is read, it is chemically translated into a specific amino acid, and all the amino acids generated in this way are then cascaded together to form a certain protein. These proteins are the workhorses of all biological processes. The path from the functional entities in the DNA – the genes –, to codons and amino acids, to proteins, used to be called the 'central dogma of biology'. Recently, important exceptions to this central dogma have been discovered, so we no longer think of this dogma as so universal.

The more organisms for which the genome has been sequenced, the easier we can make estimates on the number of genes in each genome. Some examples: the bacteriophage *Lambda* (genome size 5.0E+04 base pairs[14], 60 genes), *Escherichia coli* (4.6E+06 bp, 4,290 genes), yeast (12.0E+06 bp, 6,144 genes), the fruit fly *Drosophila melanogaster* (1.0E+08 bp, 13,338 genes), the worm *Caenorhabditis elegans* (1.0E+08 bp, 18,266 genes), the plant *Arabidopsis thaliana* (2.3E+08 bp, 27,000 genes) and, finally, *Homo sapiens* (3.0E+09 bp, 'only' 25,000 genes)[15].

Because a codon consists of three letters, each of which is part of an alphabet of four letters (A, C, T and G), it is easy to calculate that there are 64 different codons. However, in nature there are only twenty different amino acids. But some amino acids can be 'generated' by more than one codon. Nature is not mistaken

[13] One way to identify candidate genes in the genome is to detect so-called 'Open Reading Frames' (ORFs), that start with a start codon and end in one of the three stop codons. For ease of explanation, we do not distinguish between DNA and mRNA (in which the base T is replaced by Uracil, denoted by the letter U).

[14] The notation 5.0 E+0.4 means '5 times 10 to the power of 4', i.e., 50,000.

[15] Now that we almost have the complete DNA sequence of the human genome, a systematic study has begun to map exhaustively all genes of man. Complete maps are available now for the chromosomes 7, 14, 20, 21 and 22. Because of the immense amount of information, and the complexity, this type of research is typically done by large research consortia. Chromosome 7 of the human genome was unravelled by a team of 90 scientists from 10 countries, the findings of which were published in *Science* on 11 April 2003 (cf. www.chr7.org). Chromosome 7 numbers 158 million nucleotides and by the year 2003, 1455 genes had been identified, some of which play a role in leukaemia, autism and mucoviscidosis.

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

here, as this redundancy is the basis of a certain genetic robustness: in this way, once in a while, there can be a mutation of one single nucleotide, or there can be an error in the reading mechanism, without any noticeable effect on the resulting protein. Of course, things can also go wrong: sometimes a point mutation in the DNA is not innocent at all. This is the case with certain mono-genetic diseases, which are caused by deviations in one single gene, and which can be quite catastrophic.

Despite the finite number of only twenty amino acids, the number of possible proteins is astronomically large. For proteins that consist of concatenated amino acids, say 'L' of them, the number of different proteins is large, equal to '20 to the power of L'. For a length of L=5, in principle there are '20 to the power of 5', which is 3.2 million different proteins!

Proteins, which in essence are linear chains of amino acids, typically have a complicated three-dimensional geometrical configuration. This geometry very much determines the precise interaction with other proteins and molecules, such as binding properties, enzymatic effects, signal transduction, cell-cell communication and many other functionalities and processes in the cell. A lot of research is being carried out to predict the precise geometrical form of a protein, starting from the DNA sequence that codes for it (the so-called protein-folding problem).

Proteins glue cells together into tissue, organise these tissues in organs and, from there, compose living organisms. Amongst other things, proteins control cell division, repair damaged hereditary material, and play an important role in oxygen delivery.[16] Failures in the functioning of a protein are felt very rapidly at the level of tissues, organs and soon the general wellbeing of a patient.

The genetic code we have just described, starting from DNA sequences, to codons, to amino acids, to proteins is quasi-universal for all organisms on our planet. This offers interesting perspectives to 'synthetically' exchange certain pieces of DNA sequence between organisms (as nature has been doing 'spontaneously' over millions of years) in order to obtain certain 'improvements'.

[16] The Nobel Prize of 2004 (cf. www.nobel.se) was awarded to three scientists who discovered how all kinds of 'waste material' in our bodies are labelled for transport to 'waste-treatment factories' in our cells (the proteasomes). A typical label is a protein, called 'ubiquitin', because it is ubiquitous in living cells. The failure of some of the mechanisms in which ubiquitin is involved, can lead to cancer and other diseases.

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

In the genome of humans, mammals and plants, we also find sequences of DNA that do not code for proteins. Till quite recently, these pieces were called 'junk-DNA', but over the last couple of years we have started to realise that this name is quite inappropriate. In these non-coding areas, there are many other functional entities, such as regulatory elements and motifs, on which we will elaborate on below. These are 'switches', which can switch a gene on or off, and which can do this in a continuous way, similar to the way we can switch a light on or off, or do it continuously using a 'dimmer'. They can up- or down-regulate a gene, i.e., increase or decrease the number of mRNA it generates (and hence increase or decrease the number of corresponding proteins). These switches can also act as 'timers', i.e., they control the activity of a gene as a function of time, as we see in biorhythms or in cyclic or seasonal behaviours. Also, junk-DNA contains pseudo-genes, genes that somewhere during evolution played a role, but which are now not switched on any more. Today, a lot of research is being done to unravel the interaction between genes and regulatory elements, which are organised in so-called 'genetic networks'.

All of these biochemical reactions happen on very small scales: viruses are the size of a couple of hundred nanometres (a nanometre is one-millionth of a millimetre, or 0.000000001 m). But also the time scales of biological systems can vary widely. In ideal circumstances, the bacteria *E. coli* can divide itself in twenty minutes. This means that after eight hours, one bacteria can generate a population of '2 to the power of 24' bacteria, or 16,777,216 bacteria. Genetic clocks responsible for biorhythms can have a periodicity of 24 hours. These are just two examples of the widely varying time-scales that we find in living organisms.

TECHNOLOGICAL BREAKTHROUGHS: MICROARRAYS AND BIODATA

The complementarity of DNA that we have described also lies at the basis of an important new technology: DNA chips, or microarrays. In the path from DNA to proteins, of which we have given a simplistic description, messenger RNA (mRNA) plays an important role. mRNA contains single-stranded copies of DNA. It can migrate from the nucleus to the body of the cell, where it provides the protein-generating mechanism with the instructions it needs. This is illustrated in Figure 1. The more mRNA that is transported from the genetic epicentre, the harder the protein generating mechanism will work. The volume of mRNA molecules – and their

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

concentration – is an important indicator for the molecular biological activities of the genes, and, from there, we can recognise, in principle, the difference between health and disease.

