***PRESENTATION*:**

THE UNIVERSITY OFFICE OF TECHNOLOGY TRANSFER: THE INVENTOR / RESEARCHER'S VIEW

Leroy E. Hood*

I would like to address intellectual property in the context of the scientist and inventor, to give you a very different view of intellectual property than what you've had before. As many of you know, there has been quite a profound revolution in our understanding of biology in the last ten years. In part it has been due to the Human Genome Project, which we will discuss, and in part it has been due to a convergence between information technology and biotechnology. This has led to fundamental changes in how we look at biology. So what I would like to do with this presentation is to give you some sense of how profoundly our view of biology has changed.

I would also like to suggest something quite radical, namely that intellectual property in biology really should be based on the simple concept that biology is an information science, and what we should be patenting is information not mechanical entities or compositions of matter. So in a sense this presentation is partly hypothetical, but I would predict that the information approach represents the direction that intellectual property in the biological arena will be taking in the next ten years.

When I came to the University of Washington in 1992, I had an opportunity to meet Bill Gates. I remember, during one of the first conversations we had, he stated that information technology (IT) and biotechnology were going to be two of the dominant technologies in the 21st century, both from a scientific point of view and an industrial point of view. What I found interesting about this was the simple idea that both of these disciplines are basically about information. Information technology is about

---

* Professor and Chair, Molecular Biotechnology, University of Washington, Seattle.

the digitized information of the real world and our ability to manipulate it and understand it. Of course, biotechnology is about biological information, and I will talk about that more in a moment.

What has happened in the last five years or so, propelled primarily by the Human Genome Project, is that there has been a profound revolution in our ability to decipher biological information. I would predict that in the next twenty-five to thirty years there will be an even more profound revolution in our ability to manipulate biological information. Manipulate it in the best sense, to understand biology or to lead to a revolution in preventative medicine.

So in this context it is relevant to ask: what exactly do we mean by "biological information?" I would say that there are three types of biological information; and the essence to understanding this new view of biology is understanding each of these three types and, equally important, their interrelationships.

The first type of information is the information in your chromosomes and your genes, its DNA. This is a one dimensional language, four different letters. It's a digital language, just as your computer language is a digital language, encoded in the variations in those letters along the long strings we call chromosomes. One type of unit of information present on chromosomes is genes. Genes then encode ultimately what we call molecular machines. They make an intermediate product called messenger RNA, and that in turn can be translated into a final product which again is another string, a protein string—only this time the alphabet is much richer. There are 20 letters in the protein alphabet, rather than just the four letters in the DNA alphabet. What those letters do, what the particular order of letters in a protein has the ability to do, is to cause that string to fold into three dimensions and create a molecular machine. It is these molecular machines that catalyze the chemistry of life and give the body shape and form. When you look at another individual, virtually everything you see is protein. So in a sense proteins are the informational manifestation of genes. Proteins are the executive agents that carry information to the body. Indeed, what biology has been about for the last thirty years is the study of individual genes and individual proteins. This second type of information, in proteins, is three dimensional in nature.

The third type of biological information is going to be the information that will dominate the next century, that of complex biological systems and networks. The subset of your 1012 neurons, a million million neurons, with their 1015 different connections in the human brain, is a classic network. What is true about complex systems is that the interaction of their elemental

components confers the ability to give emergent properties, or systems properties. For the brain, these systems properties are memory, consciousness, and the ability to learn. The key point about complex systems is that you cannot learn about them by studying the characteristics of their individual elements in isolation.

This is an area that the technology is moving toward: the one-dimensional information of the gene, the three-dimensional information of the proteins, and essentially the four-dimensional—that is, time variant—information that is present in complex biological systems and networks. All this forms the framework for thinking about this in biology.

A second event that has absolutely transformed biology, of course, is the Human Genome Project, with its mission to decipher human heredity, to analyze each of the human chromosomes, and to find the order of individual protein letters in each of the long strings. There are 24 different human chromosomes. The number of these strings range from 50 million to 250 million, so deciphering the human genome is indeed an incredibly gargantuan task.

