DIRECTOR'S REPORT

There are times when it is exhilarating, but also humbling, to be a member of a privileged fellowship who collectively call themselves scientists. The year 2000 was such a moment, for we saw extraordinary progress in understanding the fabric of life, the sequence of bases in DNA that encode our genetic information. Following past successes in the sequencing of the genomes of many microbes—the yeast *Saccharomyces cerevisiae* and the small worm *Caenorhabditis elegans*—a flurry of important genome sequences emerged during 2000, all with great fanfare. Early in the year, we saw the publication of the sequence of the genome of the fruit fly *Drosophila melanogaster*, which was a result of a successful collaboration between academic scientists and the private sector, a harbinger of future genome projects. In the middle of the year came the highly coordinated, multinational announcement that the drafts of the human genome were complete, following intervention and direction from President Clinton in this country and Prime Minister Blair in the U.K. The publication of the draft human sequences appeared in early 2001, again with much public attention. With lesser blowing of the trumpet, the last month of the year saw the publication of the complete sequence of the first plant genome from *Arabidopsis thaliana*. Thus, for the first time, it is possible to gaze into the intimate genetic details of organisms from all kingdoms of life. As a consequence, public interest in biology and medicine is at an all-time high and probably rivals public interest in science following other monumental scientific and engineering feats, such as landing a man on the moon and the splitting of the atom. We have been fortunate to witness and participate in these dramatic advances. A new era of biology has blossomed that will long have profound consequences for science and indeed for society as a whole.

The revolution of recombinant DNA that emerged in the early 1970s made it possible to obtain unlimited amounts of virtually any DNA. Imaginations ran wild, but like so many endeavors in science, technical limitations still kept biologists at bay. Perhaps the clearest and most difficult technical challenge was to figure out how to read the sequence of individual genes. By the mid 1970s, just as I was entering graduate school in Australia, two vastly different approaches to DNA sequencing emerged. One from Wally Gilbert's laboratory in Cambridge in the U.S. used a chemical sequencing approach, whereas that from Fred Sanger's laboratory in Cambridge in the U.K. employed enzymatic methods. The two different technologies from the two Cambridges were complementary, but early on, each had its advocates as to which one was best. It was rather amusing to listen to young Australian scientists returning from studies in either the U.S. or the U.K. tout the virtues of *their* favorite approach to gene sequencing, depending on the country of origin. In those days, both methods were important, and it was necessary to learn both to be a successful graduate student. In retrospect, these developments proved, unwittingly, to be the first multinational collaboration in the DNA sequencing era, for the sequencing problem was essentially solved and now could be scaled up.

Progress was rapid, and within a year, the genome of a bacteriophage (a virus of bacteria) called ϕX174, which contains approximately 5375 nucleotides, was reported from Sanger's group. It was equally exciting to learn in 1978, during my first visit to Cold Spring Harbor for the annual Symposium, of Greg Sutcliffe's determination of the complete sequence of the then-favorite bacterial cloning plasmid, pBR322, using the Maxam and Gilbert technique. It was only

a matter of time before the concept of sequencing the whole human genome began to be discussed, notably at a meeting organized by Robert Sinsheimer in 1985 at the University of California, Santa Cruz. But the idea of determining the complete sequence of the human genome was controversial, as many biologists saw the cost as being too high. This was a time when grants were particularly difficult to obtain because of limited funding, and there was considerable concern that such an ambitious project could not be completed for technical reasons, even if sufficient funds could be found. During the Cold Spring Harbor Symposium in 1986, Jim Watson brought together some of the leading biologists to discuss the genome sequencing proposal: Was it feasible and, of particular importance, who should fund the considerable cost? Those days now seem far in the distant past, but it was only 14 years ago, a short time in the history of molecular biology. Wally Gilbert's prediction at that meeting that we would all carry our genetic sequence on a credit card then seemed far too fanciful, but it is now well within the realm of possibility.

Shortly after the 1986 meeting, spurred by the start of a human genome project by the Department of Energy and helped by discussions organized by the National Academy of Sciences, the National Institutes of Health began the project with considerable enthusiasm. Much of the early success of the project in the U.S. was due to the outstanding leadership of Jim Watson in establishing in this country the NIH Human Genome Research Center (later an Institute at the NIH). All biology benefited by his promoting the sequencing of model organisms, such as yeast, *Drosophila*, and *C. elegans*, and by focusing early money on ethical issues. The introduction of new techniques was of great importance to the success of the project, particularly the contribution by Lee Hood and Mike Hunkapiller in developing the automated DNA sequencing machines that eventually got the job done. Francis Collins, the second and current director of the U.S. effort, skillfully directed the difficult stage of scaling up the human genome sequencing project and shepherding the public effort to near completion. In the U.K., John Sulston and the Wellcome Trust, together with Bob Waterson in this country, played an essential role early on by committing considerable effort and resources (and their scientific reputations) to model organism genome sequencing and later, at a critical time to the human genome project, when it needed to scale up. Later, the introduction of a private effort by Craig Venter and his colleagues at Celera dramatically accelerated the progress, although the introduction of the private sector into genome sequencing has created the new problem that not all of the sequences are available to all those who want to use them.