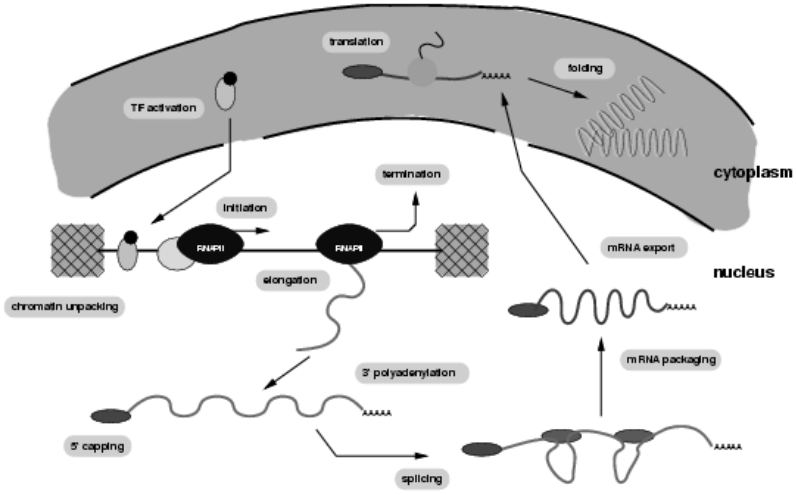

Figure 1: Representation of the different steps in going from a gene to a protein. An activated transcription factor enters the nucleus of a cell and, under certain circumstances, binds with the DNA.
Then RNA-polymerase gets involved to start transcription of the gene, into a complementary, single-stranded RNA sequence. The newly formed RNA is stabilised at both ends. In the next step, mRNA is formed by spitting out the so-called introns. The mRNA sequence folds into a more compact form that is then transported out of the nucleus. In the cytoplasm the final translation of mRNA to an amino acid sequence is performed.

One of the most spectacular technological breakthroughs in biotechnology of the last couple of years is the development of microarrays (cf. Schena, 1995; DeRisi, 1997; Lander, 1999). They have made possible the simultaneous measurement of the activity or gene expression level of thousands of genes active in a particular sample. These expression levels are a measure for the quantity of mRNA that is generated by the genes involved. Microarrays consist of a large number of spots on a small carrier surface. Each of these spots contains a string of nucleotides that is complementary to one specific single-stranded mRNA string, to which it will bind: in other words, each spot 'recognises' one specific mRNA sequence, and, at the same time, measures the quantity of that spe-

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

cific mRNA in the sample under study. Because microarrays contain thousands of spots, we can simultaneously measure the expression levels of several thousands of genes that are active or non-active in a certain sample. The speed by which microarray technology is developing implies that, pretty soon, we will be able to put all genes of a genome on one single DNA chip. While techniques to measure the expression levels of a couple of genes have been available for a while (e.g., Northern blot), the power of microarrays lies in their 'high throughput', i.e., they measure the expression levels of several thousands of genes at once! This implies that for each sample we analyse on a microarray (e.g., a sample from a biopsy of a cancer tumour), we obtain thousands of numbers, each of which quantifies the expression level of a specific gene in that specific sample. This generates a large amount of data. Suppose we have 5,000 tumour biopsies (e.g., a bio-bank), of 5,000 patients with a specific type of cancer, and that we screen those samples on a microarray for the expression levels of 10,000 pre-selected genes, and we do this every month for twelve months in a row. This will then create a database of 5,000 x 10,000 x 12 = 600,000,000 numbers; that is 600 million numbers that need to be analysed. What we mean exactly by analysis will be explained below, but the point we can make here is that this new technology of microarrays generates a lot of numbers and links the world of molecular biology with the numerical/statistical world. From this, we need to develop new types of mathematics and statistics, and also algorithms, in order to cope with these large amounts of numerical data, from which we can then try to deduce the relevant biological or biomedical information. This is exactly what bioinformatics is all about.

It should come as no surprise that the volume of biological and biomedical information on the World Wide Web is increasing exponentially. Recent estimates show that the volume of genome sequence information doubles every eighteen months (coincidence or not, this is exactly the same exponential doubling as in Moore's law). It is predicted by experts that very soon the world will produce 100 gigabytes of biological and biomedical data, *every day*!

All of these data have specific properties and features, which can briefly be summarised as follows (De Moor, 2003):

Biodata feature I: Typically, biological data are collected under difficult and nontrivial experimental circumstances, and, therefore, the measurements are not always very precise. In technical terms, the data suffer from a very bad 'signal-to-noise' ratio, i.e., they are

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

corrupted by all kinds of useless, random noise, which is quite challenging to deal with!

Biodata feature II: Biology as a science is not yet characterised by an 'axiomatic approach'; in other words, it is still very much like an 'empirical' science, where most of the 'first principles', the 'axioms', still need to be discovered. This is the reason why many of the results published in biological literature are formulated in a conditional tense, as a hypothesis. This is also the reason why there is a big need for statistics and probabilistic methodologies.

Biodata feature III: Even if biology is not (yet) axiomatic, and the overall quality of the empirical data is quite low, the qualitative, descriptive know-how of biological systems is still quite good. This requires relatively complex knowledge-representation systems that can cope with qualitative, not quite quantitative models.

Biodata feature IV: Biological systems operate on several, widely varying scales in space and time. Therefore, we need methods that can integrate and operate with information over several orders of magnitude in space and time.

Biodata feature V: Biological research on specific organisms or pathologies happens simultaneously in many (hundreds to thousands) research groups throughout the world. Therefore, the acquired scientific knowledge is distributed, not only geographically, but also over several hundreds or even thousands of websites with biological databases and with relevant publications. This requires special methods of knowledge-integration.

Biodata feature VI: Biologists typically represent biological problems and systems in a graphical manner. Therefore, they need user-friendly user-interfaces as well as graphical metaphors in order to exchange information.

BREAKTHROUGHS IN MATHEMATICS AND STATISTICS

At first sight, we think of biology and mathematics as two different, non-overlapping branches of science. However, in the 20th century, we discovered that the fundamental laws of matter, energy and information can be captured in an extremely efficient way with the language of mathematics[17]. There is no doubt that in the 21st century we will come to the same conclusion for the fundamental bio-

[17] Already Galileo Galileï emphasised how efficient mathematics is in describing the laws of nature (*'...libro della natura, scritto in carateri matematici....'*). The Nobel Prize winner Eugene Wigner called it "The unreasonable effectiveness of mathematics" (Lesk, 2000).

logical laws (Lesk, 2000). The DNA code as we have described above (albeit rather simplistically), is a clear example of this: it shows how biological systems encode information, over several thousands of years throughout generations. The way proteins interact with each other as a function of time can be described by differential equations (not that we already do this intensively, but the time that we will do so is nearby). In brief, information theory, mathematics and statistics will prove to be extremely effective in the description and modelling of biological systems.