I remember the first official meeting on the Human Genome Project was held in the spring of 1985 when a fellow at the University of California at Santa Cruz had a $35 million contribution that he was considering giving to form an institute to sequence the human genome. He called together six or seven of us, to spend two or three days talking about this then quite outrageous proposition to decipher human heredity by sequencing the human genome.

Those six individuals were the real pioneers and entrepreneurs for pushing the Project for the next five years. What each came away convinced of were really two different things. First was the idea that this task represented a completely new kind of science, which I call discovery science. Discovery science is all about taking an object, in this case the human genome, and delineating all of its component elements—that is, sequencing the strings of DNA that are called chromosomes. Discovery science did that irrespective of any questions, any hypothesis-driven propositions. Of course, classical science has been done virtually entirely by hypothesis-driven ideas; you formulate a question and you design experiments to test that particular question.

It's turned out that the discovery-driven science has enormously enhanced and enriched the hypothesis-driven science. In fact, I would argue that the intimate interrelation and integration of these two different types of science will be fundamentally essential for our being able to do systems biology in the future. But it was this element of discovery science that led

most opponents to say that this was fake science, it was a waste of money, it was a fishing expedition, and it was going to threaten small science. Of course, one could understand those views in the context of those times when they didn't understand how profoundly biology would be revolutionized by the Human Genome Project.

What has happened in the intervening time is that we have come to realize that, in some sense, the Human Genome Project is perhaps the world's most incredible software program. It's a software program that controls the wonderful phenomenon called human development. We each started as a single cell of a fertilized egg and went through the complicated process of chromosomal choreography, where in different cell types, different subsets of 100,000 or so human genes are expressed to make a muscle cell different from a brain cell. This picture of the human genome leads directly to a statement James Watson made about ten years ago. It was that we used to think our fates reside in the stars, but we know now that our fates reside in our genes. The question is: to what extent is this really true?

To give you an idea of the other side of this idea, consider the left index fingerprints of two nine-year-old identical twin girls—they are completely different from each other. The genetic program for fingerprint development manifests from identical sets of genes, but, either when it interacts with the environment or the processes of development, leads to quite a different outcome. This poses really sharply the fundamental question that philosophers have long debated: what is the contribution of nature—that is, our genes—as opposed to nurture, the environment? I would argue that for every interesting human trait we have to answer the question. And indeed, we have very ineffective ways of precisely determining those answers.

The Human Genome Project has turned out to be about a series of maps: genetic maps that sprinkle markers across the genome so we can identify the location of genes that predispose to disease; physical maps that essentially fragment the chromosomes into small manipulable pieces and then attempt, for each, to solve the linear jigsaw puzzle that puts those pieces back together; a map that defines the location of all 100,000 human genes; and a final map, which is the sequence map, that determines the order of the four letters of the DNA language across each of the 24 different human chromosomes.

Suppose we project ourselves fifteen years into the future to when the Human Genome Project is done—it will probably be finished in its entirety in the next three years or so—and ask ourselves the really simple question: what were the major accomplishments of the Human Genome Project? The first is what I call the "Periodic Table Of Life." Just as the periodic table of

the chemical elements, in the 19th century, revolutionized certain aspects of chemistry by precisely defining their interrelationships, so the Periodic Table Of Life is going to revolutionize how we do biology. It will give us a definition of the genes and the delineation of the DNA sequences of those genes that regulate the turning on and off at appropriate times in tissues. We will have the ability to deconvolute those genes into their "tinker-toy" building blocks, their fundamental components. And it will give us access to the enormous amount of natural variation that is present in human population, called polymorphisms. It is this variation that separates us one from another and predisposes some of us to diseases of one sort or another.

The Genome Project has also done several other things. It has led to the creation of new kinds of tools which I call global tools. These are about being able to look at many genes at a time, or many proteins at a time, or many cells at a time. These are the kinds of tools that we need for the interesting paradigm changes that lead to the proposition that systems biology is going to be a central focus in the 21st century.

