12-2-1988

The Human Genome Projects: Issues, Goals & California's Participation

Joint Committee on Science and Technology

CALIFORNIA LEGISLATURE

JOINT COMMITTEE ON
SCIENCE AND TECHNOLOGY
SENATOR JOHN GARAMENDI, CHAIR

Hearing on

THE HUMAN GENOME PROJECTS: ISSUES, GOALS & CALIFORNIA'S PARTICIPATION

December 2, 1988
Molecular Biology Institute
University of California, Los Angeles

BEFORE THE JOINT COMMITTEE
ON
SCIENCE AND TECHNOLOGY

IN THE MATTER OF)
)
THE GENOME PROJECTS:)
)
ISSUES, GOALS, AND) Senator John Garamendi - Chair
)
CALIFORNIA'S PARTICIPATION)

ORIGINAL

TRANSCRIPT OF PROCEEDINGS

Friday, December 2, 1988

Molecular Biology Institute
Room 159
University of California, Los Angeles

AGENDA

THE HUMAN GENOME PROJECTS: ISSUES, GOALS, & CALIFORNIA'S PARTICIPATION

Friday, December 2, 1988
Molecular Biology Institute - Room 159
University of California, Los Angeles

8:30	REGISTRATION AND CONTINENTAL BREAKFAST
9:00	**MORNING SESSION: NATIONAL GENOME INITIATIVES** JULIUS KREVANS, Moderator - UC San Francisco
9:15	THE NIH GENOME INITIATIVE JAMES WATSON - Cold Spring Harbor Laboratory
10:00	THE DOE GENOME INITIATIVE CHARLES CANTOR - Lawrence Berkeley Laboratory
10:45	COFFEE BREAK
11:00	THE GENOME PROJECT: CONGRESSIONAL & INTERNATIONAL PERSPECTIVES ROBERT COOK-DEEGAN - Office of Technology Assessment
11:15	PANEL DISCUSSION KREVANS - MODERATOR WATSON, CANTOR, COOK-DEEGAN
11:45	LUNCHEON THE IMPORTANCE OF COLLABORATIVE RESEARCH TO THE STATE'S ECONOMY JOHN GARAMENDI - State Senator
1:00	**AFTERNOON SESSION: CALIFORNIA'S PARTICIPATION** WINSTON SALSER, Moderator - UCLA
1:10	THE CALIFORNIA COMPETITIVE TECHNOLOGY PROGRAM KENNETH GIBSON - California Department of Commerce
1:25	PANEL 1: POTENTIAL IMPACT OF THE GENOME INITIATIVES ON BASIC RESEARCH NORMAN ARNHEIM, Moderator - University of Southern California DAVID COX - UC San Francisco GLEN EVANS - The Salk Institute WALTER FITCH - University of Southern California ROBERT MORTIMER - Lawrence Berkeley Laboratory
2:25	PANEL 2: DEVELOPMENT OF NEEDED TECHNOLOGIES, HARDWARE, AND SOFTWARE ANTHONY CARRANO, Moderator - Lawrence Livermore National Laboratory NEBOJSA AVDALOVIC - Beckman Instruments ELBERT BRANSCOMB - Lawrence Livermore National Laboratory TIM HUNKAPILLAR - California Institute of Technology MICHAEL WATERMAN - University of Southern California NORMAN WHITELEY - Applied Biosystems
3:25	PANEL 3: UTILIZATION OF THE DATA GENERATED BY GENOME PROJECTS ELIZABETH NEUFELD, Moderator - UCLA A. STEPHEN DAHMS - San Diego State University MICHAEL ESPOSITO - Lawrence Berkeley Laboratory MICHAEL KELLY - Intelligenetics Inc. THOMAS MARR - Los Alamos National Laboratory LARRY SIMPSON - UCLA
4:25	SUMMARY COMMENTS
4:40	ADJOURN

Sponsored by:

Joint Committee on Science and Technology
UC Systemwide Biotechnology Research and Education Program
Lawrence Berkeley Laboratory
California Department of Commerce

INDEX TO SPEAKERS

Joint Committee on Science and Technology

December 2, 1988

9:00 a.m.

- - P R O C E E D I N G S - -

DR. BOYER: I'm Paul Boyer, head of the University of California Systemwide Program of Research and Education in Biotechnology, one of the cosponsors of this meeting.

And I come with a perspective which in one way is better than anyone else in the audience, and that is of 55 years since I received my Ph.D. in biochemistry. And at that time, the possibility that some half a century later there would be a conference on looking at the possibilities and problems of sequencing the human genome was just utterly beyond my imagination.

What has been accomplished in our science has been really a revelation. The wisest decision I ever made was to become a biochemist at the time that I did.

But you don't want to hear about my background and perspective. We're here today to look at one of the most interesting challenges in biology, to examine the pros and cons.

We're fortunate to have as a chairman of an excellent morning panel, Dr. Julius Krevans, Chancellor at the University

of California, San Diego --

DR. KREVANS: San Francisco.

DR. BOYER: -- a fine clinician -- University of California, San Francisco -- Professor of Medicine for a long time there, Dean of the college before he became Chancellor, and who has had a keen interest in the development of the biotechnology and related facets that come out of our studies. And it's a pleasure to introduce Dr. Krevens who will moderate the morning session. And I apologize for moving you south.

DR. KREVANS: I'd take your land, but I don't want your science, Paul.

I would welcome all of you to this session on the Human Genome Projects. And the title right off the bat is a little disarming. Is it going to be a project? Is it going to be projects?

And what we're really here this morning and this afternoon to share is a discussion of how and where we should go in this exciting prospect of sequencing the human genome.

Now, I'm going to make the assumption that no one in this audience thinks it would be a terrible thing if we knew the sequence of the human genome. I don't think that we should make the assumption, however, that there aren't people in the United States or in the world who think it would be a terrible thing.

There remains a potent, and perhaps even growing, group of intellectual Luddites who find the prospect of new knowledge terrifying. And the idea that a group of people are interested in finding out what the sequence of the human genome to this logic is terrifying and awful and something to be resisted.

I grew up in New York City -- one of our next speakers, we were talking about it -- and I recall reading about what New York City was like in the 19th century. We used to study history in those days in the public schools. They've now given that up as you know.

And when I hear people railing against the internal combustion engine as having destroyed the world, I would say to myself that the little knowledge I have of the history of New York that without the internal combustion engine, given the means of transportation which preceded it, New York would be at least shoulder high in the same substance of which the logic of the nay-sayers is made.

At any rate, the interest of the University as a partner in this symposium is a genuine and important one. Biomedical science, biomedical research in the United States is a result of a social contract between the people of the United States as expressed by the legislature and the government and a variety of institutions, independent research entities, the National Institute of Health itself and the universities

of the United States to produce new knowledge.

So we have a stake, and an important one, in the answers to the question as to how and when we might embark on a project as exciting as knowing the sequence of the human genome.

We're fortunate to have on the program today some very interesting and, I think, informative perspectives. And the first of those will be given by James Watson who is formerly of a lot of places.

His office sent me an extraordinarily brief CV. This is the James Watson CV.

Currently, Jim Watson is looking at the human genome initiative for the NIH. He has, of course, been Professor of Biology at Harvard. He's been the Director of the Cold Spring Laboratories. He did some interesting research work at Cambridge and is one of the people who changed the way we even think in biology.

So, Jim, we're looking forward to hearing your remarks.

DR. WATSON: It's a pleasure to see such a small group of people. I got used to the fact that if I go to a university that everyone in the audience is not there to listen to me but to get my autograph. And it's very inhibiting and makes me feel rather creepy. So I realize now that I'm talking to people who actually want to listen

to what I have to say.

Now I have the position of Associate Director of the NIH for genome research.

And can you hear me all right? I speak softly and if someone could --

[At which time the microphone was adjusted.]

Is that better now. Okay. And if you can't hear me in the back, please yell because I get lower and lower, probably because I don't believe in what I'm saying so if it doesn't go very far I won't get into trouble.

Anyway, I have now a position now at NIH. I'm also at Cold Spring Harbor. And I have committed myself to spending two days a week on the Genome Project which probably won't be enough, but if I spent more, I'd have to resign from Cold Spring Harbor and that would put me in a bad position because it's very useful to have two jobs because if you don't like what you're doing, you can say that you'll resign and you'll have another one, which I found useful when I was both at Harvard and Cold Spring Harbor.

Those I have a little say because I found -- at least at Harvard, not very often, and even at Cold Spring Harbor -- that occasionally you had to threaten to resign, not very often, but sometimes issues really count.

Now, I felt -- I mean there's the two talks, one

by Charlie Cantor and one by myself -- and I feel slightly weak because Charlie actually knows what he's talking about and he's a pro in it. And I've been brought in, I think, really to make decisions as to which pro's to listen to.

So I don't know anything, and I'm going to try and learn something. Eventually, we can get the project done fast.

Now, I felt -- to me it has always been pretty obvious we should do the project. But a lot of people, at least to start with, felt it was a scandal, and really intelligent people.

So a lot of things have been written which I think we won't drag out to embarrass them later in the future.

I thought maybe I would give a history of sort of where I see the genome project and why it's so necessary and what are the phases of genetics, which I've been in some form or another, I guess, now for 45 years, since when I went to the University and decided I wanted to find out what the gene was.

Now, if one looks at genetics, I think the first phase was the classical Mendelian one, which was gene mapping. At the turn of the century it became clear that hereditary was controlled by discreet units, which got the names genes, and that these were located on chromosomes.

And really first through the work of the fly group

at Columbia, it is clear that the genes were arranged in linear order, which sort of made sense because the chromosomes were linear. And this phase of people mapping genes goes on today. And what we're talking about is the ultimate in gene mapping, sort of the map of the nucleotide level.

But sort of showing the traits were controlled by genes is the sort of dominant theme of genetics for about 40 years. And one could say that that sort of phase in which genes were discreet objects, where you didn't think in terms of chemistry, we saw the banding of the Drosophila chromosomes in the early '30s. And Calvin Bridges made a complete map of the Drosophila chromosomes in the late 1930's at Cold Spring Harbor. So that was the first ultimate genetic map.

And it wouldn't have been except that the director had money from the Carnegie Institution of Washington. Bridges came to Cold Spring Harbor and laboriously did the banding pattern.

Now, the second phase of genetics was really: What is the genetic material? What was it chemically? And that was interconnected with not only what the gene was, but what did it do? And that was going full blast when I became a student and I learned the slogan "One gene, one enzyme." And it was really at Cal Tech where I guess it was promoted the best, but the idea really goes back to

the metabolic defects in man going back to around 1910, and Haldane talked about it and other people talked about it.

And when I was a student, it was -- I heard about DNA. And then when I went to Indiana and took Luria's course on viruses, I certainly learned the details of transformation, and that focused me on DNA.

Most people weren't focused on DNA at that time, I guess for a variety of reasons. One was that even at that time one didn't know the chemical structure of DNA. And suddenly about 1951, '52, there was unambiguously established to be a 3-5 linkage. So one couldn't have made the model of DNA before '51 because the chemical structure wasn't there.

I think the number of people who were interested in DNA at the time we found the double helix was at most a couple hundred. And they were interested for a variety of different reasons. And most people had a sort of: What is the gene DNA? I think the geneticists would have said, "Yes," and the biochemists would have said, "No."

The biochemists would have said, "No," because they were all protein chemists, and they wanted proteins to be important. So they preferred to think of DNA as a sort of unspecific scaffold into which the interesting things, the genes, were sort of attached. I think that was the

conventional viewpoint.

Paul, I guess, says that the luckiest thing for him was that he became a biochemist. I guess you could say that the luckiest thing for me is that I didn't because I probably would have become a protein chemist, very interested in proteins whereas, in fact, geneticists knew that the chromosomes were filled with DNA, and so they were more likely -- I mean you could see Hermann Muller, he would have taken the discovery quite seriously.

Anyway, that phase -- What is the Gene? -- ended in '53 when we saw the gene had the complimentary double helix, because if the DNA had the structure it should have, it was going to replicate itself. And that was the question which the geneticists had raised: How did it replicate itself perfectly?

So there was Cold Spring Harbor symposium in 1953, and everyone agreed -- there was only person, and it was really odd, he didn't like it, and that was Barry Commoner. He just was sort of mad. I think in more than one way. So that phase of genetics ended as to what the gene was chemically.

And then the next phase was how the gene acted and, essentially how you translated the four-letter alphabet, four-letter messages of DNA into the 20-letter messages of proteins. And that went incredibly fast, from 1953 to

1966 when the code was established and you could say we had our Rosetta stone. You could translate it from one language to another.

There are still details being worked out, but it was sufficiently a big event that people reacted to it in several ways. Some people thought the field was over and so moved on to the brain or embryology.

And in retrospect, now you can say that they were pretty silly. But, in fact, if we come that DNA hadn't come along, the DNA sequencing, you might as well be working on the brain because we weren't going to get very far.

But the next phase of genetics, which really now you see, was made possible only by the discovery of recombinant DNA and by working out the Sanger and Gilbert sequencing methods was: What controls gene functioning? And I guess that's the dominant theme of genetics now. How do genes function? What are their control? And there are all sorts of protein factors. And it's moving very fast and unbelievably competitive because now, in a sense, it's almost easy to do because you can work out the sequence of DNA and do lots of things.

When we know these rules for gene functioning, then we will now permit, you can say, the ultimate aspect of genetics, that's really genetic manipulation and genetic engineering. And that's what many people in this room are

interested in, and you could say it's the end of genetics. I mean, if we know what the rules are. And then you can go out and make new sorts of organisms.

You could then, I think, go to the next stage, whether you call it genetics or not, it is arbitrary. I guess I wouldn't.

Now one can actually approach the question of: Really what is a cell? I mean really if all the information for a cell comes from the DNA, if you can go to the DNA level, you will see the instructions for how a cell can divide, and you can go beyond that in embryology. Within the DNA are all the instructions for embryologic development. So if you can look at the DNA level, you actually ought to be able to work out the way development occurs. And that will be very complicated.

And you have to -- if you say that a fly develops, you have to know what is a fly. And I guess you could look at all its proteins and say that a fly is a sum of 20,000 proteins or something like that. But now that you can go to the DNA level, if you wanted to find what a fly is, the simplest thing is just to get the DNA sequence of Drosophila. This will tell you how many genes there are. You can begin to say that some are DNA binding proteins and others will be any number of -- anyway, you can begin to get an idea of what a fly is.

And if you have all the genes plotted out, you can find which time they function during development, and we can completely describe a fly. And so there's really nothing more than the instructions in its DNA. And I like that idea, maybe it's because I have no affection for flies and the fact that there's nothing but DNA seeing that the right proteins appear at the right time, and we'll understand the fly. It's complete reductionism.

And there probably isn't anyone who would say there's more to the fly than its genetic instructions manifested itself through development. But it will be extraordinarily exiciting, but to really do it, you need the DNA sequence.

And if you go to humans and say, "We really want to understand ourself. We understand all sorts of different aspects of ourself," but you want the total picture, you might as well go the DNA and say: What is man or woman? And it's really the instructions from the DNA.

Some people say that it's more than that. I don't want to get into any arguments, but I'll be satisfied to know what's in the DNA.

So as a biologist, the prospect of now seeing problems other than genetic problems, in terms of just the instructions of the DNA, is to me very exciting. And nothing could be more exciting than to -- you know -- if you were trying to say, "Well, how does the Drosophila brain function?"

Well, you want to know all the proteins in it.

And you can lead to some people saying that they are dull projects. But I don't see any other way to go about it. So I think we have to go to the gene.

If we want to go to these big problems: What is a cell? How does it divide? You've got to know what a cell is. In the same sense, we had to know what DNA was before we could say what a gene is.

So if you want to ask what human being is, well, it's all these genes correctly functioning. So you better find out the genes. Cancer would be just still as bad as ever to work on if we hadn't been able to go to the DNA level. When you can go to the DNA level, you can define your problem.

So if you want to define embryology, you just better go the DNA level.

And so I think working -- and, of course, we've done it in the case of viruses where we got the complete sequence alligned. And it's only in the -- you could say unsatisfactory because lambda replicates in the coli cell, and we don't have the sequence for coli. And until you have the coli sequence, you won't really be able to understand it completely.

So I guess that's my general reason for saying we have to know the sequences if we want to, say, go on

to embryology.

So if you want to say, "Well, who's going to control the yeast world? Who's going to be on top of the Drosophila world?" I would think that I wouldn't want someone else to own the sequence in the sense that I think you really have to get at it because I think we're going to get lots and lots of surprises. And we ought to go forward.

Now, if you say, "Well, why until recently was all this controversy as to whether we should do it?" Well, I think common sense was that we should. I think the reason was that we talked about sequencing the human genome. And no one can really sequence it today, so it's not really a question of it's too expensive or we don't have the right techniques. And so as long as you see the sequencing, you see people wasting money, and knowing that the only people who are going to do massive sequencing are sort of idiots. And no one wants to see a group of idiots consuming a large fraction of the national budget.

So it was probably very bad public relations, or at least confused the issue by saying that one is going to sequence it; whereas, in fact, one should say, "Who's going to map it?" and really a good high-resolution genetic map. And when you say that to people, I can't see anyone who doesn't want the genetic map.

And that has to be done first, and it needs an

organized program. So you're going to have to set up a program to do the map; whereas you got into all this controversy: Why spending all this money sequencing junk? which Joe Gall, a very intelligent person, managed to waste his time by writing that letter to Science saying that it was silly to sequence junk.

The only problem is that you don't know where the junk is. So whereas it sounds sensible, now that you have introns and ectrons, and you don't know the borders between genes, it would cost more money to find out where the junk is than to sequence the genome. So that would be an even more ridiculous sort of effort, I think. You have to do the map.

And not all -- you could say -- why do we -- what are the problems in getting the map? And I guess there are several -- and Charlie will really talk about it -- one is that we don't really know how to do it yet. But there are a lot of ways that might work. And I think the sort of perspective that we should have is that we really want the human map in roughly five years.

If you said 10 years, really you will waste five not making real decisions. But if you say five years, then we will then have to even organize it and you will be sensible. And that says that if we get the map in five years, then, hopefully, someone will have clever ideas on DNA sequencing

which will reduce the costs so that we can do it.

That doesn't mean that we shouldn't try and sequence a couple megabase fragments because that probably won't inspire people to do it, and you'll actually find out the true cost. Everyone says a dollar a base pair, but I think it's really two if you look at it closely. And if you do it twice, it comes out to -- you know -- you're up around $10 billion. And no one wants to spend that. At least I can't see the reasoning.

The map -- if you can't do it for $200 million, we're really in pretty bad shape, which means three to five years effort. And we have to do three maps. And Charlie will say -- and I'll just emphasize too -- we really need the high-resolution genetic map.

And there we have to, I think most of us, rely on Maynard Olson's experience from yeast; and he says you have better have a mark for every million base pairs. And so to do that, there are various estimates that will cost, say, $50 million based on how much it costs now, but maybe we could do it for less.

What has happened so far has been fairly interesting. It really got off the ground only because of the Howard Hughes supporter, Ray White, in Utah. And, secondly, Collaborative Genetics spent some of their own money to do it in Boston. And they spent their own money only because some stupid

study section at NIH turned down the application to pay for it out of NIH money under the theory that I guess it was dull repetitive work and didn't show intellectual sparkle.

And so one has to have great respect and see the need for peer review committees but also to be aware that sometimes they make terrible decisions and generally don't think heroically, and sometimes rather pedantically, I think.

So we're lucky that the Howard Hughes money was there to support Ray White and lucky probably also that Collaborative made the business decision and spent $10 million getting it going.

The only problem is that we've got two competitive groups. And as everyone knows, we've got to get one map. And we're going to support both of these groups to go on, and we hope that they will integrate their data.

If they don't, then we have to in some way, I guess, either decide to support one at the expense of the other or withhold the money until they get together. But I think that we should just assume that they'll get together, and we'll get the map. And in five years, we'll have it.

Then we have to have the overlapping fragments and those models. And later whether that can be done by clever tricks involving cosmids, or whether YAC's would be the procedure is what we have to find out. And grants have been given out, and I would think there's a time frame of

about within one to two years, we're going to have to decide what to push. And right now all you can do is push everything. But at some point, I think it would be irresponsible if you really think one method is much better than the other, just to support the other because it happens to be in some state. You want to spread the money around. It's nice to spread the money around, but on the other hand, we want to get the job done, and there's got to be some balance between actually seeing that someone can do it and then putting your faith in them and getting it done.

Now, as to the organization of NIH, we have money first in fiscal '88, and now we have money in fiscal '89 to the tune of roughly $30 million. And the projection and hope is that it will be above $50 million in fiscal '90, which should allow us to do quite a bit.

Up to now NIH has just sent out -- sort of notified people that will support genome studies, and a lot of grant applications have come in and special study sections have existed, and the money has been passed out in the absence of any real plan for how to do it most efficiently.

Now there has been created an office of genome research at NIH, and there's an advisory committee that has been appointed, and we're to have our first meeting on the 3rd and 4th of January. And I would tell you the members, but I can't because until they've all accepted,

there's some government rules that we can't announce them, and some of them are obvious, and I think it's a very strong and good committee.

And the function of this committee, I think, is really to see that a plan is drawn up by the right people for how to proceed and get the job done, so a real plan. And at the NIH-sponsored meeting, and less than a year ago there was talk that we should have a rolling five-year plan and that we should update it every year. And I think that's a good idea.

Now, it would be very simple if it was only NIH, but it's also parallel effort in DOE. And there's been a memorandum of understanding between the two agencies for a joint committee. And I would hope we will meet fairly soon. And since there is really one genome, I think there should be a joint plan. I can't imagine the two groups having two plans. So this joint committee may be more important than the individual committees. I hope it is because that would indicate that we're working together.

The money of NIH is -- we have an advisory committee, but there's no authority to give out the money, which now has to go through General Medical Sciences. If the money increases, we would hope that the authority will be given so that the genome office can make grants and contracts itself.

That's a big step because you really have to create a new bureaucracy. And you don't really want to get into it until there is sufficient money to justify it. But if the money goes up, then I think such an office will be created, and then this office will begin to really look at the study sections and probably give them more direction.

And to use a term which frightens a lot of people, I think we'll have to go to contracts because it's really -- we're not giving a grant for $1 million to do what you please. We are saying that we're giving $1 million to get the coli sequence or something like that. I know that there are a lot of people for whom the word contract sounds -- smells of mediocrity.

But a lot of this work is going to be, hopefully, routine. And it should be managed, and probably a contract is the best way. So I'm sure that there will be a lot of controversy to start with as to whether we should go to contracts or keep it all on grant money. It all depends on your contract offices. So it depends on whether your bureaucracy is good or bad.

Besides the mapping, and something which really is contract things, is going to be the whole question -- at least to start with -- of the data bases which have been supported in a variety of ways by General Medical Sciences, by Howard Hughes. They were started when the problems were

relatively simple. And everyone has agreed that they're going to have to be totally reorganized. And who is going to handle them isn't clear. A bill was passed giving the National Library of Medicine considerable sums of money, and in a sense they're a natural body to have something to do with the -- at least the dissemination of the data once it is collected, and I think that may be the area where we're going to have to make firm decisions the soonest.

On mapping, we don't really know how to proceed, but in something like a GenBank and so on, I think a major problem, which is how to give what is inherently very dull work -- interest very intelligent people in very dull work, which is sort of what data bases are. If you say, "Well, it's too dull for me to be worried about," then it's almost like nuclear power plants, they're in charge of people who know what intelligent would be in the control room waiting for a crisis which develops every 20 years.

And we have a separate data base system for the protein sequences and for the nucleic acid and so on. And they have got to be put there, which means we somehow have to get people that everyone generally respect to make the decisions so that they will stick.

If the decisions are made by ad hoc committees of people who really aren't that interested, but just show up, and then you decide whether GenBank has another year,

and on and on, and I think a lot of -- we could waste a lot of money.

And I think that we have two things that we want to ask for and need a lot of money, but we really have to realize that it would certainly be irresponsible to other people if it could be pointed out that our research -- we give money to grants to people who wouldn't get it through an R-01 system.

That's why I'm actually pleased that we have such a good advisory committee for NIH because I think that we can't stand as the conscience to see that this dull work is not done wastefully.

Now, maybe the last thing I should say is that all our discussions of the relative roles of NIH and DOE all sounds like we're the only people who are doing it; but, in fact, the project has excited enormous interest throughout the world, and we don't own the human genome. And other nations are going to be involved, and we're going to have to collaborate with a lot people.

And toward that end, we've set up a human genome organization. Johns Hopkins is the head of it. And it's aim, I think, will be certainly to keep everyone informed of what everyone else is doing through holding meetings and, hopefully, actually to have an even more greater world plan for how to do it so that we don't get totally in a

competitive situation.

I see no reason why within the United States there really has to be competition, I mean in a sort of nasty sense. Naturally, some groups will work better than others, and they will be rewarded by having bigger roles. But in the world we really have to somehow get together. And the first place that we're going to have to do something is on the data bases and seeing that they're set up.

So I expect that I and I'm sure Charlie will spend more time than we want to going to meetings about seeing that the data bases are done correctly.

Ideally, the cost of the project could be shared, one-third by Japan, one-third by Europe and one-third by the United States, or roughly some sort of thing. Whether this will happen, I don't know.

We both, I guess, have a sort of fear, unless we act intelligently at least, that the whole sequencing end of the project will be done by the Japanese. On the other hand, when you get near them they say, "We don't have any money to do anything, much less help pay for anything."

So I don't know what the final answer should be, but I think that if we're going to provide data to the other parts of the world that it should be reciprocal at least in terms of cost. We just shouldn't send all our genetic maps across the Pacific at the expectation that it's a free

gift.