This is no coincidence, as can be seen from some historical examples. When Gregor Mendel discovered his laws of inheritance, he was not so much inspired by biological insights, but by mere statistical inference (basically counting occurrences) (Henig, 2000). The 1940 PhD thesis of Claude Shannon was entitled *An Algebra for Theoretical Genetics.* The very same engineer created a brand new branch of mathematical statistics ten years later, called 'information theory'. The famous British mathematician Alan Turing, who during the 1930s and '40s made important contributions to computer science (the 'Turing-machine'), and who also cracked, during the second World War, the secret Enigma code of the Nazis, wrote in around 1950 a famous manuscript in which he explained the cell division of embryos ('morphogenesis') using reaction-diffusion equations. And there are many more exciting examples of the interaction between mathematics and biology. In a new scientific field, called 'biomimicry' (from the Greek, *bios*, 'life', and *mimesis*, 'to imitate') (cf. www.biomimicry.org), scientists derive inspiration from nature to find new solutions for technological problems. The underlying idea is that nature is a gigantic computer, which over the last 3.8 billion years has experimented with 'survival' strategies so as to keep only the best 'solutions'. Well known examples of biomimicry include the research that engineers do to create 'artificial neural networks', which manage to find nonlinear relationships to model and predict observations. We also use 'genetic algorithms', which are inspired by Darwin's 'survival-of-the-fittest', to solve difficult optimisation problems. Or we use sophisticated search algorithms that are inspired by the way ants communicate with each other, using pheromones. Our computer-virus-detection methods behave in very much the same way as natural immune systems. And recently, researchers developed DNA computers[18], in which the complementarity of DNA strings, explained

[18] Recently, Israeli scientists described how they managed to put about three million DNA computers into one thousandth of a millilitre of a salty solution.

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

above, is used to solve complicated combinatorial optimisation problems (Kari, 1997).

Due to Moore's law, the computing power of our computers doubles every eighteen months. Therefore, nowadays we can also implement numerical algorithms so that they can perform calculations on databases of a very large scale[19], the size of several mega/gigabytes. Many algorithms were invented more than 100 years ago, but it is only because of the technological breakthroughs in the design and manufacturing of siliconchips that we can now really use them. An algorithm[20] is a certain sequential procedure, implemented in software on a computer, to solve a given problem. As an example, we all extensively use algorithms to sort a list of names alphabetically, called sorting algorithms, which are very efficient. But our computers use many algorithms every second. A special type of algorithm is a numerical algorithm. These numerical algorithms operate on numbers and use the language and properties of numerical mathematics and linear algebra. Very simple examples are an algorithm for calculating the square root of a real number, or an algorithm for multiplying two matrices together. A more complicated algorithm is one that calculates the eigenvalues and eigenvectors of a large scale matrix. Nowadays, there are algorithms for many problems in all branches of science, and the research area of algorithmic design is a very lively one.

Bioinformatics has emerged from the interaction between these breakthroughs in biology, technology, mathematics and statistics. But increasing communication via the internet has also pro-

---

These small calculating units perform an estimated 66 billion of elementary calculation operations per second. These small computers measure the concentration of specific mRNA molecules, take certain decisions based on these concentrations and then proceed in releasing or not, certain molecules that can act as a medicine. Each of these modules – to measure (input), to diagnose (processing), to output (release medicine) – are typical parts of a computer. There is a lot of ongoing research today to synthesise such elementary computers using DNA (cf. Benenson, 2004).

[19] In the first half of the 20th century, algorithms like this were executed by 'batteries' of human calculators - in many cases women, who were called 'computers'. Each of them took a small piece of the computational problem just like today we split up a difficult calculation in many tractable pieces. The word computer in due time got transferred from its meaning of 'human' to that of 'machine'.

[20] The word algorithm derives from the name of the mathematician Mohammed ibn-Musa al-Khwarizmi, a member of the royal court in Bagdad, who lived from around 780 till 850 AD. It is in the work of Al-Khwarizmi, that the word *algebra* is used for the first time.

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

ven to be a strong catalyst for the development of this new discipline: the number and size of biological databases containing genomes of organisms, and databases with scientific insights and publications is increasing very rapidly.

APPLICATIONS

Although the number of applications is nearly unlimited, we will restrict ourselves to just two, relatively easy proof-of-principle examples that illustrate how bioinformatics has dramatically changed the way biological and biomedical research is being done.

The first application describes how several types of leukaemia can be distinguished from each other, using genetic information derived from microarrays. This example shows how physicians in the near future will use 'decision-support-tools' that assist them in determining a correct diagnosis.

In the second application, we illustrate how microarrays are used to gain new insights in biology. We will show how one can discover so-called 'regulatory elements' in the DNA.

CLINICAL APPLICATIONS IN ONCOLOGY

In recent publications, it has been shown how data generated by microarrays can be used in clinical diagnosis applications, especially for cancer research and diagnosis (oncology). Some early examples can be found in van de Vijver (2002) and van 't Veer (2002), but since then there have been many papers that describe the same idea. We now know that microarray data can be used in the diagnosis, prognosis and therapy planning of malignant cancers. In addition, since the price and selectivity of microarray devices is improving exponentially, it is expected that we will soon be applying microarrays in every day medical practice. This may then lead to a revolution in the clinical treatment of cancer.

Cancer is a process in which the genome plays a critical role. Under the influence of many potential external factors, such as radiation, viral infections, etc., mutations in certain genes can be induced, causing an uncontrolled proliferation of cells, which in due time leads to invasion and metastasis (the spreading-out of the malignant cells). These genetic mutations can also lead to the malfunctioning of other genes, even though they themselves are not modified at all. But their expression might be regulated by the product or protein of a gene that is malfunctioning. The collection of

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

abnormal gene expressions, determines the phenotype of the tumour cell. It also determines the prognosis or the reaction to a certain therapy. Measuring the gene expression levels of the genes involved in these processes leads to a better understanding of the mechanisms underlying 'carcinogenesis', the origin and growth of tumour cells. It will also lead to better therapy monitoring. There is, however, a scientific challenge. Every microarray generates thousands of numbers (one number per gene per patient at a certain instance of time). Processing so many numbers, and deducing the correct medical conclusions from them, requires advanced mathematical and statistical techniques. In order to explain the idea, we will take a relatively simple example from literature, which has often been discussed in recent years, and which illustrates very well what we have in mind.