Congratulations are due to the community of genome sequencers, particularly those supported by public funds, from the heads of the various genome centers down to the technicians who did much of the work. They have shown that it is possible to undertake large, internationally based science projects and get the job done with a remarkable degree of cooperation. But more importantly, they have shown that it is possible for very talented scientists to participate in large multi-investigator science projects that benefit all of science, even though individuals might not have received the recognition that might have come had they continued to run a conventional, small laboratory. I have observed at the annual Cold Spring Harbor genome meeting a real sense of purpose and camaraderie among the scientists who cooperated in this project. Unfortunately, this message has been lost to the public, who have instead been inundated with the public-private competition issue by the popular press. One hope is that history will focus on the truly remarkable collegiality that was evident during this time.

The genome projects have also benefited us all in another direct way. Now that many genomes are complete, funding has picked up because we have captured the imagination of the public and Congress and because there is a more realistic hope of successfully tackling human diseases, particularly cancer.

The first eukaryotic genome to be sequenced was that of the baker's yeast, *S. cerevisiae*, completed in 1996. Done in cottage industry style by investigators from many countries, the yeast genome became the model for how complete genome sequences might be used. My own work on the replication of chromosomes, in part using yeast as a model organism, has greatly benefited from the freely available genome sequence. Now that we have the human genome sequence almost complete, it is perhaps worth reflecting on how the yeast genome has helped progress in yeast biology and biology in general.

Knowing the entire genome sequence immediately establishes a closed system for understanding information flow within the organism. Since all of the genes are known, when a new activity or function is discovered, it is possible to search the entire genome to discover whether there are other similar proteins that might perform related functions. In this manner, families of proteins are rapidly discovered. Because biology reuses protein domains and activities for multiple purposes, progress in understanding one pathway will often lead to insight into other biochemical processes. This iterative accumulation of biological knowledge will not only help us to understand the biochemical processes in yeast, but, because most yeast genes have functional or sequence-related homologs in the human genome, also make it possible to predict with high accuracy protein function in human cells. For this to occur efficiently, it is necessary to know the entire human genome sequence, and as more genomes are sequenced, including those of other model organisms, the information flow from one system to another will increase rapidly. This is, perhaps, the strongest argument for sequencing the genomes of all of the important organisms used in modern biological and biomedical research.

Developments in the methodology to rapidly analyze proteins by mass spectrometry have also dovetailed with whole genome sequencing in a most productive and informative way. Very small amounts of proteins can be fragmented and the mass of the protein fragments determined by mass spectrometry. These masses can then be compared to a computer-generated database of all predicted fragments from all predicted proteins encoded in a genome. If sufficient fragments are determined experimentally, then the proteins can be uniquely identified. So it is now possible, and indeed routine in the yeast community, to use antibodies to isolate a protein present in a complex mixture of proteins from cells and then rapidly identify all of the proteins with which it interacts. Since most biological functions involve proteins working together in biochemical pathways, we will move toward the possibility of knowing many of the protein networks in cells, and even the protein modifications that occur in response to signals that the cell receives. Such an analysis and the insight it provides were not possible without knowing the entire genome sequence.

Another important technical advance from studies on yeast that has been effectively integrated with whole genome sequencing is the development of DNA microarrays. The concept of arraying small amounts of DNA for biological analysis emerged from the pioneering studies of Ed Southern in the 1970s, when he showed that it was possible to transfer DNA to fixed membrane supports and hybridize DNA or RNA to the arrayed DNA. Once the entire genome of yeast was known, it became practical to array all the predicted genes on glass slides so that all of the genes of yeast could be analyzed simultaneously, instead of one gene at a time as was done previously. In this way, the dynamics of gene expression patterns could be followed as cells responded to experimentally induced signals such as nutrient starvation, changing of carbon sources, and commitment to divide or exit from the cell division cycle. The method has become one of the most powerful tools to study the reaction of an organism to biological perturbations. It has only been possible by having the complete genome sequence available.