I'm sure that there will many tortured negotiations, which always have some reason for not succeeding. But I can just hope that we can work together. And at the level of individual scientists, the Japanese that we know, we like and they're a pleasure to be with.

It's just that everyone sort of expects that the United States has paid for science, and NIH has paid for science, and they helped the rest of the world. But the rest of the world is as wealthy as we are and, therefore, I think we've got to work to try and actually get some real support.

Europe, I think, will come together. There's an initiative in the EEC and probably, I would guess, would be the body that would finally provide some money.

Now, as to my last thought, it's on this sort of ethical issues, which I think are important because the general public -- we should work hard to educate them as to really what this information is going to tell us. Many of them will remain -- no matter how we try -- essentially ignorant of it. And we're going to have to constantly fight a battle of people who don't understand what we are doing or really don't want us to do what we are doing.

And I think that we should, at the level of the genome program, pass out money, and I think it should be

one of the functions of our advisory committee as to how much, under the circumstances, to discuss the ethical issues through the holding of courses, through the writing of books, through mass meetings in one which talks genetic screening and what it means and doesn't mean. And who does your DNA belong to, the police or to yourself? These are important issues.

Actually, at Cold Spring Harbor this week, we are having a meeting that I had to miss on DNA fingerprints with joint funding partly from legal bodies, or the police, and partly from the companies who are selling the fingerprint methods. And we have a federal appeals judge and we had some legal types to really come in to the question. And I think we're going to have to fairly soon have laws that at least define that if some of your blood is taken what rights someone else has to look at your DNA, not only to see whether you are actually the son of your parents, which could be great mischief if in the wrong hands.

It would be silly to go under the assumption that since we now can screen for things, we won't be able to. And I feel very strongly that no screening should be done without someone's consent. And that's not as easy as it seems because many people give consent knowing not what they are doing. And, in fact, should you be asked to give consent?

Should a parent have the right to look at their children's DNA? You could say, "Yes," but maybe not because you might uncover knowledge that you can't do anything with, but it will sure make people feel very apprehensive.

And certainly all the problems that one sees with Huntington's disease. Do we want to be screened? I think it really has to be individual choice. But people have to be educated as to these choices.

Though I'm aware of the difficulties of saying that we should have laws, I would personally feel that maybe the sooner we get them to finding who can look at your DNA the better. And it will reassure people that this data that we're going to be collecting won't be used against people, but it will be used for them.

So at the level of the University of California, I think lots of discussions -- even at the level of undergraduate courses -- are part of the medical curriculum just discussing what life is going to be like, say, 10 years down the line or 20 years down the line because right now everyone -- I think we can say that the opposition is not yet formed, and everyone is enthustiastic, Congress is enthusiastic. If you go and say, "I don't see any way to get a schizophrenia or manic depressive disease except for getting a high-resolution map," everyone will say, "Get that high-resolution genetic map, and we have to get the high-resolution genetic map."

If we found that schizophrenia was due to some transposable ailment or something which gets such high frequency, and we could really detect it, then you don't want someone -- you don't want the right before you hire someone to look at the person's DNA to spot the things that your group health insurance doesn't cover.

We raised the ethics -- it was fairly interesting -- the ethics question in front of HUGO, and I was really amazed at some of the clinical geneticists, ethics was terrible. I mean really it was sort of a pain because you had to deal with ethicists, and they're a phony lot.

But maybe we don't have to deal with ethicists. We have to become one ourselves. I think maybe rather than passing it off to other people, we need to become part of the discussions instead of thinking that we can hire ethicists who are going to solve the problems.

The people in the room, I don't think that you're going to have to spend much of your time talking about ethics or thinking about it. But I think that at my level I'd better think about it a lot because we certainly don't want to mislead Congress. We'd better be very frank with where we're going. And I think that what we're doing is completely correct.

But just the thought of all the complications of a national register of DNA, it sounds very good. Upon

birth you take a little DNA and you know who everyone is. So in the hands of the wrong director of the FBI you could -- at the simplest, you'll spot a lot of illegitimacy, which can be embarrassing. And it would be so foolproof. You can't argue your way out of it.

So I think that's a very immediate problem of making quite clear that you need a Court order to even look at the DNA and for a specific purpose and that it doesn't haunt you in the past.

Well, I've talked rather generally, and I think now you can hear the real science from Charlie.

DR. KREVANS: We are going to have our next presentation from Charlie Cantor. And I'll tell you a little bit about him in a minute. And then we're going to have a coffee break, and then a final presentation in the morning from Dr. Cook-Deegan. Following that we'll have three of our speakers available as a panel to discuss among themselves, answer questions, get comments from the audience on the whole range of issues which were brought up.

Our second speaker of the morning is Charles Cantor. And, obviously, James Watson has set a pattern. Dr. Cantor also has two jobs. He is the Director of the Human Genome Center of the Lawrence Berkeley Laboratory in Berkeley, but he's held onto his job at Columbia University as Professor of Genetics in case something goes wrong in Berkeley; although

apparently the time on that is running down, Charles, so you better make up your mind quickly.

Dr. Cantor had the good sense to be born also in New York City, and then had his education at Columbia, undergraduate, and then took his Ph.D. at the University of California in Berkeley. He has had a distinguished career as a geneticist, and brings to the problem -- and the exciting initiative of the human genome project/projects, et cetera -- an extraordinary record of accomplishment and a promise of an exciting colleague for those of us at the University of California.

Charles.

DR. CANTOR: In trying to figure out what to say to you this morning, I called Jim last week in New York and asked him what he was going to say, and he said that he didn't know. So that made it a bit of challenge. And I think, as you'll see, I've guessed at least partly right in figuring out how to follow him.

Jim has really painted a beautiful overview of the history of this project and stressed sort of really way into the future some of the ethical implications.

What I'm going to try to do is to fill in the middle ground and try to give you a feeling, in a little bit more detail, for the scope of the project. What exactly is the Humane Genome Project? What does it entail?

I will talk a bit about the -- not so much simply the DOE view, but sort of what I view as the complimentary roles of NIH and DOE, and maybe in passing sort of mention at least some of the California participants I expect to see as major players in the Genome Project, and then I want to look a little bit into the future.

I was told that one of the major purposes of this meeting was to be concerned with issues of technology development and technology transfer as and the possible relationships between the Genome Project and the biotech industry, and so I thought I'd have a few words to say about that.

Let me start by showing you the human genome. This is the human genome. These are human chromosomes stained with a dye that makes this beautiful banding pattern. People in the audience like Bob Sparkes and Barbara Hamkalo will instantly recognize that this individual is a male becaue there's one 'x' and one 'y' chromosome and that, unfortunately, this individual has Down's syndrome because there are three copies of chromosome 21.

This is the whole human genome, which is 3 billion base pairs of DNA. And each human chromosome is a single DNA molecule, linear, as far as we know, that runs end to end, and an average chromosome would have about 150 million base pairs of DNA, and as a single molecule, if extended, would be about two inches long.

If we actually look at such a chromosome by electron microscopy, rather than by a light microscopy, what you see is shown here. This is the protein scaffold that makes up the sort of core of the metaphase chromosome, covering almost the entire rest of the slide, probably not visible to anyone over the age of 40, but younger people can see little hints of single DNA fiber on the left.

I estimate that on this slide, you're looking at somewhere between 50 and 100 million base pairs of DNA. So these are really incredibly large molecules by anybody's standards. Our goal is to learn the structure of a molecule like this.

We can't work with these molecules today. The kinds of molecules that we're much more comfortable with are shown here. Jim actually mentioned bacteriophage lambda. This is a single molecule of lambda DNA. It's 50,000 base pairs long, give or take a factor of 2 or 3. This is the kind of material that we really can manipulate biologically today. This is the kind of stuff that we can clone, have as much of as we want to, understand the genetics of and so on.

Now the problem in trying to describe the Human Genome Project is that we're crossing many, many orders of magnitude and size and dealing with a number of different descriptions on the human genome. And I'll try to define

these for you here just pictorially to keep it in perspective.

The ultimate map of the human genome is the sequence. That specifies every one of the 3 billion base pairs of a typical referenced human genome. As Jim said, and I agree, we're a long way off from that, as I hope I'll convince you.

One of the things that we can do to get an idea of what's involved here is to make a sketch of that sequence which is, for example, a restriction map. We simply mark the places where particular enzymes can cut the DNA of a single chromosome. And this is convenient because simply measuring the size of the fragment generated by a particular enzyme tells us the distance between these bench marks. And this represents a sketch in the way that would represent a finished oil painting.

Now both of these are physical maps. They deal directly with the DNA. And they're both abstract maps. They really exist only in a computer data base.

The kind of map that is of great practical value that you need to actually do experiments, most types of experiments with the human genome, or any other genome, is to actually have that genome cloned and amplified as discreet pieces of DNA arranged in order. And here I named this as a cosmid map, but there are other possible ways one could clone or amplify DNA segments and have those in

order. And this is, I think, the area where the technology is in the greatest state of flux at the moment. And I'll return to that theme a few times.

But this is still a physical map because you're dealing directly with DNA molecules.

Now, the other type of map, which is really totally different, is a genetic map which for a human is a linkage map. You measure biotic recombination, that is you measure where the two inheritable markers -- Huntington's disease, blue eyes, or what have you -- where they are coinherited and passed from a parent to a child where they can separate. And we know, as Jim already mentioned, that the order of the markers on such a map is linear and must correspond to the same order as their gene on the physical map.

Unfortunately, the relative distances along these two maps in the human, and in other mammals, is variable. It is the meaning of this that prior to gening things in base pairs can vary by a factor of 100 on different regions of the human genetic map. And this is a very serious complication which will have to be dealt with.

A second problem is that recombination does not occur uniformally throughout the human genome. And so, in fact, the genetic map in many respects is a discontinuous map. It will have clusters of genes that are rather difficult to discriminate between genetically, followed by other clusters.

That's a major problem, and we need new technology to try to deal with that, and people like Dave Cox in the room are really in the process of developing that technology which could produce maps that are sort of almost hybrids between genetic maps and physical maps.

Now, one of the other problems in thinking about these different types of descriptions of the genome is that they occur on very different size scales, and that's summarized for you here. The Human Genome Project overall, the goal, is to get a fine genetic map, complete your physical map, acquire the genome as a set of samples in freezer, clone DNA, or amplify DNA and determine the complete sequence. Find all the genes.

There are somewhere between 50 and 100,000 genes in the human genome. Each one codes for a protein. Roughly half of these, as far as we know, are brain specific. And we probably have no other way to get those genes except by this type of global approach. It's rather difficult to do biochemical experiments on the human brain.

The goal is to find all the genes, and really the most important goal, and I think the goal that many of us have stressed since the onset of this project is in addition to getting all this information to develop the tools to use it for interesting biology in medicine.

And that really means two things. That means

to have improvements in technology that allow us to do rapid comparative studies and to have parallel records on a series of different organs because a human is not an experimental animal. If we have a human gene, it's very difficult to do direct experiments to test what it does. So we need parallel studies on the mouse, Drosophila, even though Congress is probably less interested in those organisms, or less excited anyway, than they are in the human. And that's the goal of the project. And this project covers almost 10 orders of magnitude in scale.

And I want to quickly give you a feel for that because I'd like to convince you that the types of thinking, the types of information we're likely to get, and the types of technology that we need to work across these different areas vary tremendously because we're covering such a wide size range.

So you'll find a model of the human genome as the earth as seen from outer space. That corresponds to the whole genome. On this scale, this is the genetic map. It's an aerial view of the city of Chicago, which I don't know if you can see these. So this is the genetic map. It's very interesting, but it's still fairly coarse.

The physical map is about a factor of 20 in higher resolution. And you're now looking at Soldier Field in part of Lake Michigan. This is the physical map, but you

can't sequence directly from such a scale. This could only correspond to a typical clone, at least that's ready for sequencing. We're now down to the scale of about 5,000 base pairs, and the actual sequence itself is represented by this, which is a 10x magnification of man's hand. So the problem with this project is that it covers a very large range of size scales.

Now, there are several things about this project that I think are unique. And this slide is really made to -- more or less for a different audience, but I think it's worth going through.

Unlike most other large -- first of all, I should say this is biology's one large science project. Unlike most large science projects, this one is basically closed-ended. It's feasible, and we know where the stopping point is. Unlike many large physics projects, I think that we can guarantee that the results will be interesting, no question about it. And that, unfortunately, has not been true for all large accelerators, right?

I think we could also guarantee that an important new technology will emerge as the project proceeds. And the reason for this is that I think all of us have accepted, and Jim has really underscored already, that we're going to do this project as an evolving technology model. That is that rather than try to slam into the project with today's

technology, even though it is feasible, we're going to basically develop new technology over half a decade, and maybe the one percent of the overall project with at least sequencing with today's technology. And then, hopefully, with technolgoy in order of magnitude better do about 10 percent of the sequencing project. And then finally, maybe in 10 years from now, start the completion of the project with technology that we can't even anticipate today.

The other thing that is unique about this compared to most other large science projects -- and this is really a challenge -- is that coordination of work at this first location is necessary. There is no particular reason to do all of this project in one place. And I think it would make it politically unsalable.

So we have to coordinate it. And I think scientists are, by and large, individuals and not terribly used to being coordinated. And we need to develop some effective mechanisms of doing this. And I think that's probably the biggest challenge for the project.

Let me give you a scorecard of where we are and where we are going. I think this is helpful in thinking about how far we have to go. This slide was made about six months ago. It's out of date.

The largest continuous DNA sequence now known

it's 250 kilobase. It's about the same size as the smallest chromosome that has more or less all the goodies we think of when we think of what a chromosome is biologically, which is roughly a quarter of a million base pairs.

The largest genome where we have a complete contiguous library and a fine restriction map, where we really have a good description, and probably the first genome that will be sequenced is the E. coli, which is .46 megabases. The largest autonomous piece of DNA, where we have any kind of complete map at all is just a little bit bigger than that. It's the largest S. pombe chromosomes, just under 6 million base pairs and, in fact, we have a complete map of the S. pombe genome. And I think Maynard also virtually also has a complete map. So mapping is feasible on this size scale.

But the human genome is 3 billion base pairs, and it's not so easily subdividable. The smallest human chromosome is 50 megabases. The largest human chromosome is a quarter of a billion base pairs and, unfortunately, these don't grow individually.

A number of groups in the national labs have developed methods of sorting individual human chromosomes to hand you this. And if you could get enough of such material, it wouldn't be that much more complicated than the yeast, but you can't get very much of it, unfortunately.

And people -- many people have developed methods basically of cloning human chromosomes in hybrid cell lines so that you have a single human chromosome living in a rat cell, or a hamster cell, but the problem is that you still have the rat or hamster's background so that these numbers don't look very much bigger than those numbers. It looks like almost a factor of 3, a stretch from yeast to human. But that's very misleading because this purified and this is contaminated with all of that. And working on human maps is much more difficult than working on yeast maps. And I'll come back and show you that.

Now, the human genome is three gigabases, in the physicist's language, but that number doesn't impress physicists. If you want a really large number, our goal in the long run is not to understand simply a single-reference human being, because that's basically almost useless, but we want to be able to compare human beings because that's how we're going to learn about interesting human diseases.

And if you realize that there are 4 billion humans on the earth, and we differ each of us from another, at the average of 3 million base locations, if you multiply 3 million by 4 billion, you get 12×10^{15}, 12 picabases. That's the size of the data base we really need to describe human diversity. That's ultimately the information we want to get. And that is really tough. That is a long, long

way off. That's a large data base from anybody's point of view at the moment. And to even dream about such data acquisition, we need to have vastly superior technology.

Now, let me spend a few minutes just showing you how we make maps to give you a feeling for what's involved, and why a human is much harder than bacteria.

There are basically two approaches. They are called in slang top down and bottom up. This is the divide and conquer. This is the constructionist approach.

In the top down approach you take a chromosome, you cut it into pieces, you figure out the order of the pieces. If you want more information, you take each piece, you subdivide it again, reorder it and so on. This is pretty feasible, but it works at relatively low resolution.

The bottom up approach, you take a principal randomly selected piece of DNA, you try to fingerprint them in one way or another, you find identifying marks on them, you look for correspondence between identifying marks, and you link up by finding such correspondences, adjacent clones. Such maps can be extremely detailed; at least in the beginning, they are very, very efficient to construct, but they are almost impossible to complete, probably impossible to complete for any higher organ. And, in fact, most of us, I think, now feel that a combination of these two approaches is going to be what is effective for human mapping.

Let me show you just what this looks like at two extreme levels to try to convince you that humans are not bacteria.

If you take the E. coli genome and you cut it into pieces, the genome is small enough and the pieces are big enough so, in fact, we actually can fractionate -- what you're looking at is separated DNA pieces ranging in size from 50,000 base pairs up to about 1 million base pairs -- we can separate every one of those pieces -- in this case the genome has been cut into just 22 fragments -- and making a map is simply determining the order of those pieces. The critical thing is that you can see all the pieces. You know how many there are. Putting them in order turns out to be relatively simple.

If you do the same experiment with the human genome, the human genome is cut in the same way into 3 to 5,000 pieces. And right now we don't have existing techniques that would allow us to subdivide the genome into a smaller number of discreet fragments. People like Dave Sigman are working on such techniques, but they're not functional today. And this is an area where great technology development is needed.

Unless we have it, we're faced with the fact that we have to do a jigsaw puzzle with 3 to 6,000 pieces, and we can't see the pieces. So that makes it into a great

challenge.

How do we do it? Well, we do it slowly and painfully at the moment. We start with the genetic map because that already exists. And today's genetic map provides anchor points for us that is based on the average about 10 million base pairs apart. The pieces that we can cut from the human genome are average about 1 million base pairs in size.

We can take each gene whose location we know on the genetic map and use it as a physical DNA probe to identify the corresponding large DNA piece on which that gene resolves. And the trouble -- when we do such an experiment in coli, what happens is we link up the physical map, and it's complete because the genes are close together.

In the human genome, the genes we know of today are far apart, and so these fragments are very distant from one another, and we have to somehow bridge the gap between them.

And this slide is meant for a fairly technical audience, and I'm not going to go through it in detail. Some of the methods are easy to understand, some of them are difficult, but they all work. But I'll show you just a simple one because I think it really does represent where the future is going.

These are large pieces of DNA that have been generated by cutting the genome with a particular enzyme. It turns

out that it's possible to get now quite efficiently small small pieces of DNA that overlap to large pieces of DNA, that is they contain on them the same cutting site that generated two adjacent large fragments. Such linking probes can be used as analytical tools to identify the two large DNA fragments and prove they're adjacent. This turns out to be a rigorous and relatively effective method of making maps.

I'll show you one example here for the human which I like particularly because it deals with molecules that are very, very large.

This is data from chromosome 21. These are two adjacent DNA fragments on chromosome 21. These are illuminated by the same DNA probe. One of these fragments is 2.5 million base pairs, the other one of them is about 2.7 million base pairs. These are a different set of cell lines, and the fragments are the same in most cell lines, but not all. And this single probe then actually makes a map of two fragments that cover 5 million base pairs over more than 10 percent of the whole chromosome. So this is a very easy way of making a very low resolution map.

Now, they probably need to think about over the next couple of years is how not just how to make a restriction map but how to actually make it in the order of a library, how to actually possess pieces of DNA in order in a form

where they amplify, where we have as much of them as we want. And a number of people have been thinking about effective strategies to do this.

And many of those strategies boil down to this type of an approach in which you try to make a yeast artificial chromosome library of DNA fragments and make a complimentary small library of ordinary size clones, but these are linking clones, and so they overlap two adjacent yeast clones, and then in principle such linking clones could be used efficiently to tell you that two larger clones are adjacent.

And if anybody could get this strategy to work in practice -- there's not at the moment -- then it would probably be a very efficient way to put a genome together. And a number of people are developing alternative ways to do this. Tony Carrano in the room has been developing effective fingerprinting methods for DNA fragments. Cassandra Smith and Sydney Brenner are both worried about efficient ways to actually do this step, which turns out to be very time consuming at the moment, and Bob Mortimer and Mike Esposito at Berkeley are looking at alternative ways to do some of this. So I think that this is an area where I agree with Jim that in the next year or two we're going to see explosive developments, and then we'll be able to make an evaluation.

But there's a problem, and it's this: In any mapping project, and it's also true for sequencing, it's

very easy to start but very hard to finish. And so I pose this question for you: How do you navigate if you have a map that somebody has erased 20 percent of the lines? It's very difficult, right?

What do we do about that? Well, we don't have any answers at the moment, but I'll show you what the problem is.

One of the major justifications for doing the Human Genome Project, for doing whole chromosome maps, whole genome maps, is an economy of scale. You have a chromosome, and you want to make a complete map, any piece you get from it, any clone you get from it to start with is as interesting as any other one. So you have a tremendous economy of scale. You don't have to do sorting.

The moment that you are forced -- let's say we have a map that's complete except we have a hole in this region -- the moment you are forced to focus on one region, unless you have an efficient way to go into this region and cut it out, you're back to facing the fact that you have to sift through large numbers of fragments, most of them are uninteresting, but you've done the work anyway, and so the economy of scale is lost. And you're back to problems which even for mapping are too expensive to be realistic.

So what we desperately need is some new technology that will allow us to focus, let's say, on a region of DNA

that's of particular interest.

Now, this is a realistic problem. This is a serious problem today, and it's because of the following: It turns out that a certain fraction of the human genome -- we don't actually know what that fraction is, but we know it's significant -- is not clonable. You cannot immortalize it in any vectors that we have today. It's not clear why. And so in any attempt to make a map, there are going to be segments that are simply not represented. Nobody has probes, nobody has clones.

So how do you get into such regions and work with them? Well, I think the seeds of the technology are there, but they need to be developed.

One of the things that I'm interested in trying to stimulate at Berkeley is the development of methods for working with single DNA molecules because, in principle, if you had this molecule, and that was the region you wanted, and you had the appropriate micromanipulation methods, you would just go in and cut out that piece. And then you could use the PCR technique -- and several of the developers of that technique are in the room -- to actually amplify this material even if you couldn't clone it, you could at least have some of it to work with. So that's one possible way that we're going to solve the problem of completing maps, but that is just a speculation at the moment, and it remains

to be seen.

Well, let me turn now from the science to the organization. And these are the major players and how I see them in the U.S; NIH and DOE. The role of private industry is unclear to me at the moment. I'd love to hear more about it. I'm sure that the instrument makers are heavily involved in the Human Genome Project, or will be. But it's not clear to me that a lot of the large pharmaceutical companies are interested in this. There is tremendous foreign interest.

It is my impression that thus far that interest is mostly intellectual. There has not been a large financial commitment on the part of most European countries today. Italy has probably been in the most advanced position in investing money except for the U.S.S.R., which apparently has in this fiscal year allocated 15 million roubles for the Human Genome Project. I don't know exactly what that corresponds to in dollars, but it's a lot. It's a significant fraction of the U.S. genome project. I don't think that the U.S.S.R. has a significant fraction of the U.S. talent to go along with that funding right now, but we shall see, and HUGO is going to attempt to coordinate all of this.

Let me spend just a couple of minutes giving you a sort of view of what DOE is doing. DOE loves organizational charts. And I just really want to point out to you two major things. DOE is supporting research today at both

universities and industry and at the national laboratories. Roughly, at least in the future, the support ratio between these two types of projects is estimated at about somewhere between 1-2 and 40-60. So the majority of the funding will go to the national laboratories, but not by much. A significant fraction will be in extra [inaudible.]

The work at both the national labs and universities is judged by peer review panels, a slightly different mechanism. All this is coordinated by the Office of Health and Energy Reseach which has an advisory committee somewhat affluent as far as I can tell to an NIH council. What is new and slightly different about the DOE organization today is something that needs, and I hope will be coordinated with NIH, is something called the Human Genome Steering Committee. That is actually a committee made up of some of the major contractors, and it is designed to actually try and coordinate the science at a very intense level. How it will function remains to be seen, but it is just beginning to function, and in a moment I'll show you what it consists of.

And it's going to mostly function by setting up specific task forces to try to deal with problems like data bases. And I suspect that most of these will wind up being coordinated intimately with NIH activities, Hughes activities, and foreign activities.

I want to stress what Jim has already said. There

really is only one genome project. There are many players. And the major challenge is how to coordinate all of these efforts.

Now, I thought I'd just show you what DOE is actually supporting today. The large support at the national labs is mostly at three, Los Alamos, Livermore and Berkeley. There are smaller levels of support at Oak Ridge and Brookhaven. There is support at a number of universities, one foreign grant in Yugoslavia, and all the rest in the U.S., and then a number of companies, a number of small businesses have some smaller levels of support. That is the current program.

DOE's budget for the Genome Project in FY '89. It was $18 million. It will probably be something around $28 million in fiscal 1990, we hope.

This is the Human Genome Steering Committee, which I chair for the first two years. It consists of representatives from -- George Bell of Los Alamos, Tony Carrano who is here from Livermore, myself. It has ex officio representatives from OHER, that's EOE, Elke Jordan who is Jim's assistant at NIH, and Diane Hinton from Hughes.

And its charge -- and it remains to be seen whether it can live up to this charge is to coordinate at least the DOE funded effort striving for the same degree of cooperation and efficiency that would occur if everything was in a single location; obviously, it isn't.

This committee actually exists. And to prove that, it has met once. This slide was kindly sent to me just last week by Tony Carrano, who I'm sure never thought I'd show it. This shows one of the great strengths of the national labs. I think it's important to keep in mind that they have always had great photography. That is the committee. It's just getting started. We're in our infancy, and obviously things look quite frightful.

Now, let me spend a couple of minutes first seriously and jocularly talking about the differences between NIH and DOE. I think that both agencies are bringing to this project different points of view, and that there really is in the true Hegelian sense, that the opposites will actually complement and make a stronger project overall.