A sample of 72 patients is taken, some of which suffer from acute lymphatic leukaemia (ALL), others from acute myeloid leukaemia (AML), and a third group from a form of acute leukaemia, called MLL leukaemia. Blood samples of these patients are then analysed with a microarray, containing about 12,600 genes (these results can be found on the internet at: http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi). The data from these microarrays can now be used for the following:

- for selecting individual genes or combinations of genes that 'characterise' one of these three forms of leukaemia;
- for performing predictions relevant to the clinical treatment;
- for discovering unknown classes of a certain disease, and identify the genes that play a role in these classes.

Let us now elaborate on these themes.

1. Feature selection

Typically, a first step in data treatment is the reduction of the mere amount of data. This is called 'reduction of dimensionality'. Our dataset under study contains 72 patients, monitored over about 12,600 genes. This is a lot of data. Therefore, it is important to reduce the information, which will contain a lot of redundant or irrelevant information, in one way or another. We only have to select those features of the data that are relevant for what we want to do, which in this case is to recognise which patients belong to which of the three classes of leukaemia, and which genes play a role in that classification. There are several methods of achieving this, but here we will only mention two by way of example.

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

2. Selection of individual genes

The simplest way is to select individual genes, the expression of which is best correlated with a specified class. This is a very simple idea, since one can expect that the expression level of most genes is irrelevant for one or all of the three classes of leukaemia, since most of the 12,600 genes probably do not play a significant role in the disease anyway. As an example, we could select the fifteen genes of which the expression differs most in one of the three classes (ALL, AML or MLL), compared to the other two (see Figure 2). Then we can just omit the other genes. This corresponds to a dimensionality reduction from 12,600 genes originally, to only 45 (3 x 15 genes; that is fifteen for each of the three classes). In this way, we could also hope to identify the genes that play a role in a certain type of leukaemia, which might help in trying to identify the origins of this specific form of cancer. In literature, one can already find many articles that attempt to identify in this way genes relevant for a certain pathogenesis. One can not only compare gene expression levels in microarrays, for different genes and different patients, but also under several different conditions, or even as a function of time. Obviously, the technology of microarrays will help us a lot in the near future in biomedical research and clinical therapy.

Figure 2: Selection of three sets of fifteen genes, the expression of which differs the most in ALL, AML or MLL (in comparison to the two other classes). Every column of this matrix represents a patient, and each row contains the gene expression levels (colour-coded) for a particular gene in all patients (Figure taken from Armstrong, 2002.)

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

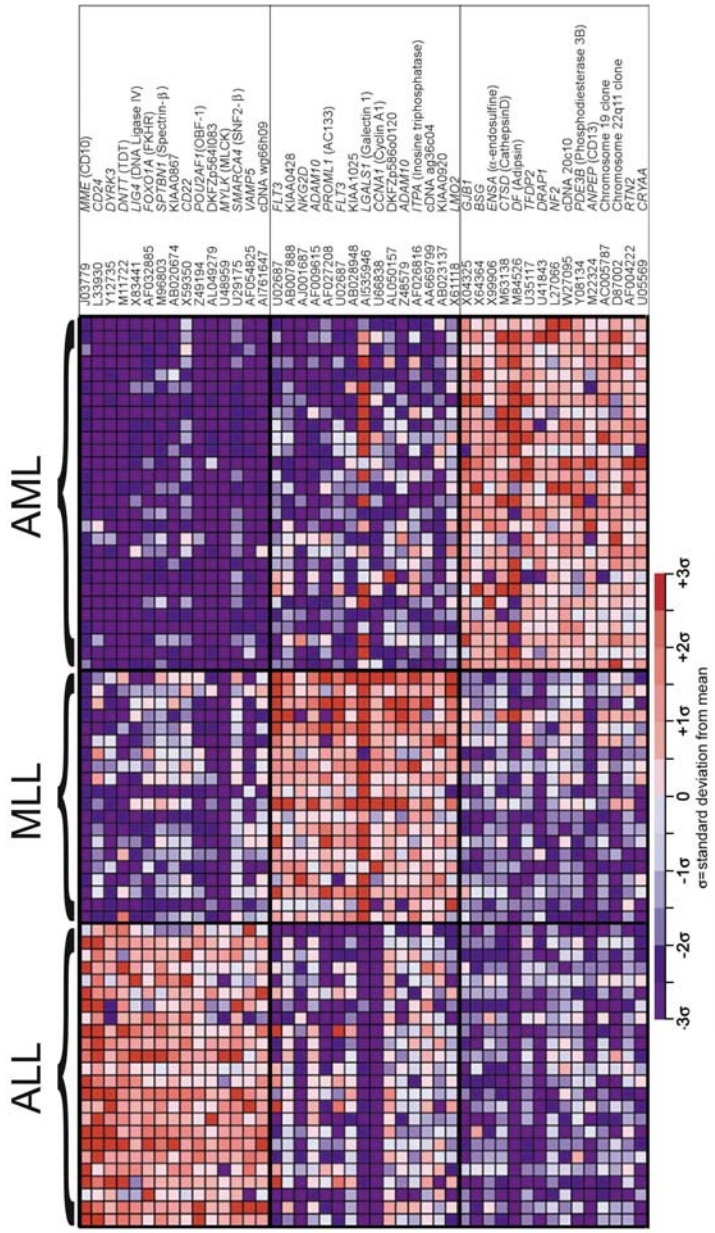

Figure 2

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

3. Selection of combinations of genes

A second, more advanced method of reducing the number of dimensions is to try to find combinations of genes that are relevant to 'recognising' and 'characterising' a certain class. Each of the combinations then results in one specific value that can be considered as what is called a 'feature'. In the simple method we have just been describing, every feature corresponds to the expression of just one single gene. But typically, a specific class is characterised, not by an individual gene, but by the interaction of several genes at once. Therefore, it is probable that we will find a better characterisation of each class by trying to combine the expression levels of several genes at once.

A well known statistical method of doing this is called 'Principal Component Analysis' (PCA). This numerical technique has been known for a long time in multivariate statistics. And these days we have excellent software available to calculate it. In PCA, the data are projected onto a lower dimensional space by using orthonormal transformations on the rows and columns of the data matrix, which, in our case, is a matrix with 12,600 rows (the gene expression levels as measured by the microarray) and 72 columns (the patients). For the leukaemia dataset, we can project the data into a three dimensional space using PCA, and each patient can then be represented by a point in a three-dimensional space (see Figure 3). The coordinate axes in this example are not individual gene expressions, but each of the three axes is a specific combination of the 12,600 gene expression levels. So, instead of scoring the profile of each patient over 12,600 genes, we can now characterise each patient, as the result of the PCA calculation, by just three numbers (the coordinates in the three dimensional space obtained from PCA). In so doing, we can clearly see that there are three different classes of patients. It can be verified that these classes indeed correspond to the three types of leukaemia we started with in our patient sample.