So let me talk briefly about several of these paradigm changes. The first we have already discussed: the simple idea that biology is an informational science, with the three different types of biological information. This is central to our discussion later of intellectual property and how our views of biological intellectual property should change.

The second paradigm change is again an explication of the simple idea of dominance of systems biology in the next century. Indeed, that dominance presupposes a particular view of how we analyze systems. So what we have to be able to do is take a complex biological system, and if it is really complex like most biological systems are, use biology to divide it into subsystems whose properties still are observable in nature; to be able to use the discovery techniques to identify the elements that are present in that subsystem; to use genetic and biological markers in model organisms to follow the flow of information in those networks and informational pathways. We can then take that information and ultimately create mathematical formulae which tell us about both the structure of the informational pathway, and about the systems properties that are created. Of course, these latter two objectives will require the invention of new types of mathematics for joining the information from global technologies and systems theory into equations that can make these particular predictions. Indeed, we here at the University of Washington are collaborating with theoretical physicists, computer scientists, and biologists to carry out these kinds of initiatives.

These global tools that I've talked about constitute an important part of the basic material for being able to approach systems biology. The tools of

genomics, of which there are many now, allow us to look at many genes at a time. The tools of proteomics allow us to look at many proteins at a time. Indeed there are a whole series of other technologies, but let me illustrate two.

In 1986 we developed the first prototype DNA sequencer, a machine that can decipher human heredity. We basically did it by color coding with flourescent dyes the four different letters of the DNA alphabet. We then devised a simple way, using previously developed chemistries, to essentially read out the color and order of each of the nucleotides as you march down a particular stretch of DNA. So, from the first machine that we developed in 1986 to today, there has been a more than 2,000-fold increase in our capacity, and there has been more than a 200-fold decrease in the expense of DNA sequencing.

A second tool that is illustrative of these global technologies is the simple DNA chip. We now have the ability, in principle, to synthesize on a single chip a short fragment of DNA of about 20 letters that represents each of the 100,000 human genes. We can then, by virtue of molecular complementarity (the two strands of DNA joined to one another), use these chips to analyze the expression patterns in normal cells and in cancer cells, to ask all 100,000 genes which change their patterns of expression in this process. So here truly is a global tool which gives us the potential to look at the behavior of every single human gene, in normal states and in diseased states as well.

A fourth paradigm change is the idea that we have to use model organisms to do these assays of biological information. Indeed, the Human Genome Project has delineated five different organisms whose genomes will be finished. Four of these are simple: bacteria, yeast, a simple nematode, and a fly—and three of the four of these are done. The fly will be finished within the next year. In all of those cases we can find a striking homology between human genes and their counterparts in these simple organisms.

For example, the genome of the nematode was recently finished. It turns out that it has 20,000 genes, and 70% of those genes have obvious counterparts in humans. So if we want to understand how a human gene works, we can go study it in a nematode that is biologically and genetically manipulable. Even more important, we can understand how that gene works in the context of its informational pathway.

Now the mouse is the fifth organism; its genomic complexity is approximately the same as our own. It is a critical model system because only in the mouse will we see complex traits of humans—the human nervous system, human development, the operation of the immune system—in life.

So these model systems truly are the Rosetta Stones for deciphering biological complexity. One of my favorite quotes from Matt Stelbrook: "Any living cell carries with it the experiences of a billion years of experimentation by its ancestor." That means that we just have to be able to unlock, in various ways, those experiences, and we have the tools to do that in a remarkably effective fashion.

The final paradigm change is the absolute revolution in biology that has come about as the applications of computer science and mathematics have been applied to it. These applications give us the ability to deal with biological information in very sophisticated ways; to be able to acquire it, store it, analyze it, display it, to integrate the various forms of information and prepare them to model these complex systems and networks, and ultimately to disburse the information.