I think one of the things to stress -- Jim didn't say it, so I will -- is NIH really represents the people who by and large are waiting with bated breath for some of this data are mostly supported by NIH. And I should say theirs is a retrospective view. And Jim's comments really suggest that NIH will change and move a lot closer to the DOE view.

But NIH, at least in the past, has distributed most of its resources in relatively small aliquots, individual research grants. There is no question that NIH is very strongly committed to study a number of different organisms,

and not just human. And there's no question that NIH is supporting extensive high resolution human mapping.

DOE is very comfortable with long-term large projects. It has mechanisms to coordinate efforts tightly, particularly those in the national labs. It's very strongly computing in engineering, which I think turn out to be major parts of the Genome Project in terms of technology that needs to be developed. And its emphasis at the moment is almost totally limited to human physical mapping and sequencing, not genetic mapping, not parallel studies in other animals.

Both agencies, I think, are strongly committed to supporting technology development, which is the future.

This is one way, I think to try to summarize what I've just said. DOE is strong in engineering and NIH is strong in biology. I think that no one would disagree with that.

This is another way, I think, to try to illustrate both the organizational problems and the differences. DOE is putting a large amount of its funds into interdisciplinary efforts at the national labs, or relatively large programs in other places. NIH is supporting a lot of terrific biology and smaller programs which may combine instrumentation and materials. But it was unlikely, at least in the past, to combine all of these at a single location.

And there are advantages to this. It allows for

a much higher probability of innovation. And there again when you're trying to get a boring job done, it helps to have a structure like this to do it.

Let me turn back to science for a second. I'm trying to decide one piece of science to highlight, and naturally I will choose something that's related to something that we're working on.

But this starts from, I think, one of the most interesting things which has emerged in the Human Genome Project in the past year is a discovery by Bob Moyzis Los Alamos of a DNA probe that detects the end of the chromosome of all higher organisms, that's both its strength and its weakness. It detects the ends of the chromosomes in a human cell, in a mouse cell, a hamster cell or what have you. Unfortunately, in the hamster it also detects centromeres, which makes it not so useful in the hamster.

Now, this is a great probe, but the trouble was that it looked at every organism. It was not human specific. And to actually take advantage of this for human studies, it was necessary to try to find a probe like this that would be human specific. And we were able to find such a probe in the last few months by using Bob Moyzis' approach.

And the scheme which we used is a very simple one. We gambled that if we took half a yeast artificial chromosome and demanded that it lie in a free-living form

that we could complement its missing head with the end of a human chromosome. If we had a bottle containing the ends of the human chromosomes then we could select these out of the bottle and clone them in yeast. Because we gambled correctly as it turns out, the human genome telomeres would be functional in yeast.

And so we've actually fished out clones, and this is one of them which have on their end Bob Moysiz's repeating sequence still in yeast, so we know where it came from.

Fortunately for us, the first clone had a human ALU repeat, the kind of thing that Carl Schmid studies here, so that we knew it had to be human. This sequence does not exist in yeast, and we were very lucky because this region turned out to be a human specific -- at least in most, if not all, human chromosomes -- the human specific DNA sequence, which we're beginning to characterize in some detail.

Why is this so important? Well, if you're going to make maps, one of the biggest problems that you get into is not knowing left from right. Since we can't see the molecules, as you start to map you don't know what direction you're going in. The only way that you would know what direction you were going is by starting at one end. If you start at one end, the maps are unequivocal, and in roughly

10 percent of the effort you can at least creep about 20 percent of the way down a chromosome. And so by having such clones we now have the ability to start from chromosome ends. And I think it's going to make things much simpler. And, of course, the ends are biologically interesting.

Now, in the last few minutes let me start with why we do this, what's coming out of it and where we are going in the future with this project.

The most interesting reason to have high resolution physical maps is to find genes. We have a physical map if we actually possess the DNA in this region. What the genetic map tells us is that Gene B is between Gene A and Gene C; if we're lucky, we can combine some information about where in between.

But, frankly, in those cases as you really bump the resolution up, it doesn't tell you where. It just tells you that it's in between.

But with the physical map, which after all is 10 or 100 times more detailed, you begin to know where to look for Gene B. The problem is that there's a lot of uncertainty in where this is. And as I mentioned earlier, techniques like those that Dave Cox is developing may start to cut down on that uncertainty considerably, but right now it's a real mess.

How do you know when you've found the gene you're

looking for? Well, that depends on what disease alleles you have that correspond to that gene. If you're lucky, there will be disease alleles that represent cataclysmic rearrangements of the DNA of the genome, translocations, deletions, insertions, something major.

This shows, for example, the case of Duchenne muscular dystrophy where roughly half of the known disease allele are large deletions; and, furthermore, the gene is gigantic. It's 2 million base pairs. So the gene is of the same order as the resolution of the genetic map. So in the case of Duchenne dystrophy once you're in the neighborhood of the gene, you're in the gene. And you know it because the disease alleles, or deletions, the material in the region is missing, and you know that you are there. That's the great fortunate case, and I fear it's not going to be all that common. It remains to be seen.

The more usual case, the case we're almost truly up against is Huntington's disease, cystic fibrosis, with many of the others that we're interested in is that the disease represents a single base change.

This presents the following problem: Genetics will locate if we are today to within 10 million base pairs. If we're lucky, eventually to within 1 million base pairs. Maybe if we're extremely lucky, and there's not too much linkage to [inaudible] within the region, to 10 percent of

that, but that's it. You just can't get any closer by anything that we know of today in genetics.

In this region, there will be 100 to 1,000 DNA polymorphism. Actually, I should mention before I go on that there is a sort of super genetics, which is potentially available from the recent work that Norman Arnheim has done. And if that proves to be generally applicable, it may allow us to go from 1 million base pairs to a tenth million base pair resolution providing that linkage to [inaudible] does not get in the way in that region. But there are real possibilities of technical advances here.

But even so, even if you shrink it to a tenth of a megabase, there will be 100 to 1,000 DNA polymorphisms in that region of the genome because we vary, 1 in every 1,000 bases. And of these 100 to 1,000 DNA polymorphisms, only one of them is responsible for the disease. The others are just noise.

To actually prove that we found the disease allele, the only sure way to do it is to sequence the region and look for the single base variation that correlates with the disease. And what you have to do is sequence lots of individuals. There will be many, many differences. One particular difference will show a 1-1 correspondence with a disease. You have the 'c' every time you have the disease, you don't every time you don't.

At the moment, we have two options. One -- I mean this means that you're doing millions or tens of millions of base pairs of sequencing every time you're trying to find a disease gene. And that either requires much faster, much cheaper sequencing technology. It's one of the major forces, I think, which is really driving this project. We can't do this kind of thing -- I mean with cystic fibrosis we're actually in this position. We have the gene -- not me personally -- people have the gene localized to about 800,000 base pairs, but nobody can conceive of doing 8 million base pairs of sequencing right now to try to prove that we've got the gene.

There is another approach -- and Rick Meyers in the room has been one of the innovators of it -- and that is that you may not have to sequence all of this region if you have specific ways of trying to look just at the bases which vary, you might be able to focus in on those one at a time, or in groups, and not have to actually repeat this kind of massive test. That remains to be seen.

Now, what's going to happen from the Human Genome Project? At the end of it we're going to have 3 billion base pairs of sequence. And I want to dwell on that a bit because it's a very large amount of information.

To put it into the scale for you, perhaps tangible, if you wrote it out in the same size type as the Manhattan phone

book, it turns out that it would fill 200 volumes of 1,000 pages each. That is what it actually takes to write out a single human sequence. And all of the available sequence today, the whole world supply of DNA sequence is one telephone book, or roughly equal to the yeast genome.

So we have to expand our data base by a factor of 200. And, clearly, in doing that, we have more than just a managerial task, we have an analytical task.

And I think that one of the areas where enormous advances need to be made -- and there are a number of people in the room who worry about such problems already --- is in fast multiple text matching.

What's going to happen progressively through the Human Genome Project is someone will sequence a region and want to know whether it's interesting, want to know whether somebody should immediately follow it up biologically or just store it into an archive and forget about it for the moment.

And the way we do that is we compare it with all of the known text. We compare it with all the known sequence that we understand and ask: Does it match? But the way we do that is tremendously crude. We tend to match text two at time without any regard for their meaning. And we need methods which will do that faster, will make multiple comparisons, and will consider the meaning of the text and

not just the raw text itself. And I think that as we develop methods to do this, which are not so simple, these methods will have a major impact on other problems where a large amount of text has to be analyzed rapidly, problems having nothing whatsoever to do with biology.

Now, let me talk for a moment about the meaning of the text. What do I mean by that? Well, raw sequences, just alphabets -- a, t, g, g, c, c -- but we know several things about it. We know that sequence is expressed in DNA structure. DNA is not a constant bordering helix. It bends, it rides, it does all kinds of things. People like Dick Dickerson in the room have been making their living on this for the last few years. And we now know that certain types of bending or arriving may be important in control of gene expression.

But nobody has really developed a methodology yet for combining our understanding of DNA structure and this text matching.

At the much more sophisticated level, we know that genes come from proteins that follow up in three dimensions. And three-dimensional structures are much more conserved than the text itself. And if we had better ways to guesstimate what those three-dimensional structures would look like -- that's a very, very difficult problem -- we'd be able to analyze this text much better.

Let me show you an example now -- a practical example -- of what I mean, and this is a very partisan example. It was made for a different audience. So don't take the message literally.

These are three texts. Text 1 and Text 2 differ or have 23 out of 35 units the same. Similarly, Text 2 and Text 3 have 25 out of 35 units the same. So in this way that we score text comparisons today, we would say these two are as similar as those two. And if you read the text it says, "I know NIH shouldn't be a lead genome agency." "I feel DOE should be the lead genome agency." "I feel DOE should study lead genome toxicity." Clearly, these two are related in a way that those two are not.

I really want to stress though that we can't do this today with DNA sequence information. We don't know how to do this kind of comparison. So we have a long way to go.

In the last two minutes, or three minutes, let me just turn to a little bit more of the technology. I believe that in the next five years these kinds of advances will be made in terms of technology; essentially, automated genetic mapping rather than the type of manual techniques that are being done today. Automative high-speed DNA sequencing is rapidly upon us. Automative physical mapping, computerized image scanning and analysis, protein structure prediction,

at least better than we do today, improvements in manipulating and cloning large DNA and multiplexy tricks that will enable us to analyze many samples in the same tube rather than one sample at a time by George Church's sequencing technology.

What's this going to do for us when we have this? Well, DOE I should say is supporting work in all these areas; and to be perfectly honest, NIH is also supporting work in all of these areas.

Data bases and clone maps. We're going to get two types of things from the Genome Project, and it's important to realize that we get them both. We're going to get technology, vast improvements in technology. Once we have it, it's applicable to all living species, not just to humans. So I think one of the major fallouts will be in agriculture, and that will also lead, I think, to substantial improvements in handling data and samples which probably will have a major impact in places like supermarkets, as well as the Genome Project itself.

The samples and the data are key because once we know the sequence of a single human genome, it allows us to have access instantly to any region of the human genome we want using the ability to amplify with PCR, any piece of sequence gives us the gene, gives us the material to do experiments with in humans. It gives us the equivalent genes in other animals if they're there.

So this is really an enormous resource, and it will form a basis of probably the next 100 years, or part of the next 100 years, of research and development.

To say it in a slightly different way, the instruments that are going to be developed as part of the Human Genome Project will be instantly usable in diagnostic testing. The methods development will be applicable to a host of plant and animal species of commercial importance. The data base design -- the major problem that we have to start with and coordinate -- I agree with Jim completely on this, it's where we have to begin.

Designing a data base and implementing it with the peculiar needs of the Genome Project requires solving a number of problems in computer science and informatics which are going to be broadly applicable to other fields. And, finally, to clone DNA samples will eventually lead to, I believe, improved pharmaceuticals because if we were to understand the gene involved in many serious diseases, we would be able to develop better therapy for them.

Let me close with one specific example which gets very much to the heart of the first problem that faces us because today we pretty much have a Tower of Babel. There are a whole bunch of different data bases. They don't talk to each other. They run on different hardware. They run on different software. Some of them -- God help us -- are

not documented. It's a mess.

We need to have what I think probably Tom Marr in the room first called a human genome work station. We need to have a package of software and hardware that sits on everybody's desk who is interested in this project, can address any data base regardless of what it looks like, contains all the tools you need to integrate and manipulate a variety of genome data. And since none of us are going to be monopolistic, this data base has to run on a wide variety of different hardware.

We would like it to allow the user to see the same screen wherever in the world he or she is. And also, because we're all individuals, we would like it to allow the user to be able to customize it.

And the typical problem that we get into -- it's a marvelous problem in the Huntington Disease Collaborative Research Project, which a couple of us in the room are involved in -- every lab uses different nomenclature. You use your own names for your own DNA samples. When you try to get together in a room and discuss the work that you're cooperating on, you're talking a different language.

And it's extremely difficult when you -- it's very easy if you have two DNA probes. You agree on a name. When you have 40 DNA probes, or 400 DNA probes, and you've been working with them for three or four years, it's very

difficult to get people to learn a new language.

So we need the types of data bases that will automatically translate from our local language, that we're not going to give up I fear, to a global language that can communicate between laboratories. We almost need computer translators in order to be able to work with them.

And there are two aspects to this. The first is that someone needs to take this very coarse description and firm it up. In terms of writing the specifications, what do we all want? And it looks like there is about to be created a joint DOE, NIH, Howard Hughes committee to try to do that, which would be a very great start, at least for the U.S. And I hope that group will have foreign representation so it will sort of encompass the world at the same time.

That's the easy part. It's just politics. People have to decide what they want.

Once we get past that part, then there's a much more difficult challenge -- and you may hear about that later from other people this afternoon -- but how do you actually implement it? Because to say "I want this" is easy. To say "Can we do this with current computer technology?" I think the answer is, "Yes," but it's not a trivial problem.

Finally, I want to close with almost the same way that Jim closed -- not quite -- the Human Genome Project, what do you do about the 90 percent which is garbage, the

90 percent of the text that we can't interpret today? Well, I love this quote from Sydney Brenner -- I think that's correct to attribute it to him -- and that is that this extra stuff in the human genome is not garbage, it's junk because garbage you throw out and junk you keep. So the human genome has kept all this stuff, so it must be there for a purpose.

First of all, there are some people interested in this garbage. The human genome has 500,000 copies of a short-repeated sequence. Carl Schmid in the room has been studying that for many, many years. And it contains interesting clues about human evolution. So it's not just junk.

I agree with Jim, today we don't know enough to be sure what's junk and what's not junk. It would be reckless to simply throw all this 90 percent aside.

Furthermore, even if we knew for sure, we don't really have easy methods right now to sort the junk from the good stuff. And so at today's cost, in terms of sequencing, I think nobody is going to sequence 500,000 repeated sequences that are all the same except for a few bases. There's simply no one that is going to take on that project voluntarily.

But as sequencing becomes much more automated, I think people will become more enthusiastic about prowling through the junk. And I think that when they do so, there

will be some interesting surprises.

I think with that I'm going to stop. I hope I have given you a fair overview of this project.

Thank you.

DR. KREVANS: To show you the effect of the advanced technology that is a part of anything we do about the Human Genome Project, unbelievably, we're right on time. We'll have a 15-minute break. We'll reconvene at 11:00 o'clock.

Recess 10:45 a.m. to 11:00 a.m.

DR. KREVANS: The final paper this morning is on the status of federal legislation and appropriations. And we are really very lucky that we have someone who actually knows something about it as opposed to someone who will criticize it. There are plenty of people available for criticisms of federal policies and appropriations available, but relatively few who know something about it.

Robert Cook-Deegan is currently a Senior Associate with the Office of Technology Assessment. As I remember, you also do something else.

DR. COOK-DEEGAN: I'll get into it.

DR. KREVANS: Yes, he's got two jobs. Everyone has two jobs. He was educated at Harvard College in his undergraduate school, and took his medical school degree

at the University of Colorado, was trained in the clinical specialty of pathology and then was a post-doctoral student with Erikson on the molecular biology of retroviruses, and did some interesting work in the early understanding of the biochemistry and the enzymes in the energy systems of the oncogenes.

Following that, his career has taken him more in the direction of an analyst and someone who understands and studies public policies, ethics, the interaction of those, working at the Kennedy Institute, as well as in the Office of Technology Assessment.

So it's a great privilege to introduce Dr. Robert Cook-Deegan.

DR. COOK-DEEGAN: Thank you.

Again, my name is Bob Cook-Deegan. And for the rest of this week I work at the Office of Technology Assessment, OTA.

The reason that I'm here really is principally because OTA did a study of federal policy relating to human genome initiatives that began about two years ago and culminated in the release of a report in April of this year.

We get into the act, typically at OTA, when there is a policy question relating to either science or technology that is clearly going to be a congressional concern. That usually means that it's going to last for a couple of years

and that it's likely to be fraught with some controversy with lots of different opinions about how some things should be done or how much of some things should be done, or questions of that nature.

It became clear in the middle of 1986 that the Human Genome Project meant many different things to many different people, and it was causing a lot of controversy since there were two federal agencies involved already at that point.

In what was by then already beginning to become a major topic of discussion in biology, it was immediately apparent that Congress was going to have to get involved because Congress is where budget decisions are made. And when there are two agencies in separate departments of the federal government, there are only two places where those can meet, one is in the Executive Department, which is essentially the President, and the other is in the Congress.

And in recent years, in fact, most science policy decisions have been made by a combination of the Office of Management and Budget, which is under the President, and the Appropriations Committees of the U.S. Congress.

The history of biomedical research, in fact, is very largely the history of Congress kind of forcing money down the throats of the executive agencies. That's my perspective as a congressional employee. That's one reason I'm here.

There's a second reason that I'm here that grew out of the OTA project, and that's that I'm currently doing a kind of -- it's not an informal -- but it's almost a verbal history that will soon be written down into a book. The Sloan Foundation has given me a grant to record the politics and the science of the early genesis of Human Genome Projects. So I've been doing that since June of this year half time.

And relating some of the questions that both Dr. Watson and Dr. Cantor have raised earlier this morning, sometime next week, I'll be starting a new job. It's for the Biomedical Ethics Advisory Committee and the Biomedical Ethics Board. This is a new congressional entity whose mandate is to look at ethical questions in biomedical research and in health care.

One of the topics that that committee has to deal with initially is the topic called "human genetic engineering," which is being interpreted by the committee to include gene therapy, use of genetic diagnostic tests -- both medically and nonmedically -- and also following what's happening in the Human Genome Project because a lot of people in Congress are interested in it but are concerned about the ethical implications of mapping the human genome.

What I would like to do -- I was going to show some slides, and I'm not going to do that now since I think

most of what I was going to say on the slides has already been said more eloquently by somebody else.

What I think is probably the most useful for me to do now is to talk about kind of the national context for human genome projects. And I won't take the scientific point of view or the technological point of view because I think those have already been articulated to some degree.

What I would like you to do is transport yourselves into being a member of the U.S. Congress or into being a high-level administrator in NIH or DOE or some other executive department that has something to say about genome projects, and think about: Why would you be interested in this project or set of projects?

It seems to me that there are principally five things that you'd be worried about on the one hand, or very supportive of on the other hand.

The first thing -- I think the first reflex of most people in Congress is to support the science. They don't fully understand the scientific implications, but they hear from the experts, the best experts, that having a map of the human genome is very important to understand the human genetic disease and, in fact, just to understand other diseases that are not necessarily genetic.

That has obvious implications. That's the main basis for the rather generous support of biomedical research

in the United States compared to any other country in the world. So there's a reflex to say "Yes, this is a very important thing to do." And I think the fact that both the DOE and NIH budgets have been relatively unscathed in the last couple of years in times of great fiscal austerity is testimony to the fact that people in Congress and in the executive agencies think this is very important.

There's a second line of support for anything relating to biotechnology. And that is the economic implications of human genome projects. The thinking there seems to be that work on human genetics is related to biotechnology. And biotechnology at least eventually is going to be related to jobs and wealth. And if you're a member of Congress you care about your district, you care about what you're paid to do is to represent the people in your district. And what they care about most is how they're doing. And jobs and wealth are the things that are on the top of that list. The linkage is not very direct, but that is something that is in the front of most people's minds as they're making decisions about human genome projects at the federal level.

There's a third issue, and that's national pride. And this is, in fact, driven some of the decisionmaking in Washington; that is, there are lots of arguments about if we don't do it then the Japanese are going to beat us or the Europeans are going to beat us. And those arguments

actually play relatively well for short periods of time. Generally, you eventually get down to making judgments about the technical feasibility and the actual pragmatic applications of something you're spending money on. But you don't have to necessarily do that initially.

The fourth concern relates to the other three. And this is that obviously the very No. 1 dominant issue in Washington for the next year is going to be the deficit. And now that the election is over -- I shouldn't say election, the multiple elections -- members of Congress also worry about things on a two-year cycle. The President worries about things on a four-year cycle. And all the cycles have stopped simultaneously now. So what you're going to see in the next year is that people will worry less about getting elected and more about managing the government. And that is going to thrust the deficit problem right to the top of the list.

The human genome initiative is somewhat unfortunate in it's coming out of its embryogenesis and really into its fetal stage right now, I think. And it's coming to that stage at a time when there are tremendous pressures to restrain federal spending. And I, frankly, don't know what's going to happen. But that is a major concern for anybody in Congress.

The fifth area is the social and ethical implications of human genetics research. Dr. Watson has already mentioned

a lot of the things that people in Congress are worried about. They're worried about people being able to get insurance, the use of genetic information for insurance purposes for hiring and firing decisions for eligibility for federal programs or non-federal programs, the criminal investigation uses -- that is forensic applications. They also worry about the costs of the medical uses of a new technology. And perhaps more important than all of those is the spotty history of the use of genetics in politics. And here I'm referring to eugenics.

Most people in the audience think that the issue of eugenics was something that happened early in the 20th century and went away. That is, in fact, not the case. There are people in prison in Malaysia right now that because of their beliefs relating eugenics. In fact, the Prime Minister of Malaysia right now is a -- what would at the turn of the century have been classified as a classical eugenecist; that is, he bases theories of race and justifies those in kind of a pseudo-scientific jargon.

But that history does overlay human genetics, and it's something that people in Congress worry about. And I think that what you're going to see in the next few years is that all these five major interests are going to be thrown into the same pot. And in the usual chaotic American mode something will come out at the other end. And I'm, frankly,

not sure what that's going to be.

My guess is that the projects -- well, let me finish one thing on the social and ethical implications before I go on because I don't want to leave it on a negative note.

Frankly, most of those social and ethical implications have nothing to do with the Human Genome Project. The Genome Project is the creation of information. And almost all of those issue relate to the use of that information and are, in fact, things that are part of the normal social fabric, and they are things that legislatures are well prepared to deal with. Who has access to this or that thing? Who owns this or that thing? Those are legislative questions. Those are legal questions. The system of government is actually pretty well prepared to deal with questions of that sort once they are understood.

The creation of the information is really on separate track. And the only question that has been raised to date is: If this information is going to be abused, why would we create it in the first place? But, generally, that sort of argument doesn't go very far because most people in policy-making positions understand the distinction between generation of information and its use. That is, again, a distinction between discovery and application.

The final thing I wanted to talk about -- although it has nothing to do with the legislative status these days -- is

the management prospects for genome projects. You've heard about the DOE and NIH efforts to organize the projects. I think, actually, if you compare development of the genome projects in the United States to development of any other biomedical research program. It's been remarkably efficient, frankly.

Two years after the beginning of discussions about how much controversy there was about genome projects, NIH and DOE seemed to be living somewhat compatibly on the same planet with the same budget mechanism and, in fact, have a piece of paper that you can point to and it say, "This is how we're going to cooperate." And that has resulted in things happening in the real world.

If you follow the history of a project within NIH, that's not always the case. And that's within a single executive agency.

There are, nonetheless, going to be lots of complicating factors. And Dr. Watson and Dr. Cantor referred to this. So far France, Italy and the U.S.S.R. have got pieces of their federal budget, their national budget, devoted to genome work. The Japanese government is now kind of in the stage of trying to formulate its policy.

There's a lot of talk from people in the four major research agencies in Japan about who's going to take the lead. It's kind of the debate that was going on here two

years ago. It's going on right now in Japan except that in the United States it's done on the pages of Science and Nature magazine. In Japan we'll probably only know about it when the results are pretty well established.

One overarching comment about the level of effort, however, is in order. The U.S. effort right now is at $50 million this year, and it sounds like it will be somewhere around $80 million next year, and maybe even higher than that. That totally dwarfs all the other efforts put together from all the other nations. It's probably more than double that all other nations are going to put together in the next years. Perhaps it will change over time.

But one thing about the international efforts is they are much more difficult to coordinate because there is no point where all decisions converge as there is in one government. And HUGO is, I think, the only hope for fulfilling that role. But I'm not sure that unless it's fiscally healthy and has some sort of agreement from the various governments to cooperate with its planning efforts, I'm not sure how that's going to work out. This is totally unprecedented in biology that there would be a concerted well-planned effort involving more than one nation. There are examples in other sciences, but the history of their success is somewhat spotty.

A final comment on the international aspects of the genome project. I think the key issue there will data

sharing. And I think there is widespread agreement among the scientists of all nations that data should be shared. There are problems, however, in implementing that very traditional scientific value. And there is a tendency to blur distinctions. It's very clear that scientists left to their own devices would share journal articles and the sort of normal pieces of scientific communication. It's less clear that they would share data base structures, but that has been done with some success in the cases of GenBank and, in fact, in the RFLP mapping efforts, the efforts of CEPH have been really pivotal in keeping the groups doing RFLP mapping unified at least to some extent.