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

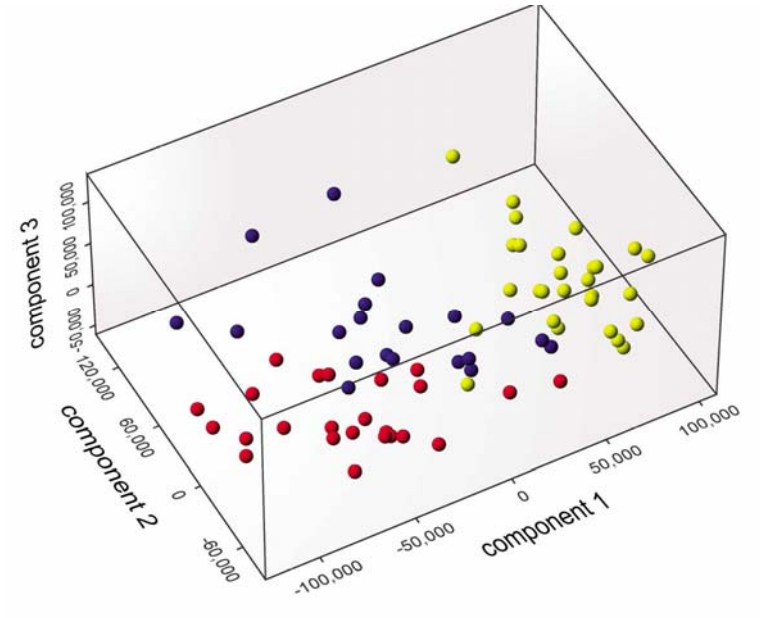

Figure 3: Principal component analysis of 72 patients with leukaemia. This analysis results in three 'features' (which are the coordinate axes). One can clearly see that, when patients are represented by their coordinates according to these axes, there are three clearly distinguishable classes of leukaemia in this dataset: the left class ALL, the middle class MLL and to the right AML. (Figure taken from Armstrong, 2002.)

4. Predictions

In a clinical environment, it is important to predict the response of patients with respect to a certain therapy. This can be done by making models using the features that have been selected to describe a certain disease. The parameters of these models are determined by using data from samples obtained from patients that are known to suffer from the disease under study. This set of samples is called the 'training set'. The trained model can then be used for trying to carry out predictions for patients for whom classification is not yet known. Some examples of models we use are 'artificial neural

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

networks', 'Bayesian networks', 'Linear Discriminant Analysis' and 'support vector machines'. In this way, we could use our sample of 72 patients, for each of whom we know which of the three types of leukaemia they suffer from, to try to diagnose 'new' patients. This could be done by calculating for each new patient, the three coordinates in the coordinate system delivered by PCA, calculated on the sample of the 72 patients, and then deciding to which class the new patient belongs.

With these methods, we can develop models in which microarray measurements are used to assist the specialist, or to predict the clinical progress of a tumoural process.

We can train a model in such a way that the presence of metastasis can be predicted (even in a case where this is not yet traceable clinically). In this way, we could select patients that would be most helped by additional therapy (e.g., chemotherapy) or patients for whom additional therapy would imply unnecessary toxicity or mutilation. We can make models based on microarray measurements that predict whether a tumour will grow slowly or aggressively. Such models can be used to make a prognosis of tumour development. Microarray-based models can be used to predict the success of a certain therapy, i.e., whether a certain therapy leads to complete remission or progression.

5. Class discovery

As we have shown, microarray measurements can be used to cluster samples from patients with a similar behaviour on several thousands of genes. This can be done using so-called 'clustering algorithms'. Some popular methods are 'hierarchical clustering', 'self-organising maps' and 'K-means clustering'. A clustering algorithm typically recognises clouds of points in a high-dimensional space, by assessing which points behave similarly as a function of certain features. The 72 leukaemia patients we have been discussing can be used as an example. By first calculating a PCA of the 12,600 x 72 data matrix, we can 'represent' each patient as a point with three coordinates in a three-dimensional feature space, which is 'automatically' determined by PCA. We could now apply a clustering algorithm to automatically find the three clouds that we clearly see in Figure 3 (in the three different colours). The leukaemia example we have used here is an easy one, as we can easily distinguish the three classes in Figure 3. However, in most applications, the task is not so easy. To begin with, the PCA dimensionality reduction might require a larger-dimensional space, e.g., five

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

instead of three. In that case, we can no longer 'visualise' the patients. In addition, in that five-dimensional space, it might not be so straightforward clustering the patients into separate classes. For instance, the boundaries between two classes are not necessarily straight, 'flat' surfaces, but can be quite complicated, and in those cases advanced clustering algorithms are used.

Figure 4: Examples of class discovery using data from microarrays. We see two data matrices visualised. The first one contains, as its rows, nineteen patients with ALL, and the columns are the gene expression levels for 80 genes, which together can characterise this type of leukaemia. Element (i,j) of this matrix represents the gene expression level (encoded by a certain gray level) for gene j in patient i.
The second matrix contains, as its rows eighteen AML patients characterised by 87 genes, the expression level of which characterises AML. The two matrices shown here are the result of a so-called 'blind' class discovery problem, in which one is given a large matrix of patients (rows) and gene expression levels (columns). It is not known beforehand how many different classes of diseases there are (AML, ALL and MLL), nor which genes might characterise these classes. Nor is it known which and how many patients belong to each of the classes. In our research, we have been developing so-called 'bi-clustering' algorithms, which manage to find classes of patients that belong to the same type of leukaemia, and simultaneously discover the genes that characterise each class.

