THE ROYAL SOCIETY

Theoretical biology in the third millennium

Sydney Brenner, CH FRS

The Molecular Sciences Institute, Berkeley, CA 94704, USA and King's College, Cambridge CB2 5TI, UK

During the 20th century our understanding of genetics and the processes of gene expression have undergone revolutionary change. Improved technology has identified the components of the living cell, and knowledge of the genetic code allows us to visualize the pathway from genotype to phenotype. We can now sequence entire genes, and improved cloning techniques enable us to transfer genes between organisms, giving a better understanding of their function. Due to the improved power of analytical tools databases of sequence information are growing at an exponential rate. Soon complete sequences of genomes and the three-dimensional structure of all proteins may be known. The question we face in the new millennium is how to apply this data in a meaningful way. Since the genes carry the specification of an organism, and because they also record evolutionary changes, we need to design a theoretical framework that can take account of the flow of information through biological systems.

Keywords: genetics; DNA; computation

Like begets like is the fundamental law of biology and probably the oldest piece of genetic knowledge. During the 20th century—the last of this millennium—our understanding of inheritance has undergone several revolutionary changes; first with the rediscovery of Mendel's laws in 1901, through the DNA double helix of Watson and Crick, and culminating, in the last decade, in DNA sequencing of genomes. Genetics changed from a subject concerned simply with the segregation of characters in crosses to the direct analysis of the genes. This has led us to the insight that organisms are unique, complex systems in the natural world, which contain internal description of their structure, function, development and history encoded in the DNA sequences of their genes.

Parallel advances in biochemistry have provided us with detailed knowledge of how energy is converted to chemical bonds and chemical bonds to energy, and how the elementary chemical components of living cells are synthesized. We have come to understand the mechanisms of information transfer from genes to proteins. We know that the information is copied into messenger RNA, that this RNA is translated in ribosomes and that the codescript is read in triplets by transfer RNAs, each carrying one of the 20 amino acids. We know the special signals for starting and stopping the polypeptide chain and the code for each amino acid. The genetic code is universal, with some minor exceptions to this rule in a few organisms and organelles.

Several major technical advances occurred in the mid-1970s. These were the invention of DNA molecular cloning, and methods for sequencing DNA molecules and synthesizing oligonucleotides. These techniques allowed geneticists to clone their genes and characterize them directly, and gave biochemists access to large amounts of the proteins they were studying. In principle, the sequence of amino acids can be read from the DNA sequence, although the presence of introns found in the genomes of higher organisms may cause some difficulties. In any event, sequencing DNA became the preferred way of finding the amino-acid sequences of proteins, the direct determination of which had previously been a long and laborious process.

It was an essential feature of Crick's sequence hypothesis that the information contained in the amino-acid sequence was sufficient to determine how the chain folds to give the three-dimensional (3D) structure of globular proteins. For many proteins, this process occurs spontaneously, but in a large number of cases, special proteins called chaperonins are used to facilitate the folding of the molecules. Advances in X-ray crystallography, electron microscopy and nuclear magnetic resonance methods allowed us to determine the structures of large numbers of protein molecules and even complex protein assemblies, but the problem of going from the one-dimensional polypeptide to the folded, active structure remains unsolved and may even be insoluble.

These new methods came as a godsend to those studying the genetics of organisms. Cloning the mutated gene gave us a direct approach to the protein product of the gene and, as knowledge increased, to an insight of how it might function and thereby contribute to the observed phenotype. They liberated experimental genetics from the tyranny of breeding cycles and provided new approaches, particularly to human genetics, which had hitherto been intractable. They enabled us to move genes from one organism to another and allowed us to analyse the function of human genes in yeast cells, and to study how fish genes behave in mice.