So there is some hope for cooperation at the data level. It's equally clear, however, that governments to the degree that they see investment in genome projects as an economic investment see this as a way of creating new instruments, new technologies, new ways of making products. And that's going to be an area where economic nationalism, I think, will be the norm.

And the game will be to try to separate the data gathering and data sharing from the commercial aspects of the human genome projects because it's clear that on one hand there's going to be competition; on the other hand, there's an agreement that there should be cooperation.

But the decision-making apparatus in the United

States are very different from what it is in most other nations. The people in this room today are -- as I look around -- principally scientists. And, in fact, scientists have a great deal of power over the decisions over spending money for science in the United States. They have a much greater degree of control over spending in the United States than in most other nations. And particularly in Japan, most of the role of the government is in influencing private corporations spend their money. It's an indirect role rather than a direct role. And it makes the process of formulating policy in Japan much more difficult to explain some of the difficulty of the Japanese government being able to say, "Yes, we'll pay for this or that thing." They don't do that in any area.

And, in fact, the research budget in the United States is about 50/50 private/public. In Japan it's much more highly private research money in the biomedical area. So it's going to be a difficult impedance match among the various nations, particularly between the U.S. and Japan, because there's so much room for trade fiction.

I'll end there and just open it up for discussion.

Thank you.

DR. KREVANS: Thank you very much. I'd like to ask Dr. Cook-Deegan and Dr. Cantor and Dr. Watson to please come to the table. We have now a free time for comments, discussion, questions for our panelists. And they

can question one another or you all can comment and question them. And I'd like to just start out by calling on people for questions and comments.

Could you please identify yourself so that you can be recorded, and 10 years from now you can point out what a mistake you've made.

DR. SHANKAR: Robert Shankar, University of Southern California.

I haven't heard any participation from the United Nations such as UNESCO or the World Health Organization. Is anybody --

DR. WATSON: I would hope that we would have none in the sense that they're large bureaucracies, and I don't think they'd bring any expertise. And we would spend even more time in meetings if we had to bring them in. And I think that the reason for forming HUGO is to keep an organization run by the scientists instead of, essentially, civil servants.

DR. KREVANS: Do other analysts want to comment on that?

DR. CANTOR: UNESCO has been sending observers to some Human Genome meetings recently, but they've been keeping a low profile.

DR. COOK-DEEGAN: They're very interested in following the projects. They're particularly interested in if there are data generated in the Human Genome Project promoting

Third World applications of that information; for example, sickle cell disease, or something like that, the diseases that are highly prevalent in Third World countries. They want to make sure that their data are rapidly made applicable for the needs of Third World countries.

And they have been quite interested -- in fact, at the Valencia meeting about a month and a half ago, there were three representatives there. And they were extremely enthusiastic; more so, I think, than the scientists were about having them involved.

DR. KREVANS: Next.

DR. NEUFELD: Elizabeth Neufeld, UCLA.

Dr. Watson, what proportion of resources does the NIH plan to put into genomes of other organisms versus the human genome?

DR. WATSON: I don't think we've made any decision. I think --

DR. NEUFELD: Would it be minor?

DR. WATSON: No, I think it would be major. I mean in any sense of the word, we want to get the coli sequence out as fast as possible, and then you want yeast, and then you want to really see that we get the Drosophila sequence, I mean certainly a map of it.

And I think one thing is that we hope that bodies will appear which say, "We want to do something." That would

go a long ways away from the human. We won't end up being sequenced in Europe because they actually put their act together. Or will some group from the United States say, "We really want to find out what a plant is"?

So I would hope that there would be bodies in the United States, or groups, which will have rather heroic objectives.

DR. CANTOR: Let me add to that because I sit on the Genome Council, and so I have seen what has thus far come through and been funded with genome, I mean even before Jim was on board. And I think, roughly half of NIH's genome budget thus far has been spent on organisms other than human. Some direct attempts to explore those organisms, some using them for model systems to develop technology. I don't know how that pertains to the future. But that's what has happened thus far, which is very reasonable.

DR. WATSON: Maybe, I think, we sort of want to put a lot of money in the mouse. But you're going to have to find people who really can do it well. I think there's more a problem of finding people who can do it than the desire to do it. And I think there are going to be a lot of people saying, "We want to do it," and then we're saying, "Well, we're not interested."

DR. KREVANS: Is the extensive, fairly extensive, data base on mouse genetics a reason you say we should use the mouse in terms of one of the mammals?

DR. WATSON: Well, you want to compare the mouse sequence just really to identify genes in nonsense regions.

DR. CANTOR: The mouse is really, I think, especially useful for comparison because as expensive as transgenic experiments are on mice, they are more expensive than probably almost any other animal, and they will be more expensive in any other animal you can think of.

And the problem that you're going to be faced with in so many cases is: You have a gene, whether you've originally gotten it from human or a mouse almost doesn't matter, you want to know what it does. The first experiment you want to do is create a nullmutant in most cases, and knock it out. You can do that in the mouse today inefficiently; three years from now probably efficiently. You can't do it in a human ever probably, and you can't do it in larger organisms because no one will pay the bill.

DR. WATSON: The mouse world now -- and maybe I shouldn't say this -- is rather mouselike. And we really don't expect to get much. So we have to develop leaders of the mouse world who really think big and have the same ambition as the human people. And it's unfortunate that the center of the mouse world has to be in Bar Harbor which is so far removed from everyone. They have great difficulty in recruiting such molecular types. But that would have been the natural center for it. But it doesn't look like,

in fact, it would be the place where --

DR. KREVANS: Those of us who summer in Maine think that Bar Harbor is much more attractive than Cold Spring Harbor.

DR. WATSON: I think it is in the summer. But I think that we're going to find a way to encourage major groups working on it.

Yes.

DR. LAKE: Jim Lake, UCLA.

I wanted to ask -- a lot of people seem to think that one of the most informative organisms for comparison to put us in the proper background would be the chimpanzee. Is this the sort of thing that you would perceive as following?

DR. WATSON: No.

DR. LAKE: You wouldn't. You think it might have some political ramifications? I think scientifically it must be central --

DR. WATSON: We really can't do the experiments that you can do on the mouse. So it's not an experimental organism.

DR. LAKE: But in terms of comparative questions?

DR. WATSON: When someone gets all the costs down, then we can look at the chimp. But I think right now -- I personally think -- I wouldn't think we'd develop a big group working on chimp DNA.

DR. CANTOR: I've heard that the chimp is also a little too close evolutionary to the human. And that if you could ever afford to do the experiment, some more distant related monkey might be a better choice. I mean that's probably debatable.

DR. WATSON: The way things are going it would be harder to do experiments on the chimps than on humans.

DR. KREVANS: Certainly more expensive.

DR. WATSON: So I really shudder at the thoughts of the problems.

MR. HURST: Steve Hurst, UCSF.

I'm a patent counsel with UCSF, and one of the areas that I have noticed a deafening silence on is the issue of patent rights related from proprietary technologies developed by the participants in the Human Genome Project.

I see it as being anything from a minor nuisance to a major problem within the project in terms of how those rights are going to be coordinated. And I think it's a question that industry will at least have some interest in.

I wondered if the three of you could comment the coordination efforts in that area, any thinking other than the proposed solution of putting all patent attorneys on a boat and sending them off to China? What are the solutions to the potential problems associated with having to fuse proprietary rights?

DR. KREVANS: I don't want to discard that solution.

Do you want to take that on, gentlemen?

DR. COOK-DEEGAN: We thought long and hard in connection with OTA project about what we were going to say about patent rights. And, frankly, the federal policies on it are fairly straight forward.

The Institution of Human Research get the rights, whatever the rights are. The problem as I see it is that we don't know what those rights are because there's no case law in the area.

For international stuff, I don't think it's a problem, frankly. It's very clear who gets a patent. There are serious problems in harmonization of patent criteria among nations, and procedures and things like that.

But in terms of a federal effort involving DOE and NIH or NSF, it's very clear of how they're supposed to distribute the patents. And I don't see that as being a very -- from the federal perspective -- as being all that important. It's going to matter a lot to the people who are doing the research, but then it's going to be up to them to negotiate who owns the patent. The Institution and the people doing the research would make those decisions now. That's federal policy.

DR. KREVANS: Charles, do you want to comment on this?

DR. CANTOR: I just want to second what Bob said about harmonization because already a significant fraction of the effort in the Human Genome Project are international. And the differences in the patent laws and constraints about disclosure in different nations is really cumbersome. I mean that's a problem that transcends this enormously, but it's a mess. And I would hope that patent attorneys can one day straighten this out.

DR. KREVANS: All the way in the back, please.

UNIDENTIFIED SPEAKER: There was no mention of the Department of Defense in the funding role of this. Are there implications to DOD? And if so -- [Rest of the question was inaudible.]

DR. WATSON: I don't really think that their help is needed. You know, at a different level, I think they're very interested in the DNA fingerprint for body identification, and you can get that sort of a question. Do they want to DNA fingerprint everyone who enters the armed services. And I don't really feel qualified to make any comment on that.

DR. KREVANS: Charles.

DR. CANTOR: Well, they're also interested in automatic pattern analysis. And that technology is relevant to the Genome Project. There's no question about it. Patterns anywhere from DNA sequences to automatic analysis of photographic images. But, again, I agree with Jim; I don't think that

they're needed in this project.

DR. KREVANS: I just want to pursue that. Hasn't a lot of the advance in pattern analysis come out of DOD research on how to analyze patterns from satellites; the other direction, if you will?

DR. CANTOR: As far as I know, yes.

DR. KREVANS: Do you want to comment on this, Bob?

DR. COOK-DEEGAN: No, that's absolutely right. The DOD has got some interest in this. They've got a few little bits and pieces here and there. They've got the world's best treasure house of odd pieces of the human body at the Armed Forces Institute of Pathology. It's a wonderful resource that probably will eventually be quite useful to their genetics and things like that.

And I think that at the hardware end, their artificial intelligence is in large part supported by DOE. It's highly relevant to these projects. But I think at that level, I think it will probably trickle down from DOD. Frankly, they would probably be a little worried about being perceived as getting into this because everybody would immediately think that this was pretty nefarious purposes.

DR. KREVANS: Yes, right here.

DR. LONGERBEAM: Gordon Longerbeam, Lawrence Livermore National Lab.

I have a question about technical data rights excluding

patents, perhaps from Dr. Cook-Deegan, but the others may respond.

Do you see any progress at the federal level on differentiating between technical data which has economic value and technical data which is principally scientific value which should be open to share? Or is that still a very sticky question?

DR. COOK-DEEGAN: That's a mess. And it's all tied up also with the question: Who owns the data and all that? And that's a total mess right now, and I'd just as soon not talk about it too much. We don't know very much about it, and it's undoubtedly going to change in the next two or three years.

Access to technical data and who owns it and all that is very unclear.

DR. KREVANS: But this, I think, is the question that Mr. Hurst was trying to get at. As these things evolve in cooperative studies where we talk about ways in which to make data accessible, then when something comes from it, who owns it?

DR. COOK-DEEGAN: If that was the point of your earlier question, I missed the point, and I apologize.

MR. HURST: If I could maybe elaborate just a little bit on the issue.

DR. COOK-DEEGAN: Sure.

MR. HURST: I think you'll see within the industry that there is a lot of -- I think the relevant point of the whole project is that at some point the Genome Project will translate itself into a technology transfer effort that will ultimately benefit people not just of this country but the world.

Those technologies are expensive to develop and require some incentive. And usually the industry will perceive incentive as being exclusivity, or maybe even free competition where someone can't come in and close them down, all of which is relevant.

And I think that if you find that in a project as profuse as this, those patent rights rest in 40 or 50 or 60 different places, I think you're going to at some point, maybe not in the basic research, but certainly where the tech transfer efforts take place, I think you're going to find a real bottleneck that, hopefully, will straighten itself out. It will eventually straighten itself out in the sense that all the patents expire at some point in time.

But I wonder, perhaps, should we wait 30 years after the technology is developed for all the patents to be out of the way before the industry can really take off? And I think if you don't have a coordinated effort within a project of this scope, can look at exactly what you've cited, which are the federal laws on ownership and the fact

that each institution can elect ownership, and in a project of this scope I'm not sure that sounds reasonable.

And I think that's what I was getting at because you're going to find an impedance in progress at some point along the way by virtue of proprietary rights in data, techniques, products.

And I think the law is clear that gene sequences, protein probes, et cetera under the U.S. patent law are clearly patentable right now. This is a real issue.

DR. COOK-DEEGAN: Yes, I think just to make sure that I understand where you're going with this, I think the real issue is not: What's patentable and what's not patentable? Although, that's still in question.

It's also clear that if you've got something that's patentable, it's clear who owns the patent. What is unclear is what you do with information that goes into the process of documenting that you have an invention.

And I think that patent laws as they're currently construed do have -- it is a serious problem for the project in the sense that people will be reluctant to disclose sequence information; for example, that is relevant to a patent they found until they file it. And that could lead to a delay.

But if you aggregate it over hundreds of labs where you're trying to pool data, it does matter a lot. It delays things by years if you add it all up. That is a serious

problem.

There are things that the federal agencies could do to encourage data sharing if they just had explicit policies saying that this sort of information needs to be shared as a condition of accepting this grant. But there's always going to have to be a clause in there. And there's going to be inevitable uncertainty because the agencies can't cross that line if it gets in people's way when they're filing patent applications.

It's against the law for them to put any impositions on recipients of grants if it's going to interfere with their intellectual property rights.

So there's a gray area there, and it's just going to have to be thought about a lot by the federal agencies. And I think it's going to be muddled through, frankly.

DR. SALSER: Winston Salser, UCLA.

Regarding the earlier question about the interaction of the Department of Defense, it seems to me there's more likely to be an interaction and a problem with the intelligence community because one of the things that they do is analyze huge amounts of communication to sift through, trying to sift through it.

And they must have contemplated for at least 10 years the various kinds of data analyses that will also be needed here. And, presumably, they could tell us a lot about

how to do it. I don't think they will.

But more problematically, if we look at the basic researchers on breaking codes, they really interfered a lot by classifying a lot things on code breaking. And we may have them really interfering when someone gets a powerful outgrowth in decifering DNA and picking up things. They may feel that that's very threatening to them, and that would be unfortunate.

DR. COOK-DEEGAN: The Cal Tech chip that is being used with that style finding is based on one of the chips that was developed -- so there's a big area of overlap. And I think the biologists only got access to it after it was no longer the first generation technology.

UNIDENTIFIED SPEAKER: I understand that the [inaudible] of automated instrument in DNA sequencing is important for this program.

In Japan, companies where [inaudible] instruments are developing the instruments [inaudible] is support from government.

Is there any federal support for the development of instrumentation in this sequencing effort?

DR. CANTOR: Really that's one of the things that has changed as a result of the Human Genome Project is the fraction of the money at both NIH and DOE is now being spent on instrumentation development. That's what was difficult

before the Genome Project. It was relatively hard for those kinds of projects to get funded in a conventional way.

UNIDENTIFIED SPEAKER: The money is spent in [inaudible] or with companies?

DR. CANTOR: Some of the funds are used for SBIR or direct grants. We have a lot of the biotech companies are quite successful in competing for standard research grants of NIH or DOE. And so a reasonable fraction of the funds have gone along those directions. I mean I can't name names, but I know of at least one major company which has totally new ideas for sequencing and is seeking federal support for it, and I imagine will get it, based on quality.

DR. COLE: Belle, Cole, University of California, President's Office.

Dr. Watson, you mentioned that -- you discussed the ethical issues that confront the project, and mentioned the role of this advisory committee. And also you mentioned that Congress will be hearing from many different groups about the project. You also mentioned there may be some interest groups out there that will develop and that might be adverse to the project.

What are some of the things that NIH is thinking about to deal with these issues, ways of educating the public, ways of just being prepared? I just wondered if --

DR. WATSON: I think we really don't have a program, but I would think that within six months we will announce one. I think we have to -- we haven't had a meeting of our advisory committee. And when that occurs early in the year, I would hope that we would form some subcommittee to deal with these matters and to deal with things like with the Human Genome Office at NIH to issue proposals for grants in this area, but at what levels I can't say.

But I would think there should be a level of several million dollars, a fairly sizable sum.

DR. WHITELEY: Norman Whiteley, Applied Biosystems, Inc.

In response to: Does the federal government support instrument development? They financed the purchase of a lot of the equipment. Something like 10 or 12 or 14 percent of that money goes back into research. So indirectly they support a great deal of research.

I'd like someone to explain the role of HUGO.

DR. KREVANS: Charles, do you want to handle that one?

DR. CANTOR: The role of HUGO will depend very much on how successful HUGO is in fund raising. At one extreme, if it's very unsuccessful at raising funds, it's likely to be in discussion forum and very little else.

If it's quite successful at raising funds, it sees itself modelled after an organization called EMBO, which

is the European Molecular Biology Organization, which has a substantial budget which it gets from participating nations. And in addition to being very active in training various post-doctorals of workshops and so on runs a gel. And most important, it runs one large international laboratory called the European Molecular Biology Laboratory.

At least some of the major participants in HUGO at the moment are very enthusiastic about trying to have the latest stages of the project -- a discreet number of international laboratories -- that would do the real predominant efforts in a way that will allow a lot of foreign nations.

Whether that fantasy ever really becomes a reality will depend on whether the governments are willing to give substantial amounts of money to a truly international organization.

Other people in the room may know more about EMBL than I do. It's my impression that EMBL is a very delicate year-by-year political negotiation of participating nations to try to keep a budget going. It's very, very tricky.

And what I'm worried about with HUGO is that since it's the whole world, and not just a small European nation, it's going to be even more complicated.

So I've given you two extremes. I think the reality is likely to be somewhere in the middle.

DR. WATSON: Right now HUGO has $75,000.

DR. KREVANS: It's certainly not going to do a

lot of harm with that, Jim.

DR. WATSON: I would just like to say that I think it will need roughly $1 million a year to really have a real secretary and to be able to hire someone who really has this function of really trying to integrate the activities in different parts of the world.

I think if Charlie and I try and get bogged down in bilateral discussions with all the different countries involved, it would be a mess. And we would like to in the United States interact with HUGO and not have to, but that means a secretary. And I think that's about $1 million dollars a year, which means that the United States itself would have to make a contribution to it, that is our government.

And I would simply be in favor of -- if the means can be found -- of NIH making a contribution to HUGO when we really know what HUGO is going to be.

The thought, I think, was that for a year or two you should get foundations to give money to put together a secretary so that the governments would know what they're buying.

DR. KREVAN: One last question, please.

MR. HUNKAPILLAR: Tim Hunkapillar of Cal Tech.

And I was going to real quick follow up on what Bob said about the TRW. This is an idea. They're actually so far -- so far there haven't been any real difficulty.

In fact, they come looking for people to put the technology out into the world. And NSF doesn't seem to care one way or another, not just NSF but those people, don't seem to care a whole lot.

Where is the NSF in this? NSF obviously is a major contributor to funding for instrumentation and large data bases. I know they do a lot of data bases all over the world.

And why do we never hear the NSF in these discussions?

DR. KREVANS: Gentlemen? That's the question that I asked as you were gathering together.

DR. COOK-DEEGAN: NSF in terms of budget has life sciences budget that's roughly in the same ballpark as DOE's. And they have probably the largest -- it may not be as of now -- but it used to be the largest budget for viewing biology instrumentation development.

So, in fact, they are de facto in the game. And they've had a big role as you know. I mean they were the only federal agency that ever supported the DNA sequencing. So they do have a big role here.

I think that they've been out of it just because DOE and NIH both identified genome projects. And NSF made a conscious decision not to do that same thing. My guess is that eventually they're going to have to consolidate their efforts so that they can have a match to what's going on, and that's DOE and NIH.

So I think it's a political answer; that is that it made a conscious decision not to have that sort of an approach to genome projects.

DR. KREVANS: I'd like to close this with one question and an advertisment.

The National Academy Press has just released a committee report on which both Charles Cantor and James Watson were members of the committee, and which was chaired by a colleague of mine in San Francisco, Bruce Alberts, and it's on mapping and sequencing the human genome. And it's available from the National Academy Press. And the University of California, San Francisco gets a small royalty on copies sold.

In this book the question is raised: Why sequence the entire human genome? And then it says that we hope everyone agrees that it's a wonderful idea. And it says that there were only three major things that sort of stand in the way. And the third one is -- and I'll read it:

> "Even if the project is worthwhile, the intensive effort required will divert funds from other research aimed at understanding the structure and function of genes in all organisms and, therefore, there will be a net loss rather than a net gain of important biological information."

Now, the narrative goes on and destroys that argument,

but I would like to hear particularly James Watson and Charles Cantor say something on that in closing this morning's session.

DR. WATSON: Well, I think the sum of money, if you add together the NIH support of biomedical and NSF and DOE and Agriculture Department, it would only be about two-and-a-half percent of the total biology -- I think in terms which are buying as [inaudible] it would be a bargain if we can actually deliver.

The other is, I think, the breadth of the program that we hope to have. At least at NIH, we would be doing things which a lot of people would have wanted to support out of our lungs of program projects even if this program didn't exist.

So I think our -- as I said before -- our aim is to see that when we pass out the money, outsiders can't say we're passing out money to second-class people who -- that are just supported because we have a lump sum of money which is of a certain size, and we can't fill it with good people.

So I think we really have to -- if we believe in the project -- really work hard to encourage good people to come into.

And if I could say one thing, I think right now the program is structured, for the most part, at NIH in terms of almost on a one-type grants. I'm personally a real believer

in rather large program project. And I would hope that groups at universities like the University of California and otherwise would come together and put together sort of genome centers which we could fund because I think we're going to need a variety of diverse talents, and by forming centers, you'll be able to, I think, get more done than individually.

DR. KREVANS: Charles, do you have anything to add?

DR. CANTOR: I think Jim has really said it all. I would just add as a postscript that that criticism was leveled at and was valid at the original concept of the project. That current technology, the moment you broaden the project to the model which exists today, the criticism is really -- it just doesn't apply.

DR. KREVANS: What I was hoping our panelists would say is that it's better to spend the money on this than particle physics, but they're very ecumenical.

I'd like to thank our panelists very much and turn the program back to Paul Boyer.

DR. BOYER: We appreciate very much the fine panel we've had this morning. And I would like to continue to mention what our program will be. Let me just quickly correct some mistakes. It won't take 10 years. As you know, I've tried to give Julius two jobs, both at San Diego and here. That was a mistake on the thing.

I need to take 10 years off my own time since a Ph.D. I want to say that in terms of biochemistry, the area of proteins and enzyme molecules would have been far ahead if Jim Watson would have chosen to be in that area. I think it would have been a disaster for genetics if I would have chosen to try to do Jim's research.

I want to also say something about how this meeting developed. And it developed out of an interest of Senator Garamendi's office and Senator Garamendi. And what's going to be the role in the future of the activities related to the Human Genome Project.

And the people who put it together, I want to just take brief recognition of and introduce Masako Dolan from Senator Garamendi's staff.

Are you here Ms. Masako? Would you stand up just briefly.

And I was going to introduce Dr. Sue Huttner of my biotechnology staff, but she just disappeared with Dr. Watson.

But I would like to at least say these and their staff have done a fine job of bringing all the money together, and a round of applause would be in order for them.

I would like to mention just briefly our technical plans now for lunch. We will have a lunch available on the terrace in which we can pick up lunch. There will be a few

tables on the terrace where you can sit. Others can proceed with this meager lunch that you will have out through the opening of the building and compete for space in the eating service called the bomb shelter, which is just across from the front of the building.

Are there any other technical plans I need to mention?

[No response.]

And we're also very pleased that we can have Senator Garamendi here to comment on just a bit about the importance of collaborative research to the state's economy.

Now, I mentioned that Senator Garamendi here has a strong interest in the University of California; although, he took his Bachelor's Degree at Berkeley, he has learned to appreciate at UCLA, San Diego, San Francisco, Sacramento, Riverside and so forth. He had an MBA for Harvard University. He is a very successful cattle rancher, but is also the Chairman of the Senate Committee on Revenue and Taxation and the Joint Committee on Science and Technology. And I could say other on his background, but we're more interested in hearing his comments.

And Senator Garamendi, if you would tell us a bit here while we have a few moments before we move for lunch, it would be greatly appreciated.

SENATOR GARAMENDI: It's good to be with you. I'm delighted to see the size of this audience and the participants

that are here.

We, in California, like to pride ourselves as being on the top of the scientific and technology of the world. And indeed we have been. And our goal in the state legislature, certainly the Joint Science and Technology Committee, is to see to it that we remain at the top.

There are no guarantees that simply because we're in this golden state and get to enjoy this marvelous weather when the rest of the country is freezing that we deserve to be at the top. We have to earn it. We have to earn it every single day.

You don't get there, and you don't earn this ranking by letting opportunities pass. We have lost many, many major research projects in the last six or seven years here in the State of California. We don't intend to lose future opportunities.

The project, the Human Genome Project, is a huge one, and it involves -- it will involve the entire nation. And as we were hearing a few moments ago, it involves many other parts of the world in the process of figuring out what human beings are all about.

California is ahead. We need to do some things to stay ahead, just in the pure research. My points that I want to make to you are not just that area. Obviously, you are scientists, most of you. Some of you come from industry

and you provide certain tools to the scientific community. It is that part of the linkage that I want to spend a few moments on with you.

The scientific efforts of the State of California in the past have invariably been translated into an improvement in the society in California, in our economic status, in the distribution of wealth to the many citizens of this state, and to opportunities for every individual in the state to prosper, to benefit directly from that scientific endeavor.