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

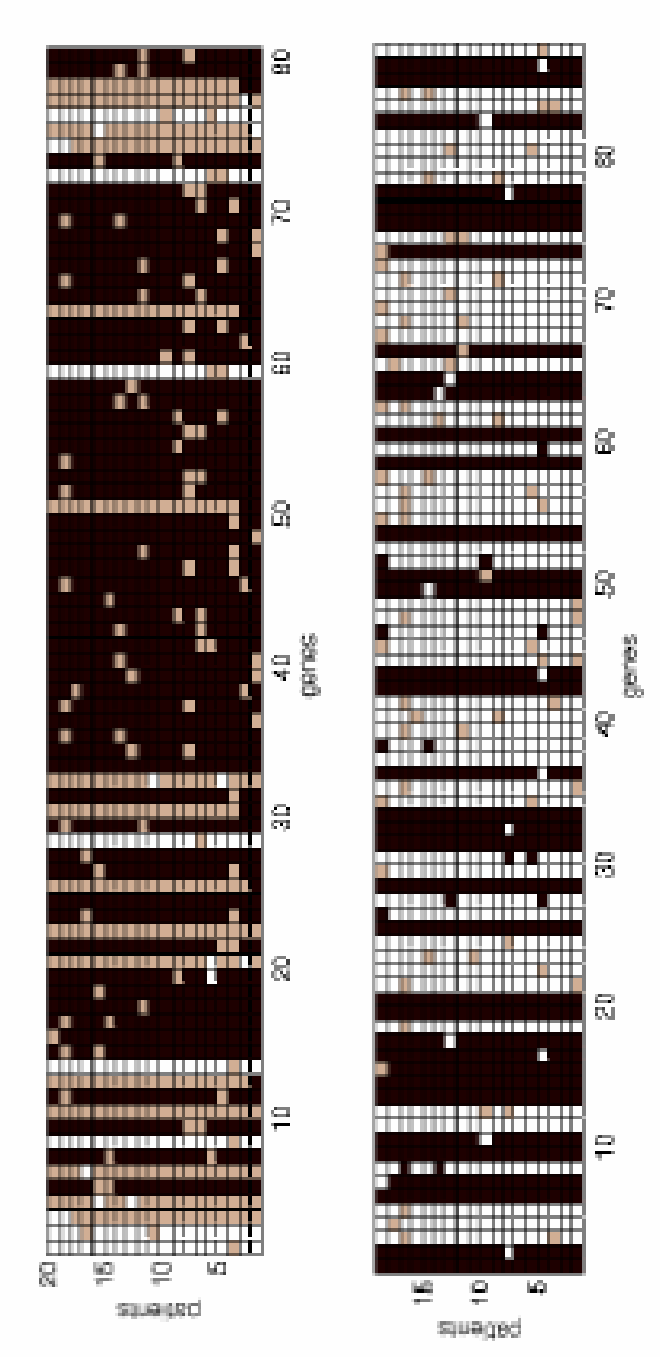

Figure 4

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

UNRAVELLING DNA FUNCTIONALITY

Let us now discuss a second example of the applicability of microarrays.

1. Context

The most important aspect of the gene expression process is that transcription (i.e., the way DNA is transcribed into messenger RNA (mRNA)) starts by the binding of an activated transcription factor with the DNA. In 1987 it was shown that there must be a certain complementarity between the active site of a protein and the DNA, before a transcription factor can bond with the DNA (Berg, 1987). This implies that, if we compare the different binding sites of a specific transcription factor, these binding sites share certain common features at the nucleotide sequence level. Nowadays, there is a lot of ongoing research to develop algorithms to detect these binding sites in the genomes of species. A typical binding site consists of a relatively short sequence (e.g., ten to twenty nucleotides) so it is extremely difficult to find them in the genome. However, the search can be focused by assuming that genes that show a similar expression level under certain conditions are regulated by the same transcription factors. As a first step in the detection of regulatory elements, we have to identify genes that show similar expression behaviour. This step is also implemented using clustering algorithms.

2. Clustering gene expression profiles

As we discussed earlier, microarrays determine the expression levels of thousand of genes simultaneously. And, as we have already mentioned, we can repeat microarray experiments with samples from different patients, samples obtained at different instances of time (e.g., during cell division, during the cell cycle or during therapy) or samples obtained under different experimental conditions. The objective of clustering gene profiles is to find groups of genes that show a similar behaviour in patients, time or experimental conditions. Genes that show such a similarity are called 'co-expressed'. In Figure 5, we see the result of a cluster analysis of a collection of genes, the expression of which was measured at eighteen different instances of time during the cell division of yeast cells (Spellman, 1998). For genes that belong to the same cluster, there is a higher probability that they participate in a common bio-

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

logical function in the cell cycle. Therefore, it is highly probable that they share the same binding sites for transcription factors.

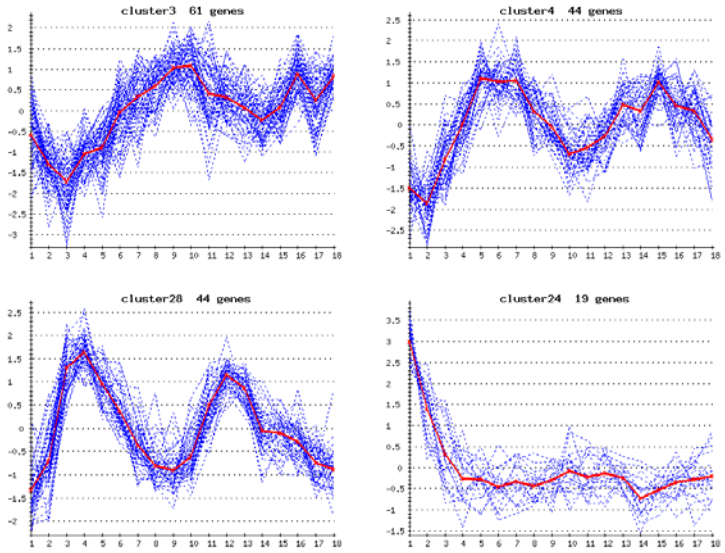

Figure 5: Four clusters of gene expression profiles measured during two cell cycles of yeast. Thin dashed lines represent the individual gene expression profiles, measured with microarrays over eighteen points in time. Cluster three (top left) contains the expression profiles of 61 genes as a function of time. Cluster four (top right) contains those of 44 genes. Cluster 28 (bottom left) contains the profiles of the expression levels as a function of time of 44 genes, and cluster 24 (bottom right) those of nineteen genes. The thick lines represent the average expression profile of these clusters, averaged over all gene expression profiles in each cluster. The fact that certain genes co-occur in a cluster is probably an indication that they are involved in a common process in the organism.

