

Towards a data monopoly?

It is a disturbing experience for an academic scientist to sit through a press conference given by any biotechnology company which specialises in platform technologies. To be told that yeast can be sequenced in 6 hours or the bacterium of your choice in only 2, makes you feel as if you had to adjust to the news that somebody had jogged up Mount Everest in the same time it normally takes to run a mile. No matter how optimistic you are, the message is clear: forget about catching up. The sequencing facility that you are planning at great expense should be reconsidered, because the sequencing machines will not be fast enough and the computers will not have the necessary capacity to process the output. The impact on proteomics will be equally dramatic as soon as phrases, such as 'a million protein sequences per day', become reality. The unlimited mass of data that can be generated by the biotechnology industry means that the 'only' task left to be performed by academics is its compilation in order to get a full description of biological events. I have addressed this in an earlier editorial ('Back to Darwin?', November 2000) and I remain convinced that the truly interesting work still lies ahead and will take place in laboratories that have more brains than computers.

But in order to carry out this new biology, scientists must have access to the necessary raw data. The work carried out and planned by Celera is indeed impressive but it also means that the key to the treasure chest of information is linked to certain conditions. The price may be within your means if you simply need a few thousand bases of sequence and if you are a genuine basic researcher working on a topic with no possible commercial consequences. But very few researchers today work on projects that are totally devoid of potential commercial applications, and even fewer can work with only small amounts of sequence data for very long. It is therefore soothing to know that a public source provides these data for free, but the relief has to be matched with the

concern about its long-term viability. Indeed, when measured by the parameters of speed, automation, computer capability, and—most importantly—single-minded organisational efficiency, the commercial capability to generate sequence is superior to the public enterprise. (As always, I welcome any correspondence correcting that assertion.)

It follows that most microorganisms of relevance, all mammals, many plants, parasites and their hosts and anything else that catches the fancy will be sequenced and added to the burgeoning stocks of data in the commercial storage rooms. And this will happen faster than publicly funded plans can be agreed upon, never mind initiated. There is the possibility that a predominantly public-funded public/private consortium will ensure that sequences that are of less commercial value than the human one will be available to everybody without restrictions. But that is dependent on the interests of the company and can, of course, change with time and market demands. Furthermore, those consortia will not be needed for long, and it is not the private sector that will be superfluous.

So how can public efforts to re-sequence any organism be justified to the taxpayer? How will the scientific merit of any such exercise pass scrutiny? Who will volunteer for the task? The outcome seems clear: countries and organisations will rush in ever increasing numbers to buy licences to access the commercial data. This is a perfectly pragmatic solution but the consequence is that funds that would have been used for research will instead be used to buy this new commodity. A downward spiral of research activity will be the consequence for some laboratories, while the well-funded ones will add to their dominant position.

The problem is not that Celera is wrong or conducts a reprehensible business—far from that. They are doing exactly what one would wish from a biotechnology company. But the recent development could give rise to a situation where there is only a single provider of an essential com-

modity. To avoid such a monopoly, policy makers need to react now. But herein lies another series of dilemmas. Can an effective alternative be established? Would the US government take actions that would damage a US company? And, if a publicly funded alternative were initiated, would it damage the private company in such a way that the generation of essential data were impeded?

So what should be done? If one focuses on the future, i.e. comprehensive proteomics, structural biology, networks of interacting proteins etc., then at least there is a choice; public funding should be directed massively towards generating these data. But this will require a level of coordination and innovation that is hard to achieve within the academic community. It will also require additional funding if such an initiative is not to deprive the current research projects of their financial resources. Such a massive undertaking seems like a task for the European Commission, if not the G8, which set up the Human Frontiers Science initiative some years ago. But, of course, other associations could and should play important roles too.

For DNA databases, however, it is probably too late to establish an alternative. One possibility could be to fund, on a large scale, private and academic enterprises, such as the EC Joint Research Centers, to re-sequence important organisms with the requirement that all data are placed in the public domain. There are few other alternatives if we do not wish to leave the future of life science research subject to the whims of a single company. Whatever the solution, there will be a major new expenditure in research budgets everywhere. A completely new paradigm for experimentation has been enabled, and this has a series of complicated consequences as well as wonderful opportunities.

Frank Gannon

DOI: 10.1093/embo-reports/kve098