We have to see that that continues to occur. I think that anyone that looks at the future economies of the world have come to the conclusion that those economies are going to be based on a large extent on the issues of biotechnology. It is a field that is growing and blossoming and holds tremendous potential, certainly in the area of human health. I think all of you are aware of that, and certainly the Chancellor is.

In the areas of pollution, in the areas of dealing with -- you name the problems that are out there in front of us -- feeding people, dealing with the environmental problems that exist today. All of those things can be reduced or come back to the issues of biotech and biology. There is a tremendous potential here.

We have to make sure that the economy of California has in its foundations for the future a very strong scientific

program in biology and biotechnologies of all sort. This particular project, the Human Genome Project, being funded by the federal government gives us the opportunity in California to maintain our leadership and to advance our leadership in this area. Out of that, if we're able to position ourselves, carry on the kinds of research that are fundamental in this area, we see a tremendous potential for all the citizens of this state and beyond that, the nation and the world, to benefit from our efforts here.

We see our industries providing the tools necessary to sequence the gene and all the other scientific tools that are going to be necessary, not only to do the research but then to take that research on into products or into medicine or whatever it happens to be.

In order for us to make each of those steps, we need to have a very, very close collaboration between the scientific community, the government and the private sector.

One of the primary goals of this conference is to pull together all of those elements so that they are all working together right at the outset.

First, to see to it that we in California have a reasonable share of the federal research dollars that are coming out.

Secondly, to encourage our state government and policy makers in state government to provide the necessary

state funding to backfill, to fill in wherever the research dollars at the federal level are inadequate or insufficient, and also to provide a base of support so that those research dollars can be used here to provide the facilities, the research grants for the scientists and the others that are going to be necessary to conduct that research.

So we're trying to build that coalition in the governmental area, to build the knowledge in the governmental area so that the policy makers, those of us that put together the budgets for the state and set the priorities are ready and willing to provide the money.

We've also brought together here major players in the private sector who have a very, very significant role and opportunity to provide, as I said earlier, the tools and also later to use the research to develop products, and in terms of our whole society develop the economy, to add value, to add wealth to this economy so that it can go back and do something in the next round of research in moving and advancing the citizens of this state and this nation.

So our hope out of this conference -- and I believe that there will be subsequent conferences -- is to move this process along. We've already made, I think, very substantial progress. The fact that so many of you are here today -- I think the first meeting we had we started with three or four of us, and now and then it went about to 20 or 30 at

Berkeley, and now we're well over a 100 here -- that kind of growth and understanding, collaberation, communication is going to be essential in building in the State of California, whether it's at the University of California or Stanford or Cal Tech or any of the other laboratories or private industries, the opportunity for this state to carry on this kind of research and to benefit directly from the research that's going to be conducted.

We know that we can do it in this state. Challenge is one of those that comes to us every day. Fortunately, we have the leaders here among us that will carry it out.

I want you to know that the California State Legislature is aware of this project. We will be doing what we can to make it a priority and to fund where appropriate. We need your input and information to help us in that process of trying to determine where our resources should be spent.

Now that I have sufficiently wet your appetite, let's go have lunch.

Thank you very much for being here.

DR. BOYER: Let me just comment, Senator Garamendi, that had we opened this for full attendance, we would not have had 100 people, we would have had 500 people at the meeting.

SENATOR GARAMENDI: Good.

DR. BOYER: What you have are the people that are

really the most interested. You have the leaders in the field. These are the important people to carry out these things.

Thank you very much for your pertinent comments.

We're adjourned for lunch.

Recess 12:00 p.m. to 1:00 p.m.

Afternoon Session: California's Participation

December 2, 1988

1:00 p.m.

- - P R O C E E D I N G S - -

DR. BOYER: I'd like to welcome you again, but now to the afternoon session, after what to me was a very interesting and intriguing morning session posing many promising prospectives and problems which come ahead.

In the afternoon session, we're going to be looking with three panels on the impact of the genome initiatives on basic research, the need in technologies and hardware and the utilization of the data.

And let me comment that in the organization here, we've had so much interest in participation by capable people that we may have put too much into the pie; that is, we may have more participants that could tell us interesting information than we have time on the panel.

So I need to comment briefly here of how we would like the panelists perhaps to proceed would be to take the information that they have gained in the morning and what they know themselves about the problems that they want to address and in a way of self-introduction speak for five and not longer than 10 minutes each about what their perspectives

are of the issues that face this on the Human Genome Project.

Then that will allow the audience to know what the panelists' viewpoints and perhaps areas of interests are, and then have the balance of the time open for discussion from the audience among the panelists or with the audience because otherwise if we were to get all of the information that would be really useful, our talkers would need to extend far beyond this afternoon.

Now, Winston Salser, Professor of Molecular Biology at UCLA who I found had an understanding of the implications of DNA long before most people in the field, not earlier than Jim Watson, who also himself has been a student and contributor to the area is going to chair our afternoon session.

And I will not try to introduce the moderators of that, but I will turn it over to Winston at this point.

I will mention one other thing. I think this morning that I mentioned our other sponsors. The Lawrence Berkeley Laboratory and the California Department of Commerce as well as the Joint Committee on Science and Technology with Senator Garamendi and the U.C Biotech Program. If I left anyone out, I thank you anyway.

DR. SALSER: I remember as a young kid driving from a farm that we farmed in western Kansas to Wichita where my dad was also a school principal hearing on the radio a popular program about -- it must have been right after

the discovery of how DNA replicated. And I was really intrigued by that. I guess I was a teenager.

Let me tell you a little bit about my perspective on DNA sequencing because I'm going to make a couple of comments that I hope will stimulate discussion.

During the past year I've been -- among other things -- setting up a DNA sequencing facility which will provide automated sequencing for the more than 60 labs here at UCLA that have need of that.

And we were very pleased that we were quickly able to get highly accurate sequences out beyond 500 nucleotides per run. And we think that things are going well and that in the near future that machines should be able to do all of UCLA's current DNA sequencing needs with a very cost-effective manner. So we are very pleased with the technology.

But also I've become aware in doing that about how much further the technology can be taken with some further automation. And to facilitate that I've organized an international sequencing newsletter that now goes out to about more than 700 people that are connected with automated DNA sequencing.

So from these perspectives and from exchange of information with quite a wide number of people, I think I have a little bit of insight into how the actual sequencing part of the genome sequencing initiative might go if it was based on an extrapolation of the current technology, things

that are now in at least prototype form without involving some new things that no one has discussed yet. A miracle happens here, it's another area of magnitude, that really would be delightful.

But I think it's useful to consider: How would it work if we just extrapolated things that we now have?

In the transparency I've broken things down into four areas, and the top is the human genetic map. And actually my own laboratory in collaboration with Dr. Gatti, was one of the 20 original labs in the CEPH consortium that's put together a number of genetic maps of the human genome, and I think it's very fair to all of us that that works. That technology is working very well, and undoubtedly it will be improved and so on.

Incidentally, here at UCLA we've just succeeded in mapping the gene for Ataxia telengiectasia, which is a disease probably unknown to you, but it's turning out to be more important than the disease Ataxia itself since in the past few years it's become clear that that Ataxia gene is responsible for an estimated 20 percent of breast cancer and a large number of other cancers appearing in the heterozygote carriers of the gene rather than, of course, the patients with the neurological disease.

But similarly in the second area of physical maps, Charlie Cantor has pointed out to you the things there that

we don't know how to do and the problems. But from the fact, the success in the Ceanorhabditis elegans case one can take encouragement that a great deal can be done, and that ultimately that will be -- one sort of has a picture of how that's going to be worked out.

There's also the problem of converting cosmids or YAC's, whatever the physical maps are made of, to sequencing clones -- sometimes referred to as the subcloning problem -- and this is a really important technical problem. And I guess we'd like to have perhaps automation of the Henikoff deletion procedure, or something like that, to solve it in a really powerful method.

But what I want to talk just briefly about is the four step, the actual sequencing. I think the report, the NRC report, put it this way: Is the future likely to lie in scaling up automated techniques that are already at the prototype stage, or does it lie in revolutionary new methods which may as yet undiscovered?

Well, as of six months ago when I did a survey as part of the intersequencer's newsletter, I think I can say that currently the best machines are actually working at a rate that would require about 4,000 machine years to sequence the human genome to a depth of 1. And I think that that's the kind of thing that Jim Watson was speaking about when he said that currently sequencing is very expensive;

although, this may already represent an advance, probably does, over what he was talking about.

But the encouraging thing is that it's clear that the machines could go a lot faster. And if we had automation of template preparation and automation of the sequencing reactions, we could go at half of 10 to the 7th nucleotides per machine per year. And with minor changes in the existing sequentures, we could go at about 10 to the 7th nucleotides per machine year.

Now, that means that 300 machine years would suffice to sequence the human genome to a depth of 1, but, of course, that's not adequate. Or you could take four years with these 300 machines to sequence both strands to a depth of 2.

But on the second one, which repeats a little bit of that and goes further to a depth of 2, and the capital costs turn out to be not totally unreasonable. It's expensive. But 300 machines would cost around $30 million whether you bought them from ABI or Dupont, and I don't know if any other machines are out, and if you have to build new space to house them. So your total capital expenditure might get up to $200 million roughly. It's more problematic how much it would cost to staff it. This is just for the sequencing part, not the other parts. But probably for an equal amount, equal the capital costs, you could supply about 15 man years for each of these 300 machines.

So I think this is perhaps doable, but I think that Jim Watson really put it very well when he said, "The question is: How do you interest intelligent people in very dull work?"

What we're doing now is extremely exciting because we're pushing it ahead. We're improving the technology. But now we're talking about -- well, supposing you got the technology, and you said that we now want to replicate this and have 300 machines. I don't know at what stage you'd make that decision. But if you did make that decision, it's immediately a very different ball game. And I just wanted people to think about that in more concrete terms than I've heard of being done before.

And with that, I will turn it over to the first talk this afternoon which will be by Kenneth Gibson.

He was appointed by Governor Deukmejian in 1987 to head the California Department of Commerce. Ken Gibson was trained at Princeton University. And he also brings us the perspective of big business because he served as Senior Vice-President for Kaiser Steel Corporation before he joined the state government.

And as Director of the Department of Commerce, he's the administrative official in the state most directly concerned with his topic this afternoon which is the California Competitive Technology Program.

MR. GIBSON: After hearing Winston's comments, I hope the group will forgive me if I revert back to English. It's always helpful. Sey Siegel invited me down to a superconductivity symposium a couple of months ago, and I had exactly the same feeling. Most of the words were --- I know they were English. They just weren't exactly the same sequence that I've always heard them.

Anyway, I appreciate being here, and I'm sorry that I missed this morning's session. I understand that it was very helpful. What I've been asked to do is describe to you a new program which is just being initiated on the state level that we think should very definitely be material and certainly to our state's technology development.

And I'd like to give you, if I may, just a brief background as to the philosophy of how this program came about in the first place, and then to bring you up to date on the status of it.

First and foremost, the Competitive Technology Program in its initial conception and really the basis for it throughout is that for us it's an economic development program. It has always intended to be that.

And I've quoted a couple of phrases out of Simon Ramo's recent book, The Business of Science, to bring this across. He says:

"As a nation we can't raise our average

personal income by shifting a constant total of assets around among our own people, nor can we count on discovering huge deposits of gold, oil, or diamonds on our land. A sure way for the United States to raise its living standards is to excel in technology.

"Technology can be applied to increase the resources of a nation, to generate wealth that would not exist if the technology were not employed."

Well, I think that obviously what is true in the United States is even more true in California. And even as rich in natural resources as we have been, we've pretty much run out of those. Coming from the steel industry, I can attest to that.

But what we have now is the cold hard fact that there's really only one natural resource remaining in California which can be converted into finished products to the degree anyway that it will ultimately increase our standard of living, and that's to convert our ideas and our brains into finished products and processes for our state.

So our whole idea, obviously, with this program is hopefully to bring the resources of our campuses and our national laboratories into the commercial marketplace in

the private sector.

Our challenge, as we see it, is to be more efficient and a lot more prolific in converting these resources into world-class products and jobs and manufacturing excellence. And it is this realization really that brought around this program and its ideas.

All the elements were here when we looked at what this program should be. We looked at what other states were doing. And other states have been very successful with programs. And we determined that we had a unique situation in California though that allowed us to not have to do some things that some other states are doing.

We didn't feel that we had to go into the venture capital business as some other states are. We didn't feel that we had to get into a position of taking equity positions in companies, which other states are doing. We felt that really the components are here. The technological resources of this state are incredible. We do have the entrepreneurs in this state that are already here. But what we have not had is a vehicle to help them bring all these resources together.

We've achieved a great deal in this state, and our topological excellence doesn't have to apologize to anyone. But we're at a point in a global economy where we simply have got to do more and do it better than we ever have in the past.

So Governor Deukmejian introduced this program in May of 1987. He issued an executive order, and he introduced the program actually with one of his Saturday morning radio shows. And in that -- and I'd just like to read you from that to give you a bit of the flavor for the idea and philosophy behind the program.

He said that with this new partnership, that is the private sector, the academic sector, and the state government, we will encourage and support technology breakthroughs that will have commercial potential through matching funds. We will stimulate industry to take these discoveries and turn them into high-quality products that will generate new markets and jobs for California.

We have neither the need nor the intention to wait until the federal government or some other entity develops a selection process to establish these problems -- projects, a process over which California has little or no control.

So the point of this program, hopefully, is to allow California to take better control of its own destiny. And that's certainly what part of this program is intended to do.

The funds that are available from this program, as far as we're concerned, are for the benefit of California companies. We don't look upon them so much as an incentive to California companies as an incentive to -- as a challenge

to California companies. The point is that the program will provide an opportunity for the private sector which it really has not had before.

We've had some excellent examples on a smaller basis, the micro programs with the University, of course, we think has been very successful, a couple of other pilot programs, RIMTECH as another example. But they are in a very narrow framework, and we simply need a much more broad-based program. And I think we're at least on the right track right now.

As I said, the governor introduced this program in May. Subsequent to that, Senator Garamendi introduced a bill and Assemblyman Chuck Quackenbush introduced a second piece of legislature which were passed in August and signed into law in September. So those were the two pieces of enabling legislation that had now made this new program a reality.

What it does is that it creates, first of all, an Office of Competitive Technology in the Department of Commerce. We are funded at a level of $7 million in this current fiscal year that runs through June of next year. The governor has stated that he will be asking for $20 million a year from then on.

I have to tell you that with the passing of Proposition 98, none of us are sure what our budgets are going to look like. I won't get into an argument with anyone here over

the merits of Prop. 98, but I can only tell you that the cold hard fact is that it is going to wreck havoc with other budgets, not the least of which is the University of California.

But in any event, we'll have to just see what that $20 million next fiscal year looks like. We think it will be substantial. Certainly the Governor is dedicated to this program. It has the very strong support of Senator Garamendi, and it has been a very, very successful bipartisan effort.

So it is not for lack of support that the program may have less funding than what we had anticipated, but that remains to be seen. In any event there are some reasonably significant funds here for us to get started.

As I said, the legislation was signed in September. Since that time, we have been writing the regulations to the program. The legislation was passed on an emergency basis, which means the regulations are also being written on an emergency basis and should be completed in about two weeks. We have a final draft. And so the regulations are virtually completed.

The application process for the program is coincidental to this so that if everything continues to go well the next two to three weeks, we should have the regulations in place and the application process ready to distribute by very early in January.

The program is obviously a technology transfer

program. It is intended to bring the private sector to the labs and to the campuses. Any given project requires, with very few exceptions, private sector participation. It is for the benefit of California companies. We want this to be done in California. And the funding will flow through either a public agency, a university, a national laboratory or a nonprofit organization of some other type.

It also comprehends intra-industry collaberation; obviously, that would be applicable to the biotechnology industry in some areas where discussions with probably three or four different discreet industries at this point who are interested in putting together intra-industry consortia that we think would fit the requirements of this program very, very well.

The program requires participation from the private sector and from California companies. It does not prohibit involvement from companies outside of the state as long as there is a California company or companies that are involved in the project. But the majority -- the great majority of the project work, if you will, does have to take place in California.

The purpose of the program again is to address the competitive needs of our state. It is to address overriding policy issues of our state such as environmental concerns. We intend to see this program -- at least hopefully see this

program addressing environmental concerns, environmental technologies that hopefully will help companies who are already here who are suffering from environmental concerns of one type or another, and find the technological wherewithal to stay in the state.

So it's going to be heavily emphasizing healthy companies who are already here, addressing the issues that companies have existing in California today and, obviously, serve as an inducement to bring other business and other companies into California.

That's the rudiments of the program. The program was written and the statute was written hopefully in a reasonably general manner so that we haven't locked ourselves in it. It is intended to be nonintrusive, if you will. We are not attempting to dictate the policies of any given institution. We simply are trying to create a catalyst and a program that will induce collaboration. And the funds are there for those who see their way to apply for them and to use them.

If there are any questions about this, or if we have time for a question or two, I'd sure be happy to answer any if anyone else has anything they'd like to ask.

UNIDENTIFIED SPEAKER: Is one of the industries that you're talking to the biotechnology industry?

MR. GIBSON: Yes.

DR. SALSER: What kind of a generic kind of a

proposal -- I know that you wouldn't want to discuss specific ones -- do you see coming from the biotechnology industry, especially related to the Human Genome Project?

MR. GIBSON: I don't know that I can answer that right now. I think that's exactly what this kind of conference is there to tell us. Obviously, there are so many steps from here to there. And the program would certainly be available for hoping to achieve the incremental steps to the end result.

So in other words, if we're talking about computer processing or a process achievement along the way, the program -- in other words -- we don't see it as always being on the tail end where a product is going to cough out and we can go buy it at Sears. It's not intended to do that.

On the other hand, it is intended to provide funding for projects that do have in mind the eventual commercialization or a product-oriented end result, or a process-oriented end result, if you will, for the good of the state and the companies in it.

But it would be hard for me to answer that question because I'm not sure where you are on that. That's really more up to this kind of a group and the industry to tell us what they need out of this program.

UNIDENTIFIED SPEAKER: Does your program have anything to say about or do with the public education of new technology, or have any interest in it?

MR. GIBSON: It has interest in it. It doesn't address it directly. Again, I think we'd be very open to that. But there is no concrete records of that in the statute, if you will.

But, again, to the degree that it facilitates the kind of project that we're looking for, we're open for suggestions. And that's really the key point of this is that the concrete is still wet on this program. And we're very anxious to have input from a very wide variety of sources.

UNIDENTIFIED SPEAKER: Could funding for your program serve to attract federal funding for a large project like this?

MR. GIBSON: That's an excellent question. When the Governor said and when we say that we're trying to go our own way here and not be dependent on federal projects or federal funds that doesn't mean that we don't want to attract federal funds.

And we really feel that if our program is successful it should tend to attract federal funding. And a major part of the project is to leverage private sector dollars, federal dollars, whatever there may be.

DR. SALSER: Thank you very much.

By the way, I've been told that there's not going to be a coffee break this afternoon, but there will be a few minutes between each of the segments for people to put

slides in the projector trays and so on. And I'm told that there is coffee on the tables out there, and so you can come in and out.

Our first panel this afternoon is chaired by Dr. Norman Arnheim. Dr. Arnheim is California trained, having received his Ph.D. in Berkeley in 1965. And since 1985, he has been chair of the Department of Biological Sciences at USC.

Many of you also know him, however, from his stint as the Senior Scientist at the Cetus Corporation where he played a major role in developing the polymerase chain reaction approach that many of us are using.

More recently, he's also played a major role in extending the usefulness of this powerful technique. His group has shown, for instance, that you can work with very difficult but very useful material such as thin sections of human tissue, of parraffin embeded tissue, or even single spermatazoa.

And as Charlie Cantor mentioned this morning, related to that technique, this technique for analysis of single sperm sounds rather obtuse, but may become one of the most powerful techniques for fine structure of genetic mapping with human chromosomes which is a very important part of what we're discussing today.

Dr. Arnheim will introduce the members of this

panel who will discuss the potential impact of the genome initiatives on basic research.

DR. ARNHEIM: As was just mentioned, the purpose of this panel is to discuss the impact of the human genome initiative on basic research. And I think it's clear that there are at least two fundamental principles that we can rely on.

One is that there clearly will be a technology transfer, that is all the technologies and technologies that we can't even possibly dream about perhaps will have an impact on people who are interested in doing basic research. And I think there's no question about that.

It's also true that one of the major driving forces for the human genome initiative is the need to understand some basic fundamental principles of biology. And it's also equally clear that the kind of data that is going to be coming out of the mapping and sequencing part of the initiative is going to absolutely require -- in order to understand the data -- some applications of basic biological research.

So I think each of the basic biology and the quest for fundamental information about basic biological principles is going to drive the human genome initiative, and at the same time the human genome initiative in order for it to understand its data is going to be driving the basic research community at the same time.

Now, three of our panelists this afternoon will be discussing how the human genome initiative will impact on general cell and developmental biology in animal models of disease in human genetics and also evolutionary biology.

And, finally, our fourth speaker will more or less bridge this panel to the next one by discussing some basic research that's needed to help in some new technology developments.

I'd like Glen Evans, who is an Associate Professor at the Salk Institute to perhaps start off this afternoon. And he's going to be talking about areas that are related to generating animal models for human disease.

DR. EVANS: One of the -- there are actually from my point of view two major important things to come out of the human genome initiative. One, of course, is simply this sequence of nucleotides that make up human chromosomes or the human genome.

But as Francis Crick is always very quick to remind us, we don't really care what the structure of the genes are. We want to know what they do, what their function is. And one can anticipate that knowing the complete sequence of the human genome is likely to give us a vast amount of information but, in fact, in many cases won't answer that last question: What do the genes do?

To approach that and, in fact, to approach what is probably the most important question in biology at the

present time, which is: How does the genome control or program development such that a single cell can turn into an individual human being requires another step in biology?

And while what I'll talk about in just this few minutes is not directly under the auspices of the Genome Project, it's something that will be evolving at the same time as the Genome Project in terms of technology and techniques.

Genetics has really undergone a major change these days in that one can now do experiments. Genetic experiments were not possible several years ago.

Could I have the slides? Do we control that, or are they controlled from there? And you need to skip forward four slides or five, I think.

One of the ways that we might begin to approach what genes do having determined them is, in fact, to reimplant them into an organism, either in its normal configuration or after making some changes. This is now fairly routine in a large number of laboratories with a technique known as producing transgenic animals.

This is a selection of mouse embryos at a very early stage, essentially a one-cell stage. And under the microscope, one can essentially hold that using micro tools and introduce back into that embryo a small amount of material, a gene, either a gene when it's cloned from a human being, from a mouse, from another organism in its natural state

or one in which one has expressed one's creativity and made a number of changes.

Those embryos can then be reimplanted back into a pregnant animal. And the next slide shows that 21 days later one has mice.

This mouse looks completely normal from the outside. On the next slide though, we can demonstrate by looking at the DNA that it has been genetically modified, and that it now has material present in its genome which was not there before.

The animals that show the extra bands here are the ones where we had planted an extra piece of genetic material, and that genetic material can be from a large number of sources.

One way that one might go about understanding what some of the genes do that will be uncovered by the Genome Project is essentially to reimplant those, having made a mutation in that, then ask: What effect does that mutation have?

This can take a wide variety of -- can use a wide variety of different approaches which can go from simply inactivating the gene using techniques which are just now becoming available, that is turning if off, and asking: What is the effect of that absence of the gene product on the final organism? Or, in fact, one can ask that that gene be turned on somewhere that it's not normally expressed.

A gene that might normally be expressed in the liver, one could perhaps turn on in the brain, and ask: What would be the consequence of that?

Both kinds of approaches are important because many of the genes that one might imagine being important to development if inactivated would be lethal. It would not allow the animal to survive and develop.

The next slide shows that this can actually be taken quite far, in fact, to the extreme in that one can model animals to simulate certain human diseases. This approach is now in its very infancy, but it has the potential of being extremely valuable for making models of diseases for which we really can't imagine the underlying cause.

This is an animal on the right which is completely normal. This is an animal on the left in which we've modeled a genetic defect in the development of the eye. This has been done by implanting a plant gene from the castor bean in the genetic compliment of the animal under the control of elements which turn on at a specific time in development and kill certain cells that are responsible, in fact, critical for that developmental process.

And a large number of labs are in the process of doing those kinds of studies to derive models for diseases such as Huntington's disease or neurodegenerative diseases in which certain cells may degenerate or die at particular

times in development but, in fact, we have no idea what the underlying cause is.

Eventually, as the Genome Project progresses, one can imagine a lot of the genes responsible for those diseases might be uncovered, but we may have no idea how they function in the animal. And these kinds of approaches might allow us to go a little bit further along that.

The next slide, I believe is the last one, is to emphasize the point that these modifications are not only the individual in which the gene is introduced but all of the progeny of that; that is it's a true genetically engineered situation where not only the initial animal but all of the subsequent progeny are also modified. And those animals can then be bred and extremely useful for a large number of studies.

One final word which addresses a couple of comments made this morning is that it is not possible to do experiments in human beings, particularly genetic experiments. Yet, it is possible to do those kinds of experiments in animals; therefore, it's very important that as one begins to understand the human genome that in parallel one begins to understand the genomes of other animals in which one can do those experiments, in particular the mouse which is a wonderful genetic system and becoming a model for manipulation.

Many of the regions of the human genome are, in

fact, colinear with the same regions in the mouse genome; that is if one finds a gene in a particular place in man and one looks in a similar place in the mouse, one finds the same gene.

That's one way to anticipate that almost anything, almost anything, one would find in the human genome, a similar gene would be present in the mouse, and that would allow one to, in fact, approach it in an experimental sense rather than in a descriptive sense.

I think that's all I want to say. And I think we'll lead into David Cox.