3. Representing binding sites

A collection of binding sites is called a 'motif'. One specific binding site is then a specific instance of a motif. A motif is represented by a motif model, which can be just a string of nucleotides or a matrix, the columns of which refer to the positions in the motif, and the four rows give the probability that we will find an A, C, T or G at the corresponding position in the motif. To construct a motif model, we start from a set of DNA sequences where the tran-

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

Motif 2: TyTTTCCAwC

	1	2	3	4	5	6	7	8	9	10
A	0.0979	0.0096	0.0096	0.0096	0.2747	0.0096	0.0096	0.8049	0.4514	0.0096
C	0.0049	0.4467	0.2700	0.0049	0.0049	0.9769	0.9769	0.0049	0.0049	0.6235
G	0.0041	0.0041	0.0041	0.0041	0.0041	0.0041	0.0041	0.0041	0.0041	0.0041
T	0.8931	0.5396	0.7164	0.9815	0.7164	0.0094	0.0094	0.1862	0.5396	0.3629

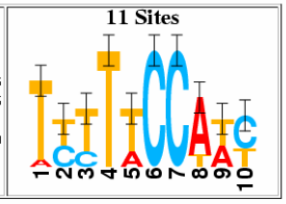

Figure 6: Probabilistic representation of a motif, i.e., the representation of a collection of possible binding sites for a transcription factor within the DNA, which can start the transcription of a gene. We show a motif, which numbers ten nucleotides. In the matrix on the left, we can see, in the first column, that the probability of having a letter T at that position is equal to 0.8931. This implies that more than probably the first letter of the motif is T. For position ten, we see that the probability of having a C at that position (0.6235) is twice as large as that of having a T (0.3629). In the figure on the right we see a visual representation of the motif, in which the size of each letter is proportional to the probability of having that letter at a certain position. Recently, we have been developing algorithms that can find motifs like the one represented here in DNA-sequences. These are algorithms that try to estimate the probability matrix per motif, using advanced statistical techniques.

scription factor binds. By grouping the samples in this set, we can deduce a so-called consensus. This consensus is formed by align ligning all segments and at each position selecting (or ranking) the most probable nucleotide(s). In one version of this method, we construct a position-frequency matrix, which is a four by N matrix, with N being the length of the motif. The rows correspond to A, C, T and G. Each element in the matrix represents the probability with which the nucleotide corresponding to the specific row will occur at the position represented by the column. An example can be found in Figure 6, where we see a probabilistic representation of a motif of ten letters.

In our research group we have been developing advanced algorithms to detect motifs in DNA sequences and represent them in the same format as shown in Figure 6. On the internet, there are several databases where motifs are stored and can be downloaded by other researchers, so that they can be verified or validated biologically.

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

4. An example

As an example of motif detection, we have used an extensive and often-used dataset of microarray data from the cell cycle of baker's yeast, *Saccharomyces cerevisiae* (Spellman, 1998). The cell cycle here consists of four consecutive phases: phase G1 in which the cell grows, the S-phase in which DNA synthesis occurs, the transition phase G2, and the M-phase in which eventual mitosis occurs. The microarray samples were obtained at eighteen instances of time in two consecutive cycles of the cell. After some data preprocessing, we use one of our home-developed clustering algorithms (called AQBC: Adaptive Quality-Based Clustering), from which we find 38 clusters. Four of these are shown in Figure 5, where one clearly sees the different phases of the cell cycle. The first three clusters (clusters three, four and 28) clearly show a periodic behaviour. The fourth cluster (cluster 24) contains nineteen genes that have a high expression level at the start of the experiment and that are switched off later on. For each of the genes in one cluster, we select a sequence of 800 nucleotides upstream of the gene. With our motif detection algorithms, we then look for common motifs of a length that varies between five and seventeen nucleotides. The results we obtain are quite different for the several clusters. The most prominent motif is found in the sequences upstream of the genes in cluster 28. Here we find in all motifs a common consensus: ACGCGT. This consensus corresponds to the well-known MCB motif, which is known to play an important role during the cell cycle. In cluster four, we find two motifs: TTTsGykT and TGTTTsTT (the small letters represent several possible nucleotides at the same time).

These two motifs are unknown in the present-day databases. For cluster three we do not seem to find any significant motif, as we only find consensus sequences containing only A's and T's, of which it is known that such sequences cannot really be regular motifs. The analysis of the non-periodic cluster 24 for short motif lengths reveals a consensus motif of ATGAAAC, which shows a striking resemblance to the STE12 motif that is found in certain dedicated databases. As a matter of fact, one of the genes, for which it has been proven that it is regulated by STE12, is present in cluster 24. This is an indication that the motif we found also influences the other genes in this cluster. If we try to find longer motifs, we find one in the form of ATATATGnnTCAGATA in seven genes. In the known databases we cannot directly retrieve what function it could have. But the fact that we can see this motif consis-

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

tently in seven genes can be a source of inspiration for further biological research and validation.

THE NEAR FUTURE: SYSTEMS BIOLOGY

Until recently, biological research concentrated on the role of single genes, proteins and other molecules as relatively isolated entities. The use of new 'high throughput' technologies, part of which we discuss in this contribution, opens completely new perspectives for biological research: we now know that genes interact with each other through complex regulatory networks, thereby influenced in one way or another by external 'stimuli' ('inputs'). From this perspective, we consider an organism as a dynamical system (i.e., a system whose variables evolve as a function of time), characterised by a certain 'state', and interacting with its environment through inputs and outputs. The whole behaviour of the organism is determined by a complex dynamical interaction between genes, proteins and metabolites in a complicated network. Through the increasing availability of data from an increasing number of model organisms, we can now start comparing the cellular mechanisms between organisms. The new discipline where we study these interactions is called 'systems biology'. It is an interdisciplinary and cross-disciplinary research domain, where we combine high-throughput molecular biology (microarrays and transcriptomics, proteomics, etc.), using system identification techniques and data mining to obtain mathematical and statistical models and acquire insight into the fundamental mechanisms of biological systems.

The reconstruction of genetic networks using molecular biological data is one of the main challenges of systems biology. A cell can be considered as a dynamical system that processes input signals, through its interaction with the environment, in an adequate and appropriate behaviour. The genetic network plays an important role in the signal transduction. The functional parts of a genetic network are the genes and proteins, each of which is connected in a more or less hierarchical way. In this sense, genetic networks can be compared to electrical circuits. Whenever a gene located at the top of the regulation cascade is switched on, e.g., through an external stimulus, the corresponding protein it generates will in turn be responsible for switching on a next set of genes. Through the hierarchical structure of a genetic network, this phenomenon of cellular signal transduction is a multiplicative process.

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

In addition, there are many non-linear mechanisms such as the presence of (nonlinear) feedback mechanisms, all kinds of synergetic effects and even Boolean logic gate mechanisms. These nonlinear mechanisms – the mathematical study of which is the branch of 'systems theory' – determine the autonomy of a cell as a self-regulated system that is quite robust with respect to variations in its environment. However, as of today, the precise causal structure of most genetic networks is unknown. Obviously, new technologies such as microarrays can play a very important role in trying to unravel the operational modes – both qualitatively as well as quantitatively - of a genetic network.