On the other side of the coin, living organisms have had 3.7 billion years of evolution to learn how to manipulate their digital strings, the chromosomes, and they're beginning to teach us, as we gain further insights about chromosome mechanics. They're beginning to teach computer scientists much better ways of thinking about how to manipulate digital information in interesting ways. Certainly one of the exciting ideas that comes from the Human Genome Project is that we can take the entire genome of an organism and, with computational tools, in time deconvolute it into all of the informational pathways that the organism uses in its logic of life. Later on, we will also be able to compare the genomes of multiple organisms and ask how those logics of life have changed.

The question of the logic of life in the intellectual property that is involved with whole genomes presents absolutely staggering problems for intellectual property, when you think about how you would patent a genome, or what should you be patenting with regard to a genome.

One of the areas that we are very interested in is the immune system. We now have cells that we can trigger to specific systems properties: an immune response, something called tolerance where the shell of the cell is shut down, cell death, or even auto-immunity. With the techniques that we have talked about today, we can interrogate the incredibly complicated informational pathways that are involved in each of these systems properties. In time we can gain fundamental insights that will lead to, on the one hand, a profound understanding of the biology, and on the other hand, the ability to manipulate the diseases that emerge from the immune system.

It is interesting to compare the growth of computer chip technology with that of DNA sequence information. Consider Gordon Moore's 1970

hypothesis that the number of transistors you could put on a computer chip would double every 18 months. In fact, that has been true for 30 years, and that simple observation, as much as any, explains the incredible explosion in the IT area. In comparison, there has been an even greater exponential increase in DNA sequence information. It will be a driver for biotechnology every bit as great as Moore's Law has been for information technology.

Clearly the challenge that stands here is that while the exponential increase in sequence information gives us information, we will still have to have the wisdom to be able to turn it into knowledge. I would argue that there is a difference between information and knowledge and that intellectual property should not be given merely for information; some of the knowledge should be inscribed into the strands.

What is absolutely critical, if we are to convert sequence information into knowledge, is the integration of all of these global technologies that we have talked about. This is an enormous task, and it's just beginning at this particular point in time. Indeed, we're in the process of setting up, here in Seattle, an institute that is committed to systems biology, and we are going to do that by having a cross-disciplinary faculty that can invent the global technologies of the future and then new types of partnerships with academia, with industry, and even with society.

One of our enormous interests in this enterprise is the whole question of intellectual property and where it is going in the future. So let me now, with this very long background introduction, give you a few of my thoughts about intellectual property and biology.

The idea that biology is an information science is the most fundamental tenant of biology today. I would argue that patents really should focus on actually patenting units of information, and even units of information that, in part, have been turned into knowledge. Clearly, however, one of the challenges is, if we can now define biological systems, how do you patent a system? Alternatively posed, if in fact others have prior patents on individual elements in the system, how do they affect the system's patents that you generate as a whole? Knowing about the individual elements doesn't, *a priori*, tell you one iota, necessarily, about the systems property that emerges from having complete a biological system or network. So knowing the system as a whole adds enormously to the information and to the knowledge—and clearly there must be a way of patenting biological systems.

The other thing that is important to realize about biology is the enormous hierarchical nature of its information. We can start with the gene, which is the fundamental coding for a particular unit of function. We can go

to the protein, which is the next higher level. We can go to the informational pathway in which that protein manifests itself. Or we can go to the whole assembled network of interconnected pathways. So the question is, if we take a patent out on a gene, at this lowest level, to what extent does it dominate patents at any of these higher levels, where you generate completely new types of information which, *a priori*, can't be predicted from any of the single elements? If we even take the unit of information at a single level, such as a protein, we can see that it has enormous potential for modification. It has a linear sequence of subunits, letters that are the protein language. It has a three dimensional structure. Its behavior can be modified by altering the chemistry of certain amino acid subunits. It can be processed in various ways. It can have different rates of turnover. It can interact with small molecules and other macromolecular molecules. It can be localized to different compartments. So does knowing one little bit of information give you fundamental insight into the many different roles and functions that a protein can play in this particular context? I would argue that it doesn't. Therefore, in some ways we have to be able to deal with the hierarchical nature of information, and if patents occur at the most primitive level, not to have them totally dominate the higher hierarchical levels.

