THE ROYAL SOCIETY

Published online 9 December 2003

Perspectives

First let me talk very briefly about the past 50 years. I think there were two revolutions. The first was the demonstration of the double helical structure of DNA. This opened up our thinking and destroyed the classical idea of the gene as an indivisible entity and a unit of mutation, recombination and function. It gave us the modern idea of the gene as a stretch of nucleic acid that was responsible for information transfer from gene to protein. Any deeper knowledge of gene structure and action had to be acquired through analysis of phenotypic expression. It was, if you like, trying to sequence genes by genetics, which is hard work.

The second revolution, which occurred about halfway in the past 50 years, was the invention of DNA cloning and sequencing. This had two effects. First, it liberated geneticists from the tyranny of breeding cycles of organisms; we could now study the genetics of anything. Second, it liberated biochemists from servitude in the cold room. They no longer had to purify proteins, they simply made them in *Escherichia coli*. In addition, the technological advances that have occurred in structure determination are amazing. I am told that a new three-dimensional structure of a protein is solved every 5 hours, so we now have an enormous amount of information in our hands.

We now stand at a closing point—50 years since the determination of the structure of DNA—at a time that is marked by the completion of the human genome sequence. I have the feeling that everybody is relaxing, thinking that they have done the job, and the machines are getting rusty. What I have to say in the next few minutes is about the tasks that lie ahead.

One major task will be to find out more about what there is in the genome other than coding sequences. These sequences we can now read pretty well, but finding and deciphering regulatory sequences, which could be considered more important, is proving to be very difficult. However, we now have a remarkable set of tools, and I will just mention one example; one can approach this problem by a method that I call 'inverse genetics'. Ordinary genetics looks for differences in a sea of uniformity, whereas inverse genetics looks at what has been conserved over long periods of time—enough time to randomize all non-essential sequences. We have been doing this using a very convenient sequenced genome, that of the puffer fish (*Fugu rubripes*), which probably separated from us about half a billion years ago. So we have available all of that 'free mutation' to analyse and therefore can ask ourselves whether there is any commonality of function of these conserved sequences. But before we do this—which means doing it on a computer—we can make a test by carrying out what is essentially a rather strange 'genetic cross', between fishes and mice, to determine whether sequences from the fish carry information that the mouse can read. Of course, we want to know what the mouse is reading, so if you take a control region from the puffer fish and you put it into a mouse and the mouse reads it faithfully—that is, the phenotype is the same—we can then say that the feature that is common is a conserved control element. We can prove that an element is necessary by deleting it, and the proof of sufficiency would be to synthesize the promoter. We could make a promoter that the mouse reads and then we would know what its important features are. I call this comparative genetics, not comparative genomics. Comparative genetics is done by interspecies crosses, which test the value of a DNA sequence in common hosts. For example, we could test the value of chimpanzee sequences and human sequences to see whether they were the same or different by putting them into a mouse. That would be the best test, using the mouse essentially as a test tube for this kind of experiment. One of the things that has become clear by having this capacity to analyse promoters in great detail is that most genes have more than one mode of expression. This is called variable splicing, but splicing variability is like meditation in that there are three kinds: accidental, incidental and transcendental. What we are interested in is the transcendental variation. In other words, what is it that determines functionality? What we find is information encoded in the genome that not only specifies in which tissues a given protein will appear and which cellular compartment it will enter, but also can provide specificity of interaction with other macromolecules by adding or removing pieces of protein sequence. For a given gene, there may be as many as five such genetic determinants, or 'instantiations', as I call them. It is possible for the cell to specify whether a protein product will go into the nucleus or into a membrane, or effectively combine with another protein. All that information is there in the genome waiting to be discovered.

I think that we also have a problem concerning how we are going to gather all of this information together into a functional model. I don't think we will learn much by more annotation of the genome; it is already totally opaque! I happen to believe there is too much text in the world already, and we should stop using text to communicate things. So, I think we are going to need what I call a map, and the correct level of abstraction is to make these maps at the level of the cell, because the cell rather than the gene is the correct level of abstraction for function. Incidentally, the cell theory is *ca*. 150 years old. Of course it is our job is to place the instantiations in the right cell. We have many questions to ask. How many different kinds of cells are there? How many non-contingent kinds? How many contingent states are there? What is the character of any molecular computation that is performed at a promoter? Is it all 'feed-forward', or is there some more

One contribution of 18 to a Discussion Meeting Issue 'Replicating and reshaping DNA: a celebration of the jubilee of the double helix'.

complex conditional calculation that would read in logical terms 'if x and y but not z'? In other words, is there a committee of transduction factors, sometimes with a very strong chairman with a veto? Or is there a committee with a consensus, or is there one protein that just obeys orders and does what it is supposed to do? I think that all of these questions about the architecture of controlling elements need to be answered.