Besides obtaining fundamental insights into the mechanistic operation of an organism, network inference also opens perspectives for a wide variety of industrial and medical applications. Think, for example, of the unravelling of certain network modules, involved in certain cancer types, which, when well understood, might lead to an improved diagnosis, prognosis or treatment.

Yet another example of a process that is apparently controlled by a complex genetic network is 'quorum sensing' or cell-cell communication in prokaryotes (bacteria). It was discovered that some bacteria can communicate with each other in a common chemical language, which serves to initiate a certain pathogenesis. Understanding the fundamental mechanisms behind quorum sensing is an important challenge in microbiology (and can, for instance, lead to a better use of probiotics as an alternative for preventing and fighting infections). Quorum sensing is possible by the production and release of signal molecules, called auto-inducers (AI). The gene that codes for AI-2 synthase, i.e., *luxS*, is conserved in more than 40 species. Bacteria adapt their gene expression as a function of changes in the quantity of signal molecules. The genetic network that is (in-)directly controlled by AI-2 is not very well understood at this moment. Systems biology can lead here to new breakthroughs.

In summary, we can state that we are at the beginning of a new revolution in the life sciences, in which biology and information technology will lead us to new discoveries and applications. It might also be necessary to modify the century-old classification system of Linnaeus, because genetic analysis has taught us a lot about the origin and evolution of species. The evolutionary relationships between species can be described much more accurately using DNA information, than based on external features (phenotypes) (Holmes, 2004). But there is more to come. In this contribu-

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

tion, we mainly discussed 'transcriptomics', i.e., the unravelling of the function of certain DNA sequences by using microarray technology. But now there are already important technological breakthrough in proteomics (interactions between proteins) and metabolomics (the biochemical processes in the cell). The wealth of information contained in databases downloadable from the web and the databases in millions of scientific publications has grown so drastically, that it has become necessary to develop algorithms that can summarise the main notions of these papers, and that can perform a correlation analysis between papers to find possible interactions between genes. This is called 'text mining'. Text mining is already used heavily by biologists and biomedical researchers to simultaneously screen thousands of articles to discover potential biological relations, which can then be validated in wet lab experiments. Bioinformatics and systems biology: we ain't seen nothing yet![21]

[21] I would like to thank all my past and current PhD students, my postdoc researchers and my research colleagues in the many projects and networks we are involved in, for their direct and indirect contributions to this article. Of course, all mistakes, inaccuracies and simplifications are my responsibility. Suggestions can be mailed to bart.demoor@esat.kuleuven.be

BIBLIOGRAPHY

If you would like to read more about the research we do in bioinformatics and systems biology, you can always consult the 'pub-engine' on our website: http://www.esat.kuleuven.be/~sistawww/cgi-bin/pub.pl.
Here, you will also find many research articles. Examples of bioinformatics software can be downloaded from: http://www.esat.kuleuven.ac.be/-~dna/BioI/Software.html, or from http://www.mathworks.com/products/bioinfo/ and numerous other websites.
Also on the website of the Flanders Institute for Biotechnology, www.vib.be, one can find a lot of useful information.

Some very readable books on biotechnology and genetics are (among many others):

Kevin Davies. Cracking the genome. Inside the race to unlock human DNA.

Robin Marantz Henig. The monk in the garden (about Gregor Mendel). Houghton Mifflin Company.

Matt Ridley. Genome.

James Shreeve. The genome war. Knopf.

Brian Sykes. The seven daughters of Eve. Bantam Press.

Brian Sykes. Adam's curse. Bantam Press.

Other references:

Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R. and Korsmeyer, S.J. (2002) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genet, 30, 41-47.

Benenson Y., GilGil B., Ben-Dor U., Adar R., Shapiro E., (2004), An autonomous molecular computer for logical control of gene expression, Nature 429, 423-429

Berg, O. and von Hippel, P. (1987). Selection of DNA binding sites by regulatory proteins. J Mol Biol, 193, 723-750.

Brown T. Genomes. BIOS Scientific Publishers, 2002.

De Moor B., Marchal K., Mathys J., Moreau Y., Bioinformatics: Organisms from Venus, Technology from Jupiter, Algorithms from Mars, *European Journal of Control*, vol. 9, no. 2-3, 2003, pp. 237-278.

DeRisi J.L., Iyer V.R., Brown P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science, 278, 680-686.

Griffiths A.J.F., Miller J.H., Suzuki D.T., Lewontin R.C., Gelbart W.M. An introduction to genetic analysis. W.H. Freeman and co., New York, 1996.

Griffiths A.J.F., Gelbart W.M., Miller J.H., Lewontin R.C. Modern genetic analysis. W.H. Freeman and co., New York, 1999.

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1

Henig R.M. The Monk in the Garden. Houghton Mifflin Company, 2000.Holmes B. Time for Linnaeus to leave the stage. New Scientist, September 11, 2004, pp.12-13.

Kari L. DNA Computing, Arrival of biological mathematics. The Mathematical Intelligencer, Springer Verlag, New York, Vol.19, No.2, 1997.

Karp G. Cell and molecular biology. Concepts and experiments. John Wiley & Sons, 2002.

Kreuzer H., Massey A. Recombinant DNA and Biotechnology. A guide for teachers. ASM Press (American Society for Microbiology), Washington DC, 1996.

Lander E.S. (1999) Array of hope. Nat Genet 1999, 21, 3-4.

Lander E.S. et al. Initial sequencing and analysis of the human genome. Nature, Vol. 409, no.6822, pp.860-921, Feb. 15, 2001.

Lesk A. The unreasonable effectiveness of mathematics in molecular biology. The Mathematical Intelligencer, Springer Verlag, New York, Vol. 22, no.2, 2000, p.29-37.

Schena M., Shalon D., Davis R.W., Brown P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995, 270, 467-470.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell, 9, 3273-3297.

van de Vijver, M.J., He, Y.D., van 't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E.T., Friend, S.H. and Bernards, R. (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med, 347, 1999-2009.

van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R. and Friend, S.H. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415, 530-536.

Venter J.C. et al. The sequence of the human genome. Science, vol.291, no.5507, pp.1304-1351, Feb.16, 2001.

Watson J., Crick F. A structure for deoxyribose nucleic acid. Nature vol. 171, pp.737-738, 1953.

Reprints from: Lectures for the XXIst Century 2008 - ISBN 978 90 5867 648 1