DR. ARNHEIM: Yes, our next speaker is David Cox who is a human geneticist, and he will give us some insight about what he feels the consequences of the project to be in that area.

DR. COX: Let me just say that my comments will sort of be focused more specifically on one aspect of what Jim Watson and Charles Cantor talked about this morning which was the overall broad-brush stroke of the Human Genome Project. But right now we're not waiting until we know where all the genes are before we start trying to apply some of the technologies and some of the approaches that we all hope to use in the future.

And in the next couple of minutes what I'd like to do is show you at least one apsect of that that's being

done in my lab and in Rick Myers' lab in collaboration in San Francisco, but just as an aspect of what's going on by lots of different human geneticists around the country.

Where do we stand now over the past five years with respect to recombinant DNA technology in human genetics? And it's no mystery to this audience that it's been a revolution in human genetics by the use of RFLP analysis allowing us to identify where in the genome various human disease genes lie even though we know nothing about the protein products that codes for those mutant genes.

Just a minor list of examples: Huntington's disease, a gene for schizophrenia, a gene for manic depressive illness, a gene for Alzheimer's disease; it goes on and on. It's truly remarkable.

But just as though that's remarkable, it's been equally dismal how in some of those situations where we've known where the link gene was and on which chromosome it lay, how difficult it's been to get the mutant gene product out.

And so while we think about developing ways to sequence the whole human genome, our approaches for coming up with that development, I think in human genetics, is being applied to present day problems. And at least in our labs, it's how to get those genes out and what kind of technologies could then be developed for those specific problems, but

also in more general ways.

So what's the experimental strategy? This is one such possible experimental strategy that one might take. The first is specifically with regards to the Huntington's disease gene, a simple dominant gene for a neurodegenerative disorder. It's known to map to the distal short end of Chromosome 4.

If you simply take purified Chromosome 4 and try and isolate DNA probes from that, the chance that you're going to be close enough to show genetic linkage to the Huntington's disease gene is very, very slim.

So what's an approach that will give you an order of magnitude boost -- as Charles and Jim said today -- an order of magnitude boost of getting probes in the region that you want?

Well, a simple approach would be to isolate the region of Chromosome 4 in a somatic cell hybrid so that it was the only human material there. If you could do that, then the next thing you could do is isolate human probes from that hybrid cell, and you'd be very close to the gene.

So the very first step is a simple minded one, how you just crack off a piece of chromosome that you want.

Once you have those probes though, you have a pot of full of probes, but you don't know what their relative order is to one another. It was mentioned today -- Charles

Cantor pointed out -- the real magnitude of the problem of getting real fine structure genetic maps to order those probes relative to one another.

What you'd like is a genetic map that wasn't 1 million base pairs per 1 percent recombination, but you'd like a genetic map that was much closer to the physical distance so that 1 percent of a genetic unit would be like 100 kilobases. So that's the second thing if one could come up with that.

The third thing is that once you've ordered all the probes, you still don't know where the disease gene is. You have to go back into the affected families. You take probes that you think off of the ordered probes are the most likely to be flanking markers, but then they have to be polymorphic. They have to be informative in the families that you're interested in. One needs a better technology for that for any particular probe. How does one approach that problem?

And, finally, once you find the flanking markers, how you clone all the DNA out between those flanking markers, one needs a better technology for that.

These are the problems that are posed by our embarrassment of riches; that is knowing where all these mutant genes are but having difficulty getting them out because of those four major problems. I'll come back to identifying candidate genes at the end.

So here's Chromosome 4. Huntington's disease is

way up at that light band on the tip of the chromosome. This is to scale, so it shows you the magnitude of the problem because the band is just 10 percent of Chromosome 4.

Can I see the next slide, please?

So one approach that we've taken is to start off with a cell line that has a single human chromosome and a hamster cell with a single human Chromosome 4, and that's shown in Panel A there with the little arrow, that's Chromosome 4. And over to the right in Panel B is that same cell line stained just specifically for human chromosomes. And so you can see that fuzzier chromosome there, that's a whole Chromosome 4.

If you take that hybrid, you expose it to radiation, and in a nonselective way isolate hybrids that have just pieces of chromosome in them, you can enrich be a factor of 10 for getting probes out around the Huntington's disease gene. And that's illustrated down in the bottom of the slide. If you have two pieces of DNA in that hybrid that it greatly enriches. It increases by a factor of 10.

Step No. 1. Now, once you've got those probes, how do you order them? And what we intend to develop was alluded to earlier today is an approach that's analogous to meiotic mapping, but using just somatic cell hybrids.

With somatic cell hybrids then, put a single human chromosome in that are irradiated, can be used like individual

hybrids as though they're individual meiotic products. And you use breakage by radiation as a measure of how far apart two genes are to one another. I don't have time to describe this to you, but I wanted to show you a slide of what the result is.

On the right is the meiotic map for region of Chromosome 21 that is in the vicinity of an Alzheimer's disease gene. On the left are additional probes that are put on the map, plus the ones that are also on the meiotic map. And it shows you that with this radiation hybrid approach the map is expanded 20-fold, and you're able to order probes relative to one another. This is with the 100 somatic cell hybrids with a medical student working two months in the summer. Step No. 2.

Step No. 3 I don't have time to have a slide for, but that's once you've had these probes ordered, then you want to make them polymorphic to go back into the disease families and make sure that you're able to tell one chromosome from another and really know what your flanking marker is.

In the past that's been done by simply looking at restriction enzymes and hoping that you find the single base change, a single base change that changes the restriction set.

My collaborator Rick Myers has pioneered ways of identifying single base changes in human DNA. You think

that can't apply very much to human genetics? It's remarkable on human genetics, not only for identifying point mutations, but specifically in this problem for taking any probe that you want and identifying the single base change that allows you to map it meiotically.

Can I see the next slide, please.

The fourth problem, which is how to clone the DNA, Bob Mortimer will tell us about. The closest flanking markers are likely to be no closer than 1 million base pairs, as Charles alluded to today. If you have 1 million base pairs and you clone 50 kilobases 50,000 base pairs at a time, you don't need to be a higher mathematician to figure out that that's going to take you a long time. But if you can clone 1 million base pairs a time, or 300,000 base pairs at a time, and yeast artificial chromosomes, then one will be able to -- another order of magnitude -- speed up the possibility of getting all the DNA between the flanking markers.

So far I've talked about three technological advances that allow us -- and not all of which have completely come to fruition yet, but which are on the horizon so that they're not just dreams. They're really going to happen, three ways that allow us in a general way to begin isolating human disease genes and solve these problems.

Now, are they going and mapping the whole human genome? No, they're dealing with one specific gene at a

time, but they're general and will be, I think, very useful to the whole Genome Project, but they're very directed to human genetics problems.

In the final 30 seconds or so I'd like to address what I think is going to be the real problem. This is nuts and bolts. We need these orders of magnitude boosts to get the genes out. But the real interesting thing is not getting genes, it's really finding out what they do. But we can't find out what they do until we find the gene in the million or so base pairs of DNA. That's the real problem for the future.

One of the things from my point of view and for right now, we could sequence 1 million base pairs of DNA easily. We could go and very directly with better software identify all the genes in that million base pairs. There's probably 20 or so genes. Right now there's no easy way of doing that.

I would think that that could be a major impact of the human genome initiative on human genetics.

What are the approaches the people take now? Glen already alluded to it. It's using comparative mapping to try and find sequences that are conserved in one organism from those in another.

This is an example that HSA-21 is -- a cartoon of the human Chromosome 21. And what it illustrates is that

there's conserved sequences of human Chromosome 21, large numbers of genes. To the left they are on mouse Chromosome 16, but not all on one mouse chromosome. Some are on mouse Chromosome 17 as illustrated to the right, three genes. And four genes are on mouse Chromosome 10.

So this is very crude comparative mapping, but it illustrates the humps of the genome in other organisms that you can use to identify genes that might be in the region that you want. This is the area of technology development that really needs to come about to get better.

But identifying the mouse sequences, then from my point of view, one can go and make the animal models that Glen just talked about and get two birds for the price of one; and that is using them as models not only as assays for the mutant genes you're looking at, in the case of Huntington's or Alzheimer's disease, if you're lucky. But if you get a phenotype, then you're able to study the biology in those animal models.

So to sum up, what have I said? I've said that I think that we don't have to wait for the impact of the human genome initiative on basic biology. It's here. Even though we don't know exactly how to approach and do the whole human genome, approaches that are stimulated by thinking about that are having major impacts on human genetics right now, and I think that they're only to get better, largely

in terms of helping us figure out where the genes are in the large pieces of DNA that we isolate, but more importantly constructing animal models or cell models that allow us to understand the function of the genes.

DR. ARNHEIM: Our next speaker is Walter Fitch from the University of Southern California. And Walter will tell us a little bit about how he thinks the initiative is going to affect evolutionary biology.

DR. FITCH: I'll stick to evolution, but I hope I'm less technical, that was my intent. And I may be thinking a little broader, perhaps over generally.

But I want to start off by first of all saying that -- well, most of you understand evolution, I think, as a process akin to a family tree. The genealogy of your family is not all that different from the genealogy of organisms and really can ask about the relationships of animals and plants and plants to bacteria and all of that.

With the sequencing of the genome, there's another kind of family tree that I'd just like to bring to your attention. Our chairman in his youth studied a protein from eggs of birds, but humans have this too; it's called lysozyme, and it breaks down bacterial cell walls and is part of our ability to resist infections.

Now, one of the things that can happen in the course of the history of a group is that a gene can duplicate, and

this gene has done it. And so you have sort of a brother and sister here. And in this case the sister evolved and changed in such a way that while we can clearly recognize that this is a brother and sister, the sister is really quite different and is absolutely required for the lactating female to synthesize milk. And your first reaction is likely to be: What possible relationship can there be between something that breaks down bacterial cell walls and making milk for human lactation? Well, I won't go into why it's there, it's obvious if you get into the details.

But this process of creating a family is something that goes on continually in the genome. There are lots of these. Some families are larger, perhaps we should call them clans. And asking: What happened? When did they happen? in the case of the duplication that I just described, not surprisingly that gene duplication that led to this lactating gene, lactation gene, occurred right at the origin of the mammals. And it's present in all the mammals, but it's not present in birds, unlike lysozyme which is in both.

So we can ask questions like this. And there are some clans in which there are hundreds of copies of genes, the duplication has occurred many times in this spread. And we can ask: How are they related to each other?

Searching for these is a simple problem in the first place. If you've got one pattern, you look to see

if there's anything like it, and that's how this lactation gene was found. You look to see if anything in your data base looked like what you already had, and that's how it was found. That process, of course, will continue.

But in terms of the future -- and here's where I'm going to be a little nebulous -- in this case, we had a simple problem. We knew the pattern that we were trying to match, and that's easy computationally. What's really going to be fun is to search for patterns that are meaningful -- because every time you look at something there's a pattern there. The question is: Does it have any significance? -- without knowing the pattern you're trying to look at. And, of course, you can do that by things like restricting it.

Supposing you would just look at pituitary cells, that is brain cells of some sort. And then you ask: What things do I know that occur only in the brain and nowhere else are present in my sample? And are there characteristics about those that are not characteristic of things that are turned on in other kinds of cells like the liver or the lung? And you start looking then for things that you don't know what you're looking for, but you know that if you find it, it's going to be a great help in terms of answering questions like development, answering questions maybe about disease and a lot of other things.

So part of the problem is unsolved, but there will

be a real cottage industry out there of people looking for techniques for finding these patterns. And how can you go about searching for them, whether they're at the protein level or at the gene level?

Another thing about this is trying to find out these relationships that I've just described. We'll be doing that across species, of course, because we're going to be learning a lot at the same time about genes from other animals. So we will be getting that classical kind of evolutionary relationship that I mentioned. But we will also be getting them for the duplications of the genes within the genome. And finding those relationships is rather a tiresome process.

It turns out that if you had 100 different genes that were all related to each other, and you wanted to know their ancestory, the number of possible histories is greater than the number of stars in the universe, maybe greater than the number of protons in the universe.

So you've got to have good efficient ways of attacking those problems, and that's going to be a part of what's going on in developing new methods. So that's another new thing.

I think it's going to have an impact in the following sense: What I've said up to here is what we sort of need in order to brouse through this large volume of material that we have.

What's going to happen in the future, I think,

is that scientists, especially in molecular biology, are going to start doing the research rather differently, that the first thing you do before you write an NIH grant is go search this large encyclopedia of material to see how much of what needs to be known is already there before you screw up on your grant application and get rejected because its been done or it's obviously there and you just didn't look for it.

And so we're going to have a whole generation of scientists who are going to be looking at this kind of prospect before they even do any work. And, in fact, there will be people like me who, in fact, will never leave that encyclopedia to go into the lab and do any wet chemistry. They're going to spend their life browsing through this looking for the patterns that maybe the people who did the hard work and got the sequences -- although it maybe it won't be so hard when everything has been done here -- maybe have missed.

And so I see that there will be a whole new pattern of the way that research is done. I think the implications are very strong for that.

DR. ARNHEIM: The next talk will be by Bob Mortimer who is from Berkeley and is the Chairman of the Department of Biophysics.

DR. MORTIMER: Former chairman.

DR. ARNHEIM: Former chairman of biophysics. He's

going to talk to us about YAC vectors.

DR. MORTIMER: Well, just by way of introduction, I'm not a human geneticist. My specialty is working with yeast genetics. I'll make a point that yeast is not that far away from humans. But I have spent a lot of time working on genetic mapping, developing the genetic map of yeast.

We have not had the audacity yet to compare the order of genes in yeast to that of a human, but maybe it's not an unreasonable thing to do.

I want to talk about YAC cloning, yeast artificial chromosome cloning, which is one of the approaches that is being considered for developing physical maps of the human genome. The other principle is cosmid cloning. And I think that after I'm through, you'll see that there's reason that both of these are still being considered and possibly other approaches.

Could I have the first viewgraph, please?

The procedure was described by Burke Carle Olson a couple of years ago. And basically there's a special vector which they constructed that contains two telemeric sequences, and if one clones into this exogenous DNA from any source and transforms it into yeast, these will behave as extra chromosomes in yeast, that's in principle. So this is the vector, and then the two arms, and then basically the chromosome here. And this then can be human DNA.

And Olson developed this method and has had a lot of success cloning human sequences into yeast. And several other groups, including our own, have been working on this as well.

Can I have the next one, please.

One of the technologies that needs development is making viewgraphs that have a sufficient range of grays. This is all just black and white, and there should be some in between. But the point is that the size of YAC's can vary anywhere from about 50 kilobases up to the order of a megabase.

This is a total [inaudible] digest of the human genome. The bottom is about 50 kilobases. Up at the top is several megabases. And the experience, at least in our group and I understand most groups, is that we're not obtaining any representative sample of the restriction map -- restriction fragments from not one. And I think it would be necessary to obtain such a representative sample for developing a proper physical map.

I guess this indicates that it would take a large number of these YAC's to cover [inaudible.]

If you want end in a particular chromosome -- can I have the next one, please -- it will be necessary to separate from this large library of YAC's those that come from a particular chromosome. And this will either by standard techniques

will involve some form of hybridization with human-specific or species-specific DNA probes.

If one can obtain purified or semi-purified chromosomes either from the human cells or from rodent human hybrid cells, then the job is much easier. In other words, if one in principle is interested in Chromosome 21, if one could have a large sample of Chromosome 21, and make YAC's from that, then the job is much easier.

Current technology -- the best technique is [inaudible] activated cell sorter technique, and it can purify chromosomes but not in sufficient quantity for YAC cloning.

I wanted to now just discuss some areas that I think call for additional research, and the first one is related to this.

Can I have the next one, please.

Related to the last point would be to explore new areas for separation of large numbers of specific human chromosomes. Several years ago it was shown that it was possible to separate [inaudible] chromosomes by [inaudible] and several groups are exploring this procedure again.

The pulsed-field gel techniques that were developed by Charles Cantor can separate DNA molecules up to about 7 megabases in size. This is still in the order of 4 or 5 times smaller than the smallest human chromosome. Future developments in this area are marketable, I can't say.

Can I have the next slide, please?

This just shows the study that was done several years ago. This is the chinese hamster karyotype.

Can I have the next one, please?

And this is after one passed through in a zonal centrifuge. at 3 runs , and three different fractions, you can see that there is an excellent purification of the smaller chromosomes. This was a study done 15 years ago. So I think it's true that this approach can at least give enrichment for particular sizes of chromosomes, but not purified samples.

Another point, the size of the YAC transformance that our group is obtaining, and I think it's a general -- is not representative of the size of the fragments that are in the restriction fragment unless one does presize fractionation. And this could either be due to selective sharing of the larger molecules or some selection in the ligation or transformation step or possibly instability of the larger YAC's in both cells. I think the latter possibility is not too realistic.

But this would call for research in manipulation of large DNA molecules and also in general studies of yeast transformation, ligation and transformation.

Well, another area that -- the size of YAC's is determined by pulse-field gel electrophoresis . For instance, this is the one that I showed you originally. These are lambda. And the upper gel -- the only difference between

these was the pulse time. The upper one is 55 seconds pulses. And you can see that the gel is spread out between 50kb up to over a megabase. If one goes to 22 second pulses, then only the bottom three of these chromosomes are resolved, but their spread out over most of the gel. If one one goes to the 15 seconds, then regions smaller than the smaller chromosome are spread out.

And so by selecting particular switch times one can expand different parts of the gel and resolve different sized YAC's with the yeast chromose.

But as I say, most of our information on [inaudible] is empirical, and I think a lot more work is needed to give better understanding of these factors.

One other area that I think calls for research; in fact, all of the physical mapping procedures that I know of involve autoradiography of gel blots and using film. If a two-dimensional data detector were available which could be just placed on top of the blot -- and I think such things are feasible from just talking to physicists -- that this would allow automation and automatic digitation of the data. And it seems to me that this is an area that calls for future research.

One area that we're quite interested in just strictly for basic reasons, if one had a diploid cell that carried two yeast artificial chromosomes, which are mostly human

chromosomes, the question is: Would they recombine at frequencies normally seen for yeast cells or frequencies normally seen for human cells. There's about a 300-fold difference in miotic recombination frequency between the two species. And I think that it would be quite an interesting point to look at. I haven't the slightest idea which result would occur.

Finally, just David Sheldon in my group, just relating to the evolutionary argument, is cloning human genes by functional complimentation of the yeast mutants. And has obtained several of the purine metabolism genes, human purine metabolism genes in this case.

DR. ARNHEIM: I think the first thing that we can do is just ask if anybody has any questions that they would like to address for our panelists.

DR. SALSER: Maybe it should wait until after we get into the next one, but I'm very interested since these are not representative, and similar problems may occur with the cosmid physical maps; are any of you in the cosmid or the YAC groups thinking about how to overlap your sets of data and piece it together from all of the -- together?

DR. MORTIMER: Yes.

DR. SALSER: Have their thoughts only been frustrating?

DR. ARNHEIM: That will be discussed in the next panel.

DR. COX: They're all at very different levels.

So going from the level of cracking off a whole hunk of chromosome to then making YAC's is -- and ordering those YAC's -- is at one level that you can well understand.

DR. SALSER: I realize it's not a trivial technical question that I'm asking.

DR. MORTIMER: I think Maynard Olson has one approach that gets around us by -- such as using size fraction equal our one partial fragments. I think that's probably obtaining a representative sample of the total genome by this approach; as we try to go for the huge one, something in the order of 300 kb.

DR. EVANS: I think a number of people have shown with yeast that by doing an awful lot of work you can begin to overlap by cosmids by picking them at random, restriction mapping them, putting all of your information into a computer with an appropriate program that will match things up.

I think there's no question that that works with a certain limitation. There are a number of people around thinking of ways of doing the same thing not requiring such a vast amount of work, number one, which involves analysis of multiple clones once, multiplexing, and now I guess to the George Church sequencing method and then having appropriate programs to sort that out.

And I can think of three or four different groups in the process of working on those things. Some of the ideas

appear to work. Some of them theoretically will work, but technical limitations such as the Hans Laroff method in using all of the nucleotides are very difficult to overcome.

In my mind, the technical advances that will really make that efficient are two-fold. One, to be able to speed up the process by doing multiplexy. Secondly, to have machines do it rather than people because it's very difficult to convince postdoc's to do that experiment, practically impossible to get graduate students to do it, but robots like that kind of thing.

DR. FITCH: As a remark of frivolity, I was reminded during this conversation of my youth which was preceded Dr. Arnheim's by quite a bit I think. But I was listening to Fred Allen who routinely visited Allen's Alley, one of whose residents was Falstaff Oppenshaw, the poet laureate of the alley, one of whose poems was entitled, "Alas, Alack, What Is a YAC?"

DR. BOYER: I had a minor question for Walter Fitch. This humanian lactation gene, what's known about its biochemical function?

DR. FITCH: A great deal. The association between the lysozyme which breaks down the cell walls and bacteria, they happen to be carbohydrates, the cell walls. And so the lysine breaks a particular bond.

In the case of the gene related lactation, what

it's trying to do is to impose the system a recognition of a specific configuration of carbohydrates so as to join them together to make lactose.

DR. BOYER: It's actually the lactose synthetase?

DR. FITCH: Yes. It's not the whole synthetase, it's a part of it.

DR. ARNHEIM: I'm wondering whether there's anybody from the industrial sector who might want to make a comment concerning what they feel the human genome initiative might -- in terms of their basic, if they have any basic research programs, but in terms of their basic research programs -- what impact they might see. Is there anybody here who would like to volunteer?

[No response.]

DR. ARNHEIM: No.

Are there any other additional questions? Otherwise, we're more or less on time, and we can go to the next session.

UNIDENTIFIED SPEAKER: It's not a question. But I'm curious. Do you think as a valid to labor, I think that what I just heard is that contrary to what some people feared, it sounds like the Human Genome Project -- technology if it's accelerating, there is also greatly accelerating assisting biological basic research. Is that a fair conclusion to what I just heard?

DR. ARNHEIM: Well, certainly, I think a lot of

us would agree with that conclusion, yes.

DR. FITCH: I mean it's absolutely necessary.

DR. COX: I would say that that's a very succinct way of saying what I've tried to say, yes.

DR. ARNHEIM: I'd like to thank the panel.

Recess 2:20 p.m. to 2:30 p.m.

DR. SALSER: I thought it would be good to give everyone about 10 minutes I think we've taken to stretch your legs. But it's time to get underway again.

It's a pleasure to introduce the moderator of our second panel, Tony Carrano. He is a section leader with the genetic section at the Lawrence Livermore Laboratories of the Department of Energy and the University of California.

Dr. Carrano also received his Ph.D. from the University of California, Berkeley.

I've known Tony for a very long time, and I've always been impressed with this group's ability to apply very elegant physical and instrumentation techniques to carry out difficult genetic analyses.

In the National Gene Library Project, they used flourescence activated chromosome sorting to make chromosomal assignments for various genes and also to make chromosome-specific human clone banks.

More recently, Tony has been successful in using the ABI, Applied Biosystems, DNA sequencer as a powerful tool for a new purpose, not for sequencing but to automate the creation of ordered sets of cosmid clones to create the so-called physical maps of the human chromosomes just as was discussed in the last talk of the preceding panel.

These physical maps are going to make clone by phone an incredibly powerful approach, and that's totally independent of the uses of the main goals of the Genome Project. But they're also, I think, an essential starting point for a well-organized sequence analysis program.

Tony is going to introduce his panel, who will consider development of the needed technologies, both hardware and software.

DR. CARRANO: Thank you, Winston. It's good to be back at UCLA again.

As Winston indicated, the charge to our panel is to discuss the development of needed technologies as well as hardware and software.

What I thought I'd do to try and set the mood for this is to actually take the biology of the human genome initiatve as we had seen displayed and try and put it against the disciplines and the technologies that might be necessary to get the initiative accomplished. And I think the first viewgraph sort of sums that up.

The human genome initiative really is at least two projects, two major projects, and separate in terms of the technologies that are needed in a sense. It's, first of all, a project of ordering; and, secondly, a project of DNA sequencing.

We start with human material in the form of a cell and we want to wind up with the genetic code, the DNA sequence. And there are really many steps that we have to go through to get there, and there's many technologies and disciplines involved.

For example, as we heard earlier, the human genome is huge. It's 3 billion base pairs in size. And to tackle that problem as a whole is extremely difficult, if not impossible, at least with the present state of technology.

And so everybody's consensus is that we've got to fractionate this human genome in some way. One way is to break it up naturally into its components, the chromosomes, the 24 different human chromosome types that we have. Another way is to break it up into yeast-artificial chromosomes.

We've heard about yeast cloning from Bob Mortimer a few minutes ago. The other approach is to use physical methods to separate individual chromosomes. We've pioneered at Livermore and at Los Alamos simultaneously the development of fluorescence activated cell sorting to purify chromosomes. And this basically separates chromosomes on the basis of

DNA content and DNA base composition.

Now, it's an interesting feature because this particular methodology can analyze chromosomes at the rate of several thousand per second and it can separate chromosomes at the rate of 700 per second. Now this works just fine -- after several years now -- it works just fine for cloning DNA into certain types of vectors, lambda vectors or cosmid vectors.

But it doesn't provide us enough DNA just yet to clone into systems which require a little more cloning efficiency such as the yeast-artificial chromosome.

For example, we can sort on good days, purifying good days, about 5 million chromosomes of any one type. That provides about 1 micrograms worth of DNA equivalent. And as I talked with Bob a little bit before at lunchtime, it looks like that to clone into yeast-artificial chromosomes it's going to require about 10^8 chromosomes or 20 times that much DNA, or on good days, 20 good days of actually sorting with the instrumentation.

That's sort of a balloon of practicability right now. It only gives you one shot. If you miss you have to do it another 20 days. And so right now we're not quite there.