Coupled with the protein analysis described above, DNA microarray experiments allow a very intimate look at the molecular physiology of a cell and how it functions. We can see, on a whole genome scale, how cells work and how they adapt to their changing environment. From what we

have learned from studies on yeast biology during the last few years, it is probable that cell, tissue, and organism physiology will return as a dominant area of investigation but studied now at the molecular level. When this type of analysis is applied to animal studies, it will be possible to see how an organism responds to all sorts of perturbations, leading to an unprecedented understanding of biology and physiology. This is happening at a rapid pace. Already, DNA arrays are being used effectively to study the response of animal cells to extracellular signaling and to drugs used in the clinic. A particularly innovative analysis is under way at Cold Spring Harbor Laboratory by Tim Tully and Josh Dubnau who have identified fruit fly genes whose expression is altered during the process of learning a task or consolidating memories of the learned task. It is almost like watching the brain think!

With so much data, the field of bioinformatics, or computer-assisted analysis of biological information, has rocketed to become one of the most exciting fields of biology. Judging by the number of applications we get for a limited number of places available, the two bioinformatics courses taught each year at Cold Spring Harbor are now the most popular of all our advanced courses. Again, the yeast community has been a pioneer in the field of bioinformatics because of the availability of the entire genome sequence during the last six years and because of the need to deal with the vast amounts of data that derive from whole genome experiments. But bioinformatics includes much more than the analysis of DNA sequences. Efforts within the yeast community have developed databases that link the scientific literature to the genome. Genes and their protein products, and even the pathways in which they function, have become a foundation for computationally organizing biological information. Searching these databases has become as routine as searching the printed literature. This is perhaps the best argument for supporting a single electronic database of all research papers so that the biological literature can be fully integrated with the large number of databases that are being developed by computer scientists. Publishers of the scientific literature must strive to make science as easy as possible by ensuring that the on-line literature is linked in a seamless and accessible way.

Because biomedical and biological research has flourished, the literature is growing at a pace that far surpasses even the most avid reader. And yet the biological community has only just begun to use computational approaches to analyzing the literature. Information science will become a dominant field of biology and computer scientists will need to be integrated as much as possible with biologists. At Cold Spring Harbor, we are strengthening our already broad bioinformatics research by adding computer scientists to our faculty so that they can develop new technologies in consort with biologists.

Within the yeast community, there have been a number of initiatives aimed at providing whole genome resources, such as creating a complete set of yeast strains in which all genes have been deleted; or tagging all proteins with various sequence tags, marking each gene with a unique bar code; or attempting to identify all possible protein-protein interactions using genetic methods. In retrospect, some of these initiatives have not been as productive as anticipated in the beginning. For example, the time it takes to delete a gene is very short compared to the time it takes to know the consequences of doing such a deletion. Therefore, the availability of a complete set of gene disruptions has not yet hastened progress. Thus, we should think carefully before whole genome approaches for other organisms are attempted.

So what of the human genome? Because it is our genome, there have been sophistic writings in both the scientific and public literature that the gene sequences will reveal what it is to be human, reveal the nature of the soul, and even explain human behavior. Most of this has been overly excessive piffle put forth by those who can only be excused for getting too caught up in the momentous occasion. Many discussions about the implications of the genome sequence have also been flavored by the advent of cloning animals from single cells and the consequent,

way too premature talk of human cloning. Knowing our gene sequence, or even a mouse genome sequence, is not going to help overcome all the very considerable technical obstacles that still exist in cloning other mammals. We simply do not understand enough about the methods for producing animals from individual cells. We do not know much about how gene expression programs are reset before development can occur. Clearly, research on cloning should proceed, but knowing the full human genome sequence will only marginally help solve the significant hurdles that exist, and the two areas of science must not get confused when there is discussion about future possibilities.

The achievement of obtaining the sequences of many genomes, including the human genome, is a major milestone in science. Certainly, when the double helix was revealed, it was unimaginable that the entire nucleotide sequence of a genome could be obtained. I still find it humbling to realize that we are in a golden age of biology that will have far-reaching consequences not only for our own science, but also for humankind. At the same time, we should have realistic expectations of what will emerge from these spectacular developments.

On a practical level for most scientists, research has been made much easier because of the reagents that have derived from the genomics age. Clones of genes and fragments of genomes are readily available, as are the predicted sequences of most proteins (we do not yet have the computer tools to predict all protein sequences accurately). These resources have been put to great use, speeding up the pace of biological discovery manyfold. This progress in itself has been a silent revolution, perhaps only appreciated by the scientists actually doing the work. Many of the advantages which have become available to the yeast community during the past six years are now available to those working on human biology, including arrays of human genes, protein analysis by mass spectrometry, and comparisons to the predicted "proteome"—the set of all proteins. It is now possible to analyze the changes in gene expression of the entire set of known human genes in response to physiological changes in cells and tissues. Although it is early in this analysis, new and exciting findings have been reported in the literature. Such experiments have already led to new methods for diagnosing human disease and to the discovery of new targets for therapy. In some cases, the cause of disease has been discovered by the power of being able to compare gene sequences between diverse species, such as those of *Drosophila* and human, or even yeast and human.