Yet another hierarchical level of information appears in so-called "gene families." Zinc finger proteins control process and transcription. They play an active in turning genes on and off. In human beings there may be as many as 1,000 zinc finger proteins. So you can look at the lowest level and get general motifs that will define all zinc fingers. Now you could patent things there and control all 1,000 of the proteins, or you could go up to successively higher level branch points and define sequences that specify the features of various different branches. You could also go up to the very terminal tips. The point I'm making is that depending on what you patent, in theory, very primitive patents can have enormous dominance over many different hierarchical levels of proteins, or genes in gene families.

These are some of the issues that I think are really worth thinking about: the idea that we ought to patent information rather than composition of matter; the idea that a second type of information in the context of human genes, which really hasn't been discussed at all, are those sequences that regulate, that turn the genes on and off. There are, obviously, perfectly reasonable sequences for patenting as well.

I think a really important point—and the place where the patent office has really caused difficulties—is that it is important to patent techniques because they cost a lot to develop. For example, to develop the DNA sequencer that we finished as a prototype in 1986 into a commercial

instrument cost Applied Biosystems on the order of $80 million. That obviously was an investment that required an exclusive right on that particular kind of technology. However, what has tended to happen with certain types of technology, DNA arrays is one, is that a patent may be granted so broadly that it dominates every single way you could even think about doing the DNA arrays. I think patents that are granted in too-broad context are enormously damaging and certainly they are suppressive of new types of innovation.

A final point that I would make is that the definition of a gene has become enormously confusing these days. For example, the gene contained within a single area—because it is made up of multiple coding regions and intervening DNA sequences that at the RNA level get spliced together—that gene can be spliced together in many different ways. There are examples where, from one gene, you get 50 different proteins. Therefore, if you have patented the gene, have you patented all of those proteins, despite the fact that they may do 50 different kinds of functions? Again, this is the issue of hierarchical levels of information and the question of whether the more primitive levels should dominate the more sophisticated levels.

So these are the questions which I think are really going to dominate the scene in the next 10 or 20 or 30 years. I would like to see some fundamental changes in the patent law, changes which move us from the mechanical view of biology (*i.e.,* composition of matter) to an informational view of biology which takes into account the kinds of complexities that I have discussed here.

Audience Member: Would you elaborate on the informational vs. the "composition of matter" concept there? If I thought of patenting a unit of information or thinking of a biological entity or component as being on software, then I would probably write claims like I would in a software patent. I think that the objective that you're describing is driving the comment—mainly that you want claims to have a certain breadth to them and you want a certain policy result to happen. Could you elaborate a little more?

Dr. Hood: Sure. Let me talk about patents for express sequence tags, the so-called EST's. An express sequence tag is the DNA sequence of a fragment of an expressed gene, and that sequence may constitute 2% of the entire chain. The idea is that companies that have done a lot of express sequence tags, on many different human genes, have really pushed the patent office to say "Look, there is a legitimate object of information that

should be patented." The argument they make is that sequence gives you the address that allows you to locate the precise gene on the chromosome, and that certainly is true.

However, the interesting question—and it looks like the patent office is indeed granting patents in that particular rubric—is that of hierarchical levels of information. If I have the patent on the EST for EPO, do I control EPO, when Amgen does the complete sequence, does all the biology, and demonstrates that this is a terrific drug? Alternatively, do I have pass-through rights such that Amgen must pay me a significant fraction of their income just because I generated 2 million EST's and this happened to be one of them? In that case, I'd be getting a patent for a piece of information that had nothing to do with biology whatsoever. What I would argue in this case is that there is information in that EST, but it is information that has nothing to do with biological function. Therefore that EST patent should not in any way dominate subsequent patents that come from a detailed understanding of gene structure and the function of the gene. That's an example of where the composition of matter patent may, in this case, give you a dominance to a patent at a much higher hierarchical level of information.