But there are other mechanical methods that can accompany that such as prepurifying the chromosomes on systems and then using that as an enriched fraction to sort.

And all these things can be looked at.

Well, once chromosomes are purified, then one has to then go into the molecular biology genetics and clone these pieces, fractionate the chromosome further and clone the pieces of DNA.

And this is an interesting area because we don't have to wait until a project is finished to get something commercialized. We immediately create a set of consumables right here. We immediately create a new set of vectors, cloning vectors, which are marketable, are being marketed. We immediately create libraries of specific chromosomes which are marketable. We immediately create from this probes from that DNA where certain genes are markers for genomes which are marketable. So this creates a set of consumables which can and are being commercialized right from the very start of the project.

Once we have these libraries, then one has to put them all back together again, take these little pieces that we've now got separated into little tubes and put them all back together again.

So we've actually come a full circle. We started with an entire chromosome, cut it into pieces and then we want to put it back together again.

This is a very labor intensive process, and it's also a process that requires a heavy dose of mathematics

and statistics. So the mathematicians, staticians and the computations experts get involved in this problem to a very large extent.

We have developed a technology, as Winston just mentioned, to automate this process. The present methodology relies upon not radioactivity but using the fluorescence tag to label these fragments and to analyze them on automated DNA sequencer, which is commercially valuable. This was a project that we did in collaboration with a group at Applied Biosystems in Foster City, California.

With the technology that we have, it allows one to progress at the rate of putting these pieces together, and I believe it's about a million bases per week. There's still a long way to go to do the whole human genome, but this can be improved further, we think, by at least a factor of 4.

Once these pieces are put together, then one can take the probes that are established here and pull out those pieces of DNA that contain the genes of interest, the genes of interest that you're working on which could then be sequenced, that could then be used certainly in diagnostics or the treatment of certain diseases. And so there's definite commercialization and human benefit at this end right here.

So what I want to impress upon you here is that this is really a multi-disciplinary project. It involves

physics. It involves chemistry. It involves engineering. It involves mathematics. It involves computer sciences. It's not just a biology project. And so what it requires is that groups get together. Fortunately, at least the national laboratories, all those components exist. At universities those components exist. It's a question of how you can get them together to talk to get the project done.

Now this group is going to address some of the issues related to these technologies. And our next speaker up I think will sum up what we're going to talk about I hope. The components of that, I mentioned they're here. They're going to continue to be here, and they change. They change rapidly. What's being marketed today in terms of vectors and libraries may be different than what's marketed six months from now. So it's an ever-changing market.

Mapping I've briefly mentioned. But an important component of this whole process is robotics, and we're going to hear a little bit about that and how that can --- other technologies in this area can help accelerate this process.

DNA sequencing, of course. We've heard some of that all ready. Multiplex is still in its infancy. It's still in what I call the first generation perhaps of DNA sequencing, and I can see at least two more generations coming beyond that right now what's on paper at least.

We need the hardware, and we certainly need the

software to be able to put all this information together and to analyze it and to disseminate it. We're going to hear a little bit about that as we go along.

So each of our speakers will address these areas. Our first speaker is going to be in the area of robotics, and that's Dr. Nebojsa Avdalovic. He graduated from the University of Zagreb in Yugoslavia where he got both an M.D. and a Ph.D. degree. He came to the United States and worked in Birmingham, Alabama and then at the Rissler Institute in Pennsylvania. And then he came to Beckman Instruments in 1985 in Palo Alto where he's associated with the Spinco Research Department there. And he has been working on exploring robotics in DNA sequencing and mapping and has made contributions in the automation of the Sanger sequencing reactions using vector robot.

DR. AVDALOVIC: I hope I wouldn't murder the English language the same way he murdered my name.

I'm going to try in these couple of minutes which are allocated for my discussion to cover certain topics.

Would you please show the first overhead.

I would like to address some need for automation in robotic devices, and then I would like to show some examples of what is available today and perhaps what is needed for tomorrow. And I will address a couple of questions of impact of the Human Genome Program on certain issues for each goal

beyond science, or direct science.

If we stripped the past of bare necessities, and you decide to invest more than two days a week in molecular biology, as some speakers mentioned today are investing, and you start really planning your research in trying to address these issues, you will now realize that this is an enormous task. And I think the best introduction of that was done by friend and colleague from Lawrence Livermore, Tony Carrano, when he fractionated that task, which is really an enormous task, into manageable components.

So if I just take two examples for this discussion, the first example I would like to mention is making an order to overlapping libraries. I'm not going to go into a discussion of how you're going to pick up your first clones. You can do it by hand today. But if you realize that to make a reasonably good overlapping library, you need approximately five genome equivalents, that's the minimum of the minimum. This is approximately 1 million clones. It's not such a big task to do it by hand, but you could. And the reason we are doing this by hand is because we would like to assign parking space in a [inaudible] for each of those clones. If you are smart, you will make this in and duplicate it.

If you don't do it in duplicates, about 10 times a month -- microtiter plates. If you would like to make replica clones, if these are by some chance [inaudible] they

would like to make this as replicas, it means 2 times million replica clones, which you will like to screen by certain means. You will also have to isolate DNA for probing. And those people like Tony, or some others, who are trying to characterize each and every clone by doing a restriction mapping, either using single or double digest, or using partial digestion maps, it means 1 million -- approximately 1 million clones isolated and labeled for screening.

You can see that this is not an easy task. It will require enormous bookkeeping to start with if you want to have a use. So you will have to have robotic devices, or a device which is capable of reading bar code on your microtiter plates, which will also manipulate those microtiter plates and understand and realize the X/Y relationship in that microtiter plates.

On the other hand, if we would like to do plating or making a very dense restriction map of your clones either on a [inaudible] or on an Agar plate, you will have to have a robotic device which will do it because the desire is to put at least 100 microtiter plates in the same size as is in microtiter plates. This cannot be done by hand. The reason you would like to do it would be because you want to decrease the amount of possible manipulations in hybridization.

So you will have to read those. You will have to have an automatic system which will recognize each parking

space of your clone and relate it to your existing data base with your microtiter plates and then play with with regard to retrieval, cataloging and manipulating.

I even didn't mention here the need for a computing abilities to make some overlapping clones. I'm not even addressing this at all.

As an example of the trouble with autoradiography or fluorescence, Tony Carrano mentioned a machine made by ABI which was run in conjunction with the University showing good collaboration with industry and the University if you choose the right partners, and which was vastly approved by ABI. They can use restriction mapping and fluorescent technology and somehow enter those data in the data base.

But if you are doing this with autoradiography -- could I see the next slide, please?

This is example of how high-resolution restriction mapping done on C. elegans. These are the number of restriction fragments, and it's standard in between, which you will have to deal with if you want to enter those data and make overlapping segments.

Well, to tell you the truth, somebody said that this technique might be working. Well, it might be working for small genomes. Here, when you are looking in an E. coli system, you are examining overlaps or relationship 1 clone versus 200 clones in a library. But when you take the human

genome, this is 1 clone versus 2,000 clones, or 200,000 --excuse me -- clones.

In that case, physically you either have to have 100,000 times more information to produce that or you have to have 100,000 times bigger fragments. Well, good luck with YAC's if you think you can do that here. So there are enormous problems.

So what I'm trying to tell you is that we have to have a little more humility in all these aspects because our aspirations at the moment are much higher than our abilities are to do that job.

Could you go back to the --

So that's a big issue with autoradiography. I will have similar remarks in analyzing DNA sequencing later.

Could I see the next -- thank you.

The other example is DNA sequencing. You have to isolate and purify DNA. And there are more and more data showing that the purity of DNA is extremely critical for good results. It's not only for radioactive sequencing, but it's even more so for fluorescene sequencing, especially if you would like to have a double-standard DNA as a source for your sequencing. As I say, this all comes from us who are spending more than two days a week in doing the work.

So now when you take a robotic device perhaps and you say, "I would like to run 100 or more templates per day,

which is possible -- I can show it to you -- is 400 pipetting steps, 100 to 400 loading for electrophoresis. I say 100 versus 400 regarding this using fluorescent sequencing protocal where you can combine certain lanes together, or radioactivity where you have to run a lane separately.

I would like to see you really more than the two days in a row sequencing films of the size 14 by 17 inches carrying 24 templates on it.

It was a good remark of a colleague of mine who said that -- when I asked him: "How do you do your sequencing, and do you have any problems?" He said, "No, no specific problems, but I'm changing my technicians." I said, "How do you change them?" He said, "Every 100,000 base pairs."

So on the other hand, again, if you would like to read those interesting films automatically, we need a device, and right now we don't have it. And if you have it, you will have to shotgun merge 50 to 70,000 base pairs in one day on a small computer, or a big computer if you want.

I did an exercise, which was an exercise in futility, I cut genome in manageable pieces of 350 base pairs as a source of possible sequencing data and did some manipulation to use single base changes, and then trying to do a base on a computer, and I can tell you this is a very tough task for just 4,000 or 4.1 kilobases of bbi.

Right now I'm trying to do a similar simulation on the EBB lighters which are 170,000 base pairs, and I can visualize what kind of problems one can get on a computer. We just don't have that technology.

So what I would like to show to you is a humble approach to the present day needs by showing a robotic workstation which Beckman Instruments has developed and which is successfully being used for DNA sequencing in many laboratories all over the world right now and is capable of addressing both fluorescent sequencing technology and radioactivity protocols.

Some of the aspects on that workstation were worked out together with Dr. Lee Hood's laboratory whose colleagues, Wilson and others, have contributed -- and Steve Clark -- contributed by designing a nice heater system which now comes with a robotic station.

And then I will show you what is in the making with some new aspects of robotic device, and then we can talk about automated electrophoresis and detection which was accomplished very nicely by our colleagues at ABI. And I will show you some data with electrophoresis which is a fully automated system.

So this is a robotic device which has pipetting device tools on the right-hand side. The tips and the heated plate on one side, the plate, platform and microtiter plate together with the solutions.

When you look at this scheme you can see that where the tubes are located by pipet tips, microfish tubes carrying the samples of the [inaudible,] and then the heating plate where you can heat inside from 10 to 90 degrees at your desire, and you also have a 4 degrees cooling system.

Then you make some patterns for those reactions. And since the time is running you can incubate this very easily at any desired temperature using any enzyme available today.

And this shows the profile of the temperature and the time which elapsed in doing all these jobs.

This will give you your chance at running [inaudible] to run 24 templates in a run which gives to you approximately about 12,000 base pairs per run on that machine.

And this is the approximate result which you see here on the right-hand side where the gel was done by machine. The left-hand side was done manually. And you can see there is practically no difference. It's the same quality done by machine and by hand.

This is what is new. This is a side-loader arm which helps you to change those microtiter plates from your workstation so that you can run multiple reactions. They are like towers, like hotels with the microtiter plates, or test tubes if you need them, or if you want to have an electrophoresis apparatus on it too.

So basically this system will allow very high throughput in both sequencing and mapping because the tools are being made which can stamp the [inaudible] both on an agar or on the filter paper, and in this way you can exchange and go through the process with not only one plate but with multiple plates.

This is one of the towers with different kinds of consumables on it.

This is something to provoke curiosity. Automation of electrophoresis is becoming a reality by using capillary electrophoresis. Where the driving force in an open tube is not a regular charge mass ratio like you have in a polyacrylamide or Agarose system. And this shows the example that you can separate in 10 minutes using [inaudible] gas system from 12 to 30 [inaudible] in 10 minutes.

Thank you.

I didn't have time to address some impacts, but we can talk about this later.

DR. CARRANO: Our next speaker is going to address the area of DNA sequencing, and that is Dr. Norm Whiteley whose from Applied Biosystems. He received his Ph.D. in '74 from Harvard University. And he's worked on the DNA synthesizers, the DNA [inaudible] and the sequencing.

DR. WHITELEY: I'd like to be in some ways a little more optimistic about the ease of some of what we're about

to do. Everybody has talked about how hard it is. I think there are some aspects of it that are very hard. I think there are some aspects of it that require a lot of work but are relatively straight forward, try and draw the distinction between research and development as it were. Development may actually require much more work in terms of man hours and effort and dollars, but it's relatively straight forward.

It comes back to the little question that Dr. Watson addressed this morning which is: How do you get all these good people to do this? I think what he was thinking was: How do you get good researchers to do this? I think the key is that you want the good researchers to do the hard creative parts and you want other kinds of people to do the other parts because they like to do it, and they're good at it.

In order to sequence very large amounts of DNA, I think everybody has almost totally agreed that we need a lot of automation, whatever the techniques. And the rest of this brief talk I'm going to focus on the existing scheme of somehow generating clones doing, if you will, Dideoxy sequencing or perhaps Maxon Gilbert sequencing, electrophoretic analysis of that and conventional data collection.

At the moment, there's essentially no automation of the clone, and that's the part that I was referring to as very hard. Part of the slide is what we need and what

people are working on are global strategies for the mapping and generating sequencing clones. The point has been made a lot. I agree with it.

The other three areas doing the reactions, we just saw a way that's been automated by our colleagues at Beckman. And there's a number of other systems on the horizon, and that's about to truly -- or has really happened. I should say it's about to truly happen.

The data collection, as you know, our company sells an automated DNA sequencing based on fluorescence. There are a number of other DNA sequencers commercial, or in the works, and there are a number of people, and I think some will address it later, who are very interested in automated data analysis.

The point is, those last three things are, if you will, in the chute. I think that they are totally within the range of being able to do the kind of work we are talking about, that the real holdup is the first step, that many people have addressed, that is truly the really hard point.

May I have the next overhead, please.

To illustrate my point that the others are in the works, this is a huge slide indicating the progress in commercial automated DNA sequencing. This is not -- this is something you can buy in 1986, or you can buy now, and just in that time from the middle of '86 to the middle of '88, about

two years, the number of bases that can be sequenced, and the accuracy with which they can be sequenced, has improved fairly dramatically, as well as a decrease in the amount of DNA involved.

I think that that -- all of that improvements has been based on changes in [inaudible] changes in software, none of them changes in hardware. That graph will continue for awhile, in terms of improving throughput.

If I can have the next slide.

To do a thought experiment -- I am not proposing that someone do this -- to do a thought experiment, you just want to take that fraction of the process and automate it with today's existing commercial, off-the-shelf technology. What would it cost you? What would you have to do? Well, you could do about 500 bases a line, and if you were to run it twice a day -- which nobody in a research lab would do because that means coming in in the middle of the night to turn it on the second time -- you can figure that to have a 1x coverage of the human genome is about 750 instrument years, or for 4x coverage, in 12 years that is 250 instruments. That is $23 million of hardware, bought one of, which again, no one would do, and it would take you something like 50 people, maybe 100 people, to run those things.

That is within the scale of the kinds of money and time the people are talking about. That is today's

technology. Easily there is a factor of three in throughput there that will happen in the next year or two. Without any doubt, the company that doesn't do it will go out of business, regardless of whether it is us or our competitors.

So, my point is you really -- this is within the realm of reasonableness already.

May I have the next slide.

Now, obviously, if you can get very clever and inventive you will do it even better.

The next point is that in actual sequencing reactions, how are they done? However clever we get about this strategy, we are going to be repetitive. They are going to be done on a large number, when we are talking about something like sequencing the human genome, and you want, obviously, routinely high quality. Whoever does it, whether it is done under an academic auspices, or university, or a national lab, or an industry, it is a manufacturing type of operation. And, again, I raise the word "contract" which offends a number of people, but it is inherently an operation in which you want to define what you are doing well enough that you don't need research scientists to do it. You, in fact, want manufacturing type of people doing it.

To use an example, DNA synthesis, a few years ago, was worth a Nobel Prize. It is now currently the reagents and the instruments are made by people who don't have

college degrees. They are made every time, perfectly, consistently, they work. It is a routine thing. The reason it is routine is that some development scientists, not research scientists, took the time to figure out exactly how to do each step, so that they could do it right every time, and then they hire some people who are very bright and care about their work, that don't know anything about chemistry or biology necessarily, but they know that they need to be careful with what they are doing to execute that every day. It works quite well. They are regularly available, and that is the kind of operation that is totally feasible, even at the scale we are talking about, if we had the strategy to do it.

Another illustration, about ten years ago I worked for a company that made clinical diagnostics, and they made 100 million reagents packs a year, and it is complex chemistry. The pipettings were done with precision of fraction of microliters, and those were made at a cost of $.20 to $.30 each. I don't think that is inherently a great deal more difficult than doing the operations that would result in conventional sequencing strateging today. However, you can only do that again, if you know the strategy itself.

The last couple of minutes, I just wanted to address the other theme that is running here about cooperation within California and the effects on industry -- I hope there are still a few of the political people in the audience.

As a practical matter, industry must cooperate with [inaudible]. The previous speaker referred to it. This is a brief list of those products that I've been involved with. They are DNA products. And it's a list of California academic collaborators. They're industrial collaborators. There are people that I couldn't think of, and I did this off the top of my head. There's lots of people outside of California.

But every product that we have -- DNA synthesizer, DNA sequencer, and acid extractor, and the mapping that Tony is working on -- has had academic collaboration. It's vital, it's important, and I'm glad people are encouraging it, but it needs to happen, and it really does happen. I think we're pretty good at it.

The last side is a competitiveness concern specific to California. Competitives really means good people. It doesn't necessarily just mean good researchers. It means good people in the operation at whatever level they're working.

We have a lot of advantages here in California. We have great universities. There are a lot of people, so it's easy to find good people if you have a lot of people to start with. It's a fun place to live for all sorts of reasons.

There are some serious disadvantages that are for the political people in the audience, I think, very important. We lose people because either we're not trying very hard to keep them or because they've got some wonderful new opportunity

to join a start-up company or become a massage therapist or whatever. Those people leave eagerly.

The other category of people we lose, and we lose them frequently, are people who are sad about leaving our company, and they're going some place where they figure their kids can get a better education, where they can afford to live close to where they work, which is a function either of transportation or cost of housing.

We've lost in the last six months people I'm involved with, not necessarily ones that work for me, somebody to Indianapolis, somebody to Philadelphia, two people to North Carolina, lost a potential recruit because his wife is a teacher and came out and checked out the schools in the area and decided they could educate their kids better in Georgia.

So it's a serious problem. It's a daily problem. It's one that people like myself -- I like to think of myself as inventive and clever -- but I am concerned about things like that instead of strategies in automation.

The last thing which is both a concern and an opportunity, I think, in terms of competitiveness for California is if we could encourage better public understanding of what's going on. I think that would be an advantage for the state and, indeed, for all of us. Your neighbors would have a little bit better idea of what you're doing and maybe appreciate you a little more.

DR. CARRANO: Thank you, Norm.

Our next speaker is Tim Hunkapillar. He's from Cal Tech.

MR. HUNKAPILLAR: Well, we seem to be bouncing back between what's being optimistic and pessimistic. And Norm was optmistic, and I'll say that I'm pessimistic, in the sense of how hard this problem is. And mine is particularly, at the current level, a computational sort of issue in handling the data and analyzing the data and this sort of stuff.

And one thing we can kind of keep hearing a 3 billion bases and all that sort of stuff. And that sounds -- well, it's an attractible and noble figure. But in reality I think if -- well, the issue is not just that size. The issue is the complexity of that data as well as the size of that data, which makes it a significantly different problem. And the size alone is down, forgetting the notion of say 10 to 15 which is like way in the future, by the time you have the 3 billion bases that are your human genome, more or less complete, you're talking about having an enormous number of other bases.

You're talking about -- well, if that's the 3 billion of the single strand -- you're talking about having to sequence thousands of clones, of course. You don't have any history of these clones. You don't have any record of where this comes from, the annotation of this sequence. And you have at that time also a lot of other genomes, or partial genomes, already

sequenced. You have a lot of pieces of the human genome sequenced multiple times for comparative analysis. So it's not unreasonable even at that point, which is relatively the foreseeable future, but when you have the first more or less genome of the human now, you will have approximately, say, 100 billion sort of base equivolents of information, whether that's annotation or whether that's bases or whatever.

Now, that already sort of ranks it as a large data base by anybody's standards. There are, in fact, larger data bases in the world; but, nonetheless, that's a huge amount of information.

And I would argue -- I've argued many times on this fact -- that there are problems with that because of its size, because of its complexity that, in fact, there are not tremendous sort of paradise for handling it. That means at the data base level and the analytical level. And that 100 billion bases, or base equivolents, of course, does not even count the sort of entire literature data bases and the other biological data bases, the NSF sort of ecological data base, stuff like this that all tie back into this sort of stuff where you end up getting terabytes of information eventually.

Now, nothing we have, either hardware or software or even strategies, come anywhere near approaching this problem. And there's no reason to believe we know how to do it even if we had the money right now. It's not the issue of "Oh, well,

we just need to build bigger computers or put work stations at everybody's desk." It's not that simple an issue.

It's an issue of rethinking the -- or thinking the process now as to what really are the issues. What do we need to do with this data? How do we manage it? And start thinking at a research level how to handle this new paradigm, or construct new paradigms, with this new model of data.

You have a chromosome that's 250 million bases long. It's not just 250 million a, d, c's and t's in a row. It's all the sort of landmarks. It's all the context. It's all the history of how it was sequenced. All these sort of things are all tied together.

And as a single sort of data object -- and I could be wrong, and I've had this argument with computer scientists -- but there is no real current sort of model for that. You're not going to go out into the industry and find somebody that says, "Oh, we've solved that problem. We have a personnel data base system that will do that for you." I've been told that actually. I don't believe it.

So there are issues of research involved here as defining these problems. It's not a matter of just applying. It's a matter of going to NIH, to the DOE, the NSF, or whatever and saying, "We need to do research. "

There's no reason to believe right now we know the approach. We have to take many approaches and see where they

lead. And I'm very keen on that notion that there's software approaches to handling these obvious problems and there are hardware. I have a keen interest in hardware. I'll talk to you about it in a second. I have this particular keen interest in that.

Now, this is a very collaborative sort of operation. It has to be. A lot of the sort of stuff we as biologists have used in the past up to this point are both sort of -- while hardware means that we all have our PC's and the lucky ones have a little work station on their desk, and we all have a package of software that was written more often than not by people like me, people who are biologists and thought that they could learn computers.

The level of software, generally -- that doesn't mean that there aren't exceptions to this -- but the level of software tools that are available in the world are not state-of-the-art technology. I mean regardless of the fact that you're not working on state-of-the-art hardware, you're not even -- at the level of hardware you're using, you're not generally using state-of-the-art software that can take advantage of that hardware.

So we're in pretty primitive sort of circumstances rignt now in dealing with this data, in managing it, distributing it. There are a lot of people -- a lot of effort obviously at places like GenBank and the NBL people and stuff like this

that are putting a lot of thought and concern about how to deal at the present time with the data that we have, which is an enormous -- even though it's not 100 billion bytes of data -- it's an enormous amount of data, again, given its complexity, given the fact that it's being sort of contributed by several thousand people, which is an issue in and of itself. So we've got to keep this thing of complexity in mind.

Now, I have a particular interest. I have a lot of particular interests, but the one particular interest -- I'll show a quick couple slides of this -- is not a solution. It's an approach to some of these questions. We're interested in some of the hardware sort of things that are not using, per se, the computers.

The problem with using sort of the general computer model is that the problem is growing you might say logarithmically whereas the speed of computers is not. The speed of computers may be getting a lot faster and a lot cheaper but not at the same rate the problem is growing. You're not going to keep up just by relying on the fact that next week you'll have a faster sun on your desk than the week before. It's not going to work.

Just a real quick general notion of what the problem is, one of the bottlenecks is that the notion is on a model with a normal sort of computer architectures that we deal with. If you had data coming in, the data that handled and manipulated

and everything and processed all within the [inaudible] a traditional sort of computer model, the variation on this is, in fact, is to separate out permanent computational question before it gets to the computer. It's going to do your post processing and report writing and inferencing and all this sort of stuff that preprocess the data, that you can make simple but very fast. They don't do very many things, but they do them very fast. And you can make those so that they are essentially the speed of that data transfer, not the speed of your computer.

Now, this is the approach that we are taking. This is the generic notion of how to do this now. We at Cal Tech have a few collaborations. One is with Mike Waterman's group at U.S.C. And it's a good paradigm for the sort of collaborative multi-disciplinary sort of issues that Mike and I collaborate. We collaborate with computer engineers at JPL and an eventually determined commercial partner as well. So there's four essential units in designing some specialized hardware for dynamic programming avenues that will allow you to compare and search through data bases.

Now, we also at Cal Tech, as mentioned by Bob this morning, are collaborating with TRW on an analogous-type system for explicit pattern matching which is the FDF -- TRW/FDF thing. And both of these systems, both are dynamic programming systems for the broader questions of relationships, and the

TWR system for dealing with explicit pattern matching, are based on systolic models.

Systolic means that you essentially -- you get a lot of little cells, a lot of little tiny cpu's stuck in a row, each one of them does exactly the same thing, and you could have a lot of them. And what they do, they do one calculation, more or less, and they outbreak by streaming the data, refer to it as systolic because of the pumping. You need to take one byte into the next, into the next, into the next byte, the clock cycle of the computer.

So you've got a pumping operation that essentially means that once a day it gets through the [inaudible] it's finished. There's no turning around and thinking about it. So you can do it, essentially, as fast as you can read the data. So systolics isn't some generic sort of inherited [inaudible]. It's constant speed which is independent of the complexity and the relationships that you'd find, whether it's gaps in insertions or what kind of data is going through there, all this sort of stuff.

The problem with most of this sort of either pattern matching or relational matching sort of schemes now is that you either have huge amounts of membrane or huge [inaudible] process going on that essentially blow up with the length of the data; hence, you can't do them unless you've got a crane sitting on your desk.