One of the most interesting aspects of whole genome sequencing in humans is the diversity of sequences that are being uncovered. It is estimated that there is one difference between individuals for each 1300 bases in the human genetic code of 3,000,000,000 bases. This means that we are all about 99.9% identical, something that itself is quite remarkable. But if turned around the other way, then it means that there are about 3 million differences at the primary DNA level between individuals. Most of this variation will not be expressed, but some of it will. This means that individuals will not only have different shapes and sizes—something that we all know about—but also have different probabilities of being afflicted by disease and, when treated, different responses to drugs and other therapy. Such variation will become valuable for predicting how patients might respond to certain drugs, allowing treatments to be targeted to individuals who will benefit from the drugs while avoiding adverse effects of the same drugs in others.

Knowing more about human genomic variation also has the potential to change how we view ourselves as a species. By knowing more about human DNA variation, we will realize that traditional ethnic and cultural boundaries will not be reflected in our DNA, but rather will be purely a human invention with no genetic (and maybe even no biological) basis. If this turns out to be true, and it is really understood by the lay public, then cultural and ethnic differences may not be as dominant in future human endeavor. But, change will occur only very slowly, and this may be an unattainable utopian goal.

Clearly, whole genomic approaches to biology are changing the way we think about dealing with human disease, and perhaps this is what the public finds the most appealing. One example at Cold Spring Harbor Laboratory is the whole genome analysis of human cancer by Michael Wigler, Robert Lucito, Masaaki Hamaguchi, Vivek Mittal, and their colleagues. By combining a method called RDA (representational difference analysis) developed in Wigler's laboratory about eight years ago with high-density arrays of human DNA placed on glass slides, it is now possible to analyze small biopsies from human cancers and compare DNAs from the normal and cancer cells. Initially focusing on breast cancer, this research is already leading to exciting new results. There are new possibilities for diagnosing the disease and for identification of new anti-cancer targets. From such genomic analyses, cancers that look alike to a pathologist can be molecularly identified and classified, and understood as being separate diseases with separate outcomes.

Such analysis of breast cancer has also led to the discovery of new gene mutations in human cancers. It is already clear from the analysis of breast cancer samples that many altered regions of the genome have not been characterized, particularly in cancers derived from patients who have no known family history of the disease, which represents the vast majority of breast cancer cases. By correlating the alterations with the type and severity of the cancer, and with the response of the cancer to existing treatments, such knowledge will lead to better decision-making by oncologists. But the more important goal is to identify those genome alterations that suggest new, cancer-specific targets for therapy. Already the identification of one genetic lesion in a subset of human breast cancers has led to a novel therapy. In about 25% of breast cancers, a gene called *Erb2* is overexpressed. This knowledge led to the development by Genentech of an antibody-based drug called Herceptin which has proven effective in treating a subset of breast cancer patients. Our goal for the cancer gene discovery at Cold Spring Harbor Laboratory is to identify the key therapy targets in all classes of breast cancer. Furthermore, the methods that have been developed and applied for breast cancer research can easily be applied for other cancers, given sufficient funding.

Such scenarios will play out for many human diseases. But a caution should be noted as we all celebrate the fantastic achievements of the recent past and speculate on how quickly genomics will change our real understanding of biology. Genes make proteins, and proteins, not genes, determine how we look, how we behave, what diseases we get and how they are treated, and even how we respond to our environment. Because of gene splicing, gene rearrangements, and developmental diversity in gene expression patterns between individuals, biology is much more complex than the genomic view would suggest. Proteins can be modified in many different ways, thereby changing their function. It is possible for a single gene to encode many proteins with diverse, and sometimes even opposite, functions. This is not always apparent from gazing at the gene sequence, even with powerful computers. Our present methods for discovering the functions of proteins and the pathways in which they operate are not yet rapid enough, particularly when applied to mammalian biology. Although integrative approaches, such as comparing gene functions across diverse species, are clearly paying dividends, biology is much more complex than the set of genes that make up a genome. We need to think about greatly improving computational approaches to biology, finding better experimental ways to characterize protein diversity, and most importantly, speeding up discovery of protein function. Once this can be achieved, and we learn how to integrate this knowledge into a molecular understanding of cell and organism physiology, the true power of the genome will be unleashed.