An important feature of living organisms is the regulation of their functions. At the genetic level, Jacob and Monod showed that there were proteins that recognized segments of DNA and turned the adjacent genes off. Repression was originally thought to be the only mode of control, but we now know that there are many regulatory proteins that act positively. In higher metazoa, there are large numbers of controlling genes, which specify the times

and locations of expression of the many genes acting in development and in adaptive responses in the cells of the adult. Different cells contain different subsets of a panoply of receptors embedded in their membranes, which serve to transmit signals delivered to the outside of the cell to the inside. The signal-transduction machinery, a complicated set of interacting proteins, converts these signals into chemical currencies, which are used to control a multitude of cellular functions including growth, movement, division, secretion and differentiation. In multicellular organisms, increased complexity has been achieved not by the invention of new genes but simply by the regulation of gene expression. This reaches its apotheosis in the central nervous system of advanced animals in which the same repertoire of molecular entities is used to generate complex cellular networks.

Finally, and unexpectedly, contemporary cells were found to contain RNA molecules that display catalytic functions. These are likely to be RNA relics, survivors from very early evolution before living systems used proteins. The discovery of catalytic functions of RNA provided a molecule that could combine catalysis and the carrying of information, and bridged the gulf posed by the present partitioned situation where information is carried by one class of molecule (nucleic acids) and proteins are the catalysts. It resolved one of the important problems in how life originated.

The databases of sequence information are now growing at an immense rate and the number and productivity of biological researchers has also vastly increased. There seems to be no limit to the amount of information that we can accumulate, and today, at the end of the millennium, we face the question of what is to be done with all of this information. This problem is now widely debated and there are plans to deal with it electronically, if only to avoid the sheer weight of paper that will be required to document it. Biologists may soon have to spend most of their time in front of their computer screens. It will take a long time—if it can ever be achieved—for computers to become intelligent enough to organize this information into knowledge and to teach it to us. Writing in the last months of this millennium, it is clear that the prime intellectual task of the future lies in constructing an appropriate theoretical framework for biology.

Unfortunately, theoretical biology has a bad name because of its past. Physicists were concerned with questions such as whether biological systems are compatible with the second law of thermodynamics and whether they could be explained by quantum mechanics. Some even expected biology to reveal the presence of new laws of physics. There have also been attempts to seek general mathematical theories of development and of the brain: the application of catastrophe theory is but one example. Even though alternatives have been suggested, such as computational biology, biological systems theory and integrative biology, I have decided to forget and forgive the past and call it theoretical biology.

Now there can be no doubt that parts of biological systems can be treated within the context of physical theories: for example, the passage of ions in membrane channels or the flow of blood in blood vessels. These are physical phenomena, which happen to occur in our bodies and not in artificial membranes or pipes. There is also a considerable body of theory dealing with the chemistry of the molecules in biological systems, and with the physical chemistry of their interactions. But none of this captures the novel feature of biological systems: that, in addition to flows of matter and energy, there is also the flow of information. Biological systems are information-processing machines and this must be an essential part of any theory we may construct. We therefore have to base everything on genes, because they carry the specification of the organism and because they are the entities that record evolutionary changes.

One way of looking at the problem is to ask whether we can compute organisms from their DNA sequences. This computational approach is related to Von Neumann's suggestion that very complex behaviours may be explicable only by providing the algorithm that generates that behaviour, that is, explanation by way of simulation. We need to be very clear that this must not simply be another way of describing the behaviour. For example it is quite easy to write a computer program that will produce a good copy of worms wriggling on a computer screen. But the program, when we examine it, is found to be full of trigonometrical calculations and has nothing in it about neurons or muscles. The program is an imitation; it manipulates the image of a worm rather than the worm object itself. A proper simulation must be couched in the machine language of the object, in genes, proteins and cells. We notice, in passing, that Turing's test, which is whether an observer could distinguish between a computer and a human being, is a test of an imitation and not of a simulation.

Our analytical tools have become so powerful that complete descriptions of everything can be attained. In fact, obtaining the DNA sequence of an organism can be viewed as the first step, and we could continue by determining the 3D structure of every protein and the quantitative expression of every gene under all conditions. However, not only will this catalogue be indigestible but it will also be incomplete, because we cannot come to the end of different conditions and especially of combinations and permutations of these. Mere description does not allow computation, and novelty cannot be dealt with. On the other hand, a proper simulation would allow us to make predictions, by performing experiments on the model and calculating what it might do. Thus, if this could be carried out successfully an immense amount of information could be derived by calculation from the minimal amount needed. This is essentially the DNA sequence, the shortest description of an organism.

To do this effectively not only must we use the vocabulary of the machine language but we must also pay heed to what may be called the grammar of the biological system. We need to be clear what kind of an information-processing machine it is. It is useful to consider two kinds of such devices. As an example we consider devices that produce the values of mathematical functions. We call one a P-machine because it contains programs. When the value of factorial (5) is requested, a systems procedure invokes the execution of a program that calculates the answer. The other is called a T-machine. It has no programs but tables, and, in response to the same query, a system procedure looks up the fifth entry in the table

labelled factorial. Now the T-machine has the advantage that the values in the table can be calculated beforehand by any method whatsoever—by hand, by abacus, by mechanical calculators—and once the answer is known it is stored and the calculation need never be done again. It is clear that at the level we are considering, biological systems are T-machines; evolution has calculated values for the system by the trial and error method of natural selection and the answers are now looked up in the gene tables. There are no imperative commands, and that is why I have avoided using the term genetic program and have called it a description. Of course organisms are P-machines at other levels, for example, in the functioning of our brains. Notice that if memory is a limiting resource, a P-machine will be preferred, as indeed was the case in the evolution of the digital computer. Today, storage is cheaply and abundantly available, and now more and more computer systems employ tables rather than waste valuable processor time in calculations.

There is a second aspect of the grammar that needs comment. Genomes do not contain in any explicit form anything at a level higher than the genes. They do not explicitly define networks, cycles or any other cluster of cell functions. These must be computed by the cell from the properties of the elementary gene products. Biosynthetic pathways exist because individual enzymes carry out defined transformations at specified rates; the pathway drawn in textbooks of biochemistry is an abstraction and does not exist in the same way as the tracks connecting stations in a railway network. We need to be extremely careful in not imposing our constructions on what exists, and it is important to structure information at the atomic gene level to avoid artificial constraints. This becomes evident when we attempt to deal with multiple parallel processes going on in the same space. The coherence of a system, which may be impossible to define at the global level, is assuredly generated by the properties of the elements because the system exists and has survived the test of natural selection. Since it is not possible to start again in evolution, every step must be compatible with what has gone before; biological systems have changed by piecemeal modification and by accretion. Natural selection does not find perfect or elegant or even optimal solutions, all that is required of it is to find satisfactory solutions.

What is the likelihood that we could actually compute a simple organism from its DNA sequence? We can obtain the linear polypeptide chains reliably from the gene sequences. However, the folding problem is unsolved and is very difficult. Indeed, there may be as many different folding problems as there are proteins. However, we can resort to good heuristic solutions in the sense that proteins are composed of smaller substructures called domains, and the sequence signatures of these could be used to compute 3D structures by analogy with other proteins where these structures have been determined. We then have the much more difficult task of computing the interactions of these proteins with other proteins and with their chemical environment. This may well be impossible, but again, we may know enough about related proteins to deduce this. The very detailed properties of proteins, their specific binding constants and, for enzymes, the rates with which they transform substrates may again be beyond computational reach from the gene sequence, since there may be many equivalent solutions to the same problem.

Building theoretical models of cells would be based not on genes but on their protein products and on the molecules produced by these proteins. We do not have to wait to solve all the difficult problems of protein structure and function, but can proceed by measuring the properties that we require. At the level of the organism we would start with cells and, again, measurement could give us what we need. The reader may complain that I have said nothing more than 'carry on with conventional biochemistry and physiology'. I have said precisely that, but I want the new information embedded into biochemistry and physiology in a theoretical framework, where the properties at one level can be produced by computation from the level below.

It may be much easier to compare two genomes. The DNA sequences of any two human genomes differ from each other in one or two of every 1000 bases. If a chimpanzee genome is compared with a human genome the number of differences rises to about ten per 1000 bases. Many of these differences are without significant effect because they occur in regions or in positions where they could be judged to be strictly neutral. It would be fascinating to ask whether we could discover the differences that do count and whether we could reconstruct our common ancestor and thus find out what mutations occurred during the course of evolution to make us different. I believe that this is what we should be trying to do in the next century. It will require theoretical biology.