And the notion of massive parallelism at the bottom means that you can make these things, essentially, as big as you want to make them. You just keep putting more and more chips in a row, essentially. I'm simplifying somewhat; nevertheless, it's pretty much the model.

Now, as a paradigm for that, again, the TRW sort of stuff real quick, I'm not going to go through what this is. Well, I'm going to say that it's an actual query going through GenBank. And I can load into the FDF system that is searching things on both the top for the biology, the history, and the sequence. The sequence has motifs of both DNA and protein and composition and proximity, all this sort of stuff.

Now, I can load that in as a query, a single query. Now that means that I can do this query in the time against all of GenBank as fast as I can GenBank off of my desk. Now, I use this as a cute example; nonetheless, it has limitation. What it does now is significantly noted.

Using systolic sorting systems has tremendous promise in speeding up complex operations. It means it allows biologists to sit there and make quiries against this huge amount of data in real time. Now, of course, at 3 million bases even, if this is running at 10 million bases a second, which is a very, very fast disk, you're not going to get it much faster than that, at least for a while, you're still not worried.

I show this for entertainment value alone. I show that Bob was right this morning about where this technology came from. This was the first slide that TRW ever showed me for where this chip came from. But I use it to illustrate the point that, in fact, there's an enormous amount of technology in the world. But the notion is that there are people -- I have companies coming to our lab all the time saying, "We have technology. How can we use it?" And it's actually pretty amazing.

And this collaborative effort between industry, between computer scientists, between biologists, between mathematicians can be amazingly productive, and it's something that in California we have a tremendous amount of, so we should encourage it.

Thank you.

DR. CARRANO: Our next speaker is Elbert Branscomb from Lawrence Livermore National Laboratory. Elbert got his Ph.D. in theoretical physics. And he's responsible for our computations at the genome project at Livermore. And he's going to tell us a little bit about hardware and so forth.

DR. BRANSCOMB: About a year and half ago at the -- I think it was the second Santa Fe meeting -- Wally Gilbert gave a talk on the Human Genome Project in which he described what biology in the 21st century would be like and emphasized that it would be entirely unlike all the biology that had proceeded

it. And the picture he drew was of large groups of people sitting in front of large monitors and not getting their hands wet. And whereas that certainly is not a very pleasant prospect, I think there's no doubt that as several people have said that the biological activity is going to -- biological research is going to see a rather large shift towards people -- towards the task of analyzing a data base, which is rather than having a bunch of data collected by someone who then analyzes it and publishes it, there will be a big shift towards large efforts, industrial efforts, to obtain a large part of the fundamental data base of biology and then that will be a public resource which is analyzed.

And that then shifts towards a very heavy burden on data analysis, and it's very likely this will prove to be for the limitations that Tim was just talking about a weight-limiting step in the process.

And part of what I was going to make a few remarks about is: What might we think are the critical problems there? And I wanted to just say something about the hardware that's now available. I was trying to describe what a current good work station can do and what its storage capabilities are. And there's a reliable predicition that about a 10-fold boost in most of those properties will occur within a few years.

A typical arrangement for a well-endowed lab is quite a few of work stations like this slave to a large compute

server over [inaudible] which has essentially the powers that work stations will have four years hence. And in four years this table will flip and the compute servers will be about 10 times bigger. And it's plain that we will see limitations both in hardware and software.

I wanted to say a word about the data base. I don't disagree with the previous speakers about what they will be though, except in some details. But they are going to be very large. And the problem is not, however, in the size of the data base. There's just a number of bytes. There are really storage techniques that not very rich people can buy and can store probably all the data we will have in this project 10 years hence on a single disk of maybe 200 gigabytes or so, and that is certainly not the problem.

I think, in general, we will for most cases be more software limited than we're hardware limited. One of the things that I think that's going to happen to biology is that it will become as compute insatiable as physics. And I wanted to just mention a few areas in which I think that's going to be the case.

There are the obvious ones about -- in particular, what Tim was just talking about -- involving the sequence analysis, pattern recognition, homology searches, consensus searches and so on where we have problems at all scales. We don't really know what to do, and we have to figure out

what we should do in principle. We don't [inaudible] for doing it economically, and even when we do, good software is generally not available for reasons having to do with the economics of this problem more than the history of computational science and of the history of the industrial production of computers.

One of the problems that's clearly -- will exhaust all the computational ability for a long time, I think, is the problem of protein folding and predicting protein function from underlying sequence. And one of the things that's impressed me about why that's such a profound problem is the fact that the difference between a properly folded and unfolded protein is only a few weak bonds. And, moreover, that that appears in many cases to be the result of excluding water access to a very small number of hydrophobic residues, a handful, and that that implies extremely subtle issue about the folding properties, but we are out of our league in trying to compute that straight ahead. And it's a very profound problem which will engage us for a long time I think.

Item No. 4 is an idiosyncratic fantasy of mine and a lot of other people for sure, but in the end of -- or pretty soon after we have gone after genetic analysis, linkage analysis to get where we can in understanding the 3,000 or so most important loci in the catalog, we can then -- we will be empowered and motivated to turn our attention to the more general issues of: How does the genotype orchestrate phenotype?

And I believe it's arguable, but if we have 1 cM VNTR map, which is characterized by about 5,000 probes in order to get it dense enough to really be functional at 1 cM map, with about 70 to 80 percent average heterozygosity, which is apparently what these -- what goods the VNTR's deliver, will be in a position to -- and if we can, as I think it's arguable that we -- as I think Charles said earlier -- automate the genetic analysis using these probes, it will be feasible to go a fair way towards genotyping individuals at the 1 cM shot all together.

The fantasy is that we characterize what is the specification of people more or less completely for a complete set of 1 cM VNTR's. The advantage of that in principle has been used in some other standard genetic analysis now is that a single 1 cM locus is quite a durable piece of DNA and has, in general, a very respectable history in time. And the human population is to one degree or another more or less clonal in single alleles. So if we characterize someone as having a particular two-point allele, Lander and Botstein have emphasized the importance of recently, that particular allele will be shared completely by a lot of other people who have the same two-point allele, so that we can in some sense make a rather complete genetic characterization of individuals at that degree of resolution. And it makes it feasible, I think, to look for genotype/phenotype correlations on a population

basis rather than on a kindred basis, and be able to ask a very general question.

As Tim emphasized, I think the overpowering problem is that we don't have good software. We don't have good cheap software to do all these problems. We don't have the software tools in place, though we certainly could, to fulfill Charles' fantasy of having a reasonable, friendly, appropriate and individualized work station which would be -- have common interfaces and be recognizable and talk to everyone and run on all sorts of -- talk to everyone in the same language and run on all sorts of hardware.

And the critical issue, it seems to me, and an issue where state policy and governmental policy have a very big role is in trying to enforce and further the development of standards which will overcome these problems. And the standards matter at a large number of levels in computer design. And I just tried to indicate a few of the dominant ones here.

There's the operating system, the window environment, what the windows look like and how do they behave. You have to sit there and learn a whole new window paradigm every time you sit down at a computer. What's the visual interface software that talks through the window? What are the graphics conventions for drawing? There's very little agreement there. It's really a Tower of Babel as someone said before. What are the network conventions?

And the last item is one that I think is particularly important and can drive an awful lot of economy here, and that is the prospect of having so-called applications by or near interfaces, which means that you don't have to recompile software to run on one computer or another, but it will run straight across.

And there's the prospect now, but stimulated by some's aggression in this field in large part that there will be not one, that's too much to hope for, but maybe five or six such groups of computers that at least within those groups you won't have to cross-compile. And the big significance of that is not, I think, just that it allows you to be vendor independent so that you can compete vendors against each other, which is important enough, but rather that it would make it economic, much more economic, for people to build software because now it will be -- there will be many, many more customers. And I think that can help us a great deal in getting good software for this problem. We are too small a user audience to justify enough software development for our problems, particularly if we split ourselves up into lots of groups.

DR. CARRANO: Thank you, Elbert, catch your breath.

Our last speaker is Michael Waterman. He's a professor of mathematics in biological sciences at the University of Southern California.

DR. WATERMAN: I'm certainly here to speak as a

mathematical scientist. And I'm glad that my education wasn't in theoretical physics because I was just told that I have negative time.

This is supposed to be about hardware and software. And I have a really hard time talking about hardware and software when we don't talk about why we need the hardware and the software. And so let me -- maybe this is a bridge between this session and the next session. So let me just run over a little bit of what I think of what the other people have said.

Most of what has been talked about today, I think, is data generation. People have talked about mapping, both genetic and physical mapping, and they talked about some of the analytical problems entailed therein but not too much. And the reason the analytical problems are interesting, I think, is to get accurate maps in the genetic case and just to decide whether you're going to have your postdoc's working five years or 50 years on the physical mapping scheme. And those are worthwhile analyses to perform.

I think as well there are certainly some hardware/software issues in reading your classical or automated sequencing projects. And that's definitely been talked about.

Tim talked something about after the data is generated how we're going to manage it. He said there was certainly some research required there. And I think there's a point I'd like

to make about that.

There's a nucleotide data base as we see it today is called GenBank or NBL's data base, which are almost but not quite the same, and there's a model of sequence that's been published, that's been annotated, at least in the literature, and is sort of a final product.

I think when we start cruising through genomic sequences, an entirely different model of how to manage the data -- that just hasn't been thought through. And the management and the access and how we correct stuff after it's out there, whatever percent wrong, is a really interesting problem that people like Cantor are going to have to face.

Well, after you've got all this data, I think there's a new piece of action that certainly wasn't around 15 years ago, and that is that you take a look at the sequence and you try to find biological and meaningful patterns and chain. We talked about that sort of thing.

There's one kind of problem -- where's the mathematics or statistics or computer science in this? There's one kind of problem, if you've got eyes like mine and it's after dark and you're trying to find your way to a specific address in Los Angeles and you pull out the Thomas Guide, it's hard work. Finding your way to an address even when you know, even when you have a map and you know where you are going is hard. And that was the Moslim slide that Tim showed basically. It's

some idea of what they were looking for.

What is a lot harder is to find your way around when you don't have a map and you don't know what you're looking for. Or perhaps a better analogy, if you have a map of Los Angeles and a map of New York, what parts look a lot alike? Again, you don't know what look alike means. So that's a hard problem. And one of the most exciting versions of it, I think, is predicting protein structure from sequence there. It's a version of that.

I think there are a couple of approaches to these kinds of problems. One is to buy a bigger computer. And NIH does this, buy a great big new Cray and run the old ideas on it. Tim and I are involved in that with generated chips. It's a lower-scale technology thing, thank goodness, but just new hardware, same old ideas. And, obviously, what I'm [inaudible] is getting people to have new ideas.

One of the things as a mathematical scientist interested in biology, one of the things that I'm impressed by is how smart biologists are at analyzing their own data. Now, that doesn't surprise any biologists, of course. But you can learn a lot about this, but it requires in the end people who know about both sides of the fence, and there aren't very many of these people around, I don't think.

And that's sort of my pitch today is that we need -- no one has mentioned education today I don't think,

and there are lots of universities here. We need to train some people who are trained to think in a nontrivial way both about an analytical side of things and about the biological side of things. And that's -- and I don't mean just two courses in each one, some real training. And I think that's a challenge for our universities to come up with these people because they'll be an important part of biology in the end of this and the next century.

Thank you.

DR. CARRANO: Thank you very much.

I'm going to defer to our moderator here, Winston.

DR. SALSER: Well, let's take a few questions and then a short break.

MR. KISSLER: I just wanted to make a statement. I'm Jerry Kissler.

While this session is going on, there's also another session going on on light sources. Light sources are another technology that might be important to this project. There is a group of people representing basically the same universities and labs that are here working on light sources.

So for any of you who might be interested in compact light sources, sort of room-size light sources, at very short close length, there will be a green paper in the back of the room. You might want to check it out.

DR. CARRANO: Anything else before we turn this back?

MR. BRENNEN: Thomas Brennen, Genomyx.

One of the things that -- is there a unit underway, anything sort of like a biological computer? When you clone a filter, or something like that, you're actually performing a [inaudible] There's a lot of stuff -- biology on silicon, is there anything -- in other words, as opposed to the sequencing part of it where we have [inaudible]. This is an area where we really need good new technology in the software. Is there anything like that in progress?

DR. CARRANO: Who wants to tackle that question?

Tom.

MR. MARR: Tom Marr, Los Alamos.

There's a group at MIT who has been working on biological materials and how those apply to computing problems. I think they've actually built some prototypes.

There's a fellow named Arvin who is a computer scientist whose been working on data flow problems, architectures, who is looking at biological materials as an avenue for implementation.

MR. BRENNEN: What I meant was actually using biology as [inaudible] probe on any kind of an actual array.

[At this point, the question went on, but the reporter was unable to hear the question.]

DR. SALSER: We're going to take a brief five-minute break to stretch your legs, and be back here at a quarter till.

Recess 3:40 p.m. to 3:53 p.m.

DR. SALSER: Before I start the next session, I want to just summarize something from the last session that I was interested in.

I cornered Tony Carrano, and he has just taken off to make his plane, so he won't be able to say it. But I was interested in his project of doing the cosmid overlaps. It wasn't clear to me just how far along it was and what the capabilities of the method were as they see it at this point.

And he says that they're starting a project on the Chromosome 19 cosmid library that they think that at the rate they're going that they could cover it to a depth of 1 in about a year. That means that after about a year's sorting on the ABI machine of the restriction digest about two-thirds of the clones they would have analyzed would have an overlap with another member of the set and that about roughly 1 over E of the sequence would be uncovered.

And so in about six months, I guess, he'll be able to tell us whether that's working out as they planned. But he thinks that they can do it on whole chromosomes. And that's where they're at.

And that is so important, either that or the YAC approach, that I thought it was really useful to summarize the status of that.

Now I want to introduce the next panel which will be moderated by Elizabeth Neufeld. It turns out that having a Ph.D. from Berkeley was an absolute requirement apparantly for all of today's panel moderators. And I asked Paul Boyer why that was. And he said that was because Berkeley guarded the education game earlier than UCLA, not to worry.

Elizabeth has been one of the leaders in studying the molecular genetics lysosomal storage diseases. She served as Chief of the Genetics and Biochemistry Branch of the National Institute of Arthritis, Diabetes, Digestive and Kidney diseases up to 1984. And at that point she got tired of all those names and came to UCLA as Chair of the Department of Biological Chemistry, which is easier to remember.

Her research accomplishments have been acknowledged by a host of scientific awards, which she told me not to go through because it would embarrass her. But in 1977 she was elected to the National Academy of Sciences in recognition of some of her work.

Dr. Neufeld will introduce her panel which will discuss utilization of the data generated by the Genome Project.

DR. NEUFELD: Thank you, Winston.

Before I introduce the panelists, I would like to introduce Ellen Philhower who is the stenographer for this meeting. And she's been assiduously taking notes. And she

has a very special request of the speakers and the panelists. If you have any notes, any transparencies, she would like a copy thereof. She says that it will make her job much easier.

DR. SALSER: Can I add to that? If you have copies but you don't want to give them up to her, give them to me and I will xerox them immediately. There are xerox machines all around, and you can have your copies and leave them too.

DR. NEUFELD: Since we are the last panel, we thought that there might be some issues which have not been touched during the day. And Dr. Simpson, who is a Professor of Biology at UCLA is going to touch on one such issue, which is: What happens when your gene product is not really included in your DNA?

DR. SIMPSON: Well, most of my thoughts on this subject have already been mentioned. So I guess I'll go directly to my one unique thought on this, and that is that I'd like to throw out a cautionary note as to the difficulty in interpretation of raw sequence data that will be generated, the huge amount of raw sequence data, that will be generated in this project.

In fact, I go along with Tim in his thinking that the most practical and the most valuable aspect of this project will come as a result of the effort required to improve the computational aspects. I think the most immedate spinoff will be the development of new Algorisms to analyze this maximum

amount of new sequence data and recognize patterns.

And in line with that, I think that the most valuable information from the scientific point of view will come from the initial sequencing projects involving the E. coli yeast and Drosophila genomes because there we have a large background of genetic information. We can do experimental testing of hypotheses, which you can't do with a human.

What can we learn from the human data, the human sequence? Well, we can look for open reading frames. We can identify known genes by similarity or hemology with genes already in the data base from lower organisms. We can perhaps identify unknown genes that have been previously mapped to precise sites or by prediction of secondary structure or a function of the proteins.

However, in the absence of an experimental genetics approach such inferences are going to be very difficult to make and substantiate.

I'd like to add one additional cautionary note stemming from work in my own laboratory, and that is that we either found that in some cases in messenger RNA's in the mitochondria of certain types of protozoa, known as tropatazomes, the nucleotide sequence information encoded in several other messenger RNA's and transcribed from the mitochondria DNA differs from that encoded in the mitochondria DNA.

Let me just show this slide. This diagram -- this

cartoon just shows what I'm talking about for the nonspecialists here. We have a DNA sequence and we have the messenger RNA sequence. And what we found is that in the mature messenger RNA sequence we have additional nucleotides, which in our case are all uridines, at specific locations in coding regions which are not present in DNA sequence. And we have uridines there present as thigmodines in the DNA sequence that are not present here.

The next slide is just an example of this for one particular gene in the mitochondria. And here is from two different species of these protozoa. Here is the DNA sequence, and here is the RNA sequence. And what you see is that dots indicate uridines that we found in the RNA sequence that are not present in the DNA sequence. And these circles indicate thigmadines present in the DNA sequence that are not present in the RNA sequence.

And this is conserved, at least the pattern of the addition of uridines, conserved in these two species that differ by about 100 million years of evolutionary history but the pattern of deletions differs, and they both give rise to exactly -- almost the same amino acid sequences.

The next slide -- this just shows how common this is. This just compares the two complete mitochondria genomes and sequences of which we know. And I just want to point out that this occurs in five genes in this mitrochondria genome.

It's not an isolated event with one or two nucleotides. It occurs in this gene. There's three here. There's 29 U's added, 15 U's deleted. It occurs here, here, here, and here with the indicated number of deletions and additions. And in [inaudible] it occurs in a very dramatic way in this region here, and now we know this region here, where of 731 nucleotide sequence, 407 are resulting from additional uridines at 145 sites and 19 uridines deleted.

So this just shows you -- this is the work of Gene Fagan and Ken Stewart, by the way, this last piece of information.

So we don't yet understand the mechanism for this, apparently, nontemplated, non-normally-templated sequence information, and we don't know the biological generality of this. But I just want to throw it out to say that you may not always be able to go directly from a DNA sequence to a protein sequence. One should keep this in mind.

DR. NEUFELD: Does anybody want to comment on this rather important cautionary note?

[No response.]

If not, we'll go on to Dr. Esposito who is from the Human Genome Center and is Deputy Director of the Cell and Molecular Biology Division at Lawrence Laboratory in Berkeley. And he'll talk about the applications of biotechnology to complex human genetic disorders.

DR. ESPOSITO: Because the hour is late and time is

short, and I think most of us have received perhaps more information than we can hopefully deal with, let me assure you that what I want to do is not to talk about the arcane aspects of human genetics and population genetics as it applies to complex human disorders but rather to focus on an issue that arises when we consider the fact that many of the human genetic disorders that are of interest to us are very difficult to study genetically.

But as we see the human genome mapping and sequencing project perceived, there is an enormous target of opportunity that develops for California-based biotechnology and pharmaceutical industries, and that is to begin a dialogue, and if I could have the next slide.

To begin a dialogue, to start an intent at applying the information to the vast array of very complex human genetic disorders that -- as well as those that are attractable.

Now, most of the applications that we've heard about today in mapping refer to wise choices of known simple Mendelian genetic defects in which almost anecdotal analysis of kindreds and pedigrees can lead you to the conclusion that this is a single gene defect. That is a very wise choice on the part of these investigators.

But we know that -- and not only will we be able to proceed to analyze such diseases very carefully, but there are many diseases that are extremely frequent as far as

diseases are concerned.

There are many very important human genetic diseases for which we have information that these genes are highly complex, complex in the following sense: There are probably more than one genetic locus involved.

Two. Complex in the sense that if you have the disease genotype, you don't necessarily get the disease, limited penetrance. And they include a collection of diseases that I've simply broken down as those associated with the human major HLA. They include insulin dependent diabetes, rheumatoid arthritis, multiple sclerosis, and some of them which are complex multi-genic unknown degrees of penetrance -- alzheimer's, epilepsy, asthma, various cancers, schizophrenia, non-insulin dependent diabetes.

But the interesting thing that has happened in the struggle to understanding complex genetic diseases is that a mathematical apparatus for dealing with such diseases emerged here because in the region of the HLA locus in humans and in other organisms, there is a tremendous degree of polymorphism, genetic polymorphism, which provided markers.

So there is a population out there of human geneticists who have the mathematical apparatus in hand. And this is part of the Branscomb fantasy, and I was glad that he mentioned it. But I think it's more than a fantasy. I think it's a very important point to note that we are on the threshold

of making an interesting combination between the mathematical apparatus that has evolved to deal with these diseases with the limited amount of genetic information that was available in the emerging map.

They are refractory because, as I mentioned, there are multiple disease loci involved. There are many disease loci involved. And when you have the genotype, you don't necessarily have the disease. This makes these diseases difficult to localize.

But with the apparatus in hand, that has allowed us to sort out the number of loci involved, the degree of heterogeneity, and make [inaudible] of penetrance; that together with the genetic map will allow us to proceed to other parts of the genome, all of those non-HLA associated diseases which are among the most common human genetic disorders and the most costly to society.

And it would seem appropriate that this organization, this loose affiliation or consortium of interested people should reach out to that community and put together some thoughts about how California-based biotechnology could exploit that development.

Thank you.

DR. NEUFELD: Thank you very much.

Our next speaker is Dr. Michael Kelly who is President of Intelligenetics and will tell us about some computational

aspects.

DR. KELLY: I just wanted to say that some of the previous speakers depressed me a little bit about the future of computational abilities for the human genome. But I'd like to be more like the gentleman from ABI and say that we think that there are a few clouds on the horizon, but for the most part I think the sun is shining as far as computational capabilities is concerned for the near future.

Today I'd like to talk a little bit about the information generated by the human genome and also the computational needs for the human genome. We'd like to talk about that in terms of not about the restriction mapping because I think that was talked about earlier in detail, but more about things such as fragment assembly, the data base hardware and software needs and sequence analysis and specialized data bases that have been generated for the human genome.

In terms of the computational needs, we have to manage the genetic and the physical maps. We have to manage the cosmids and the YAC's. We need management and analysis of all the sequences that are generated. We need to have a network communication available worldwide for all the researchers to be able to instantaneously communicate, and we need to analyze the genome and also to look at inter and intrachromosomal interactions.

In terms of fragment assembly, there are a number of commercial packages available, and they are quite adequate at analyzing and managing the sequencing of the individual fragments. We have, as was told to us earlier, we're up to 500 base pair fragments. Now how do we overlap these fragments?

Well, to give you an example of this, and that is the program GEL. It allows you to display and eliminate the vector sequences so that you can eliminate any of the cosmid sequences included in the sequencing, and also you can automatically merge these functions.

Surely Intelligenetics will have a program called the SymGel out on the PC which will be able to handle over 100,000 base pairs. And there is no limitation for the future. It's only a limitation of the size of the hardware in this particular case. We have a new Algorisms that allows us to do millions of base pairs.

In terms of the generation of all these sequences, I'd like to give you some perspective on the growth of GenBank. This is through March of '88. And you can see that it at that particular time, it was just below 20 million nucleotides, and growing exponentially.

So when Intelligenetics was awarded the contract for GenBank, we did an analysis to decide exactly what will be the needs for the next five years since this is a five-year contract. And on that basis, we started out in 1987 and there

were 15 million nucleotides in the data base at that particular time. By 1992 -- and I think this is a little optimistic -- that there will be 250 million nucleotides in the data base, but that depends on how fast Dr. Cantor's group and other laboratories move in terms of entering sequences.

But right now these are the estimates that we have. We will have the capability -- the computer capability -- of the storage of all this data in a relational format, and I'll tell you a little bit more about that later. And we estimate that it will only require about two gigabytes at that time.

There was a previous slide that I think either Tim Hunkapillar or Mike Waterman had that said -- or Elbert Branscomb -- that we had -- we need about 1 to 10 gigabytes for the whole human genome. But there is currently hardware available and software available to access that in a fairly reasonable way.

In terms of rapid retrieval, we need to develop a sophisticated network communication with easy access to a PC host-type of system. And that currently does not exist. There is the NSF map which is being developed and will have a fairly wide band width up to 1 megabyte per second, but not every researcher will have access to that currently.

But this is an area where, I think, the NSF and other organizations can take a step in furthering this particular network so that we can access quite rapidly the

information generated.

There are natural language communication devices being generated, both at Intelligenetics and also at Los Alamos among other laboratories for the sequence entry and manipulation. And I'll tell you a little bit more about the software for entry of the data in a little bit.

In terms of distribution of the data bases, currently it's released on magnetic tape. And right now it takes about two or three reels of magnetic tape to do the whole GenBank release. In terms of 1.2 megabyte floppies in compressed format, it's about 19 1.2 megabyte floppies, and that does not include all the annotation. It's actually about 80 360k floppies at the current release.

So we have a problem for people using PC's right now. Their arms get tired entering the data. So what GenBank is going to be doing fairly soon is developing a CD ROM. It will be a single entity because we have high-storage density. We can store actually close to 500 megabytes, but we said 300 in here because we're thinking about adding other data bases to the format as well.

We are currently under negotiations with PIR and other data bases to try to develop a single system so that it makes it easier for the researcher, but that will require some more negotiations.

So this is ideal for the current needs. A CD ROM

costs about $1,000, and it's quite easily accessible from a