Now, I would like to come to more global issues, of which we have seen some excellent beginnings today. The study of diversity is going to be very important and perhaps we do not need model organisms anymore. This feeling arose after a discussion that I had with people who wanted to take 15 inbred strains of mice, breed them all together to generate 30 000 strains, and then train mouse phenotypers to measure their blood pressure, saliva and all of the other things that can be phenotyped. They would then try to characterize the genomes of these mice, except they have not got the technology to do that. So, I asked myself at the time last September (2002) if we had the technology to do this, why bother with mice? We already have the 30 000 lines outside of the laboratory: they are called people! And we have very highly trained phenotypers, called doctors, and so it seemed to me that we must now have completely different ways of looking at genomes, as we have begun to see at this Discussion Meeting. We should not think of individual genomes, but of populations of genomes, and one may ask whether there could be new methods that would reduce the effort and costs of sequencing 30 000 genomes. The answer is that there are, and I am very hopeful that, in the next 17 years, we will be able to compare the entire important 5% of sequences in at least 30 000 human genomes. I pick the year 2020 because the period from now to then is about the same as the period from when we first started to think about sequencing the human genome to now. You should remember that the people who planned to sequence the human genome did this when the largest genome that had been sequenced at the time was phage lambda at 45 kbp. If we are going to sequence 30 000 genomes, we might as well do it three times and sequence 100 000; whatever we find we can validate on many more people.

Walter Bodmer always said that he would like to analyse the genetics of the human face. I think that is an admirably complex thing, but we would have to define the phenotype pretty accurately to get the correct variability. Maybe Alec Jeffreys already has the DNA with the photographs—face on and sideways—of people, so maybe we could get into the database and see if we could find any association.

When thinking about genome evolution, there are still a lot of problems at the theoretical level that need explanation. The human genome should contain a record of its past. How can we extract this? Or, can we prove that there is no such information, so we would not bother to try to extract it? I think that we need a completely different approach to look at the evolution of genomes, and when one does this one may find a completely different picture emerging. I say this because I think that people have paid little attention to elementary facts, such as that genes are not individual balls in a bag but rather that they are linked, and I think there is a lot to be learnt from looking at linkage relationships. For example, we can find out what genes have been together for long periods of time simply by looking at puffer fish DNA, where we have identified five genes in exactly the same order, in the human. We know that there has been no rearrangement in those areas and, of course, we can begin to study changes in ancestral intergenic regions. This information becomes very important if we are going to try to reconstruct evolutionary events over very long periods of time. Being able to see how fish DNA works is also highly important for conserved physiological systems: for example, the immune or neuro–endocrine systems, or in the behaviour of channels and receptors in the brain. We can compare how things work by finding pieces of DNA that have not changed at all during the divergence of humans and puffer fish from a common ancestor. One 14 bp non-protein coding sequence, totally conserved between puffer fish and man, works correctly when placed in the mouse genome. Much comparative work has focused on polymorphisms in proteins and how these relate to comparative function. I think that we are going to find more interesting things in control regions and this may explain much more about the human condition and variation than we think today.

Work that you heard about here from Alec Jeffreys, as well as that done long ago by Jan Klein on human leucocyte antigen (HLA) tells us a lot about evolution and really challenges the naive ideas about evolution that someone like myself once had. My mental image was that one day some monkey got up and said 'to hell with being a monkey, I am going to go out and become a human'. What the HLA population variation studies showed is that the actual alleles in the Old World apes are the same as those in humans, thus telling us that it was a whole population that went out to become different. Evolution of a species could not have been due to the selection of one genome, because allelic variation is very old and common to both predecessors. It is going to be very interesting to think about the dynamics of this because once you have an evolving population within which genetic changes take place, both old variation and new mutation can be reassorted by recombination in the population.

Someone said that genetics had a great future behind it. I think genetics is only just starting. We have been given the tools and many of the concepts to proceed with what I think is the greatest challenge of all: to create human sciences in the twenty-first century.

Sydney Brenner *July 2003*
The Molecular Sciences Institute, 2168 Shattuck Avenue, 2nd Floor, Berkeley, CA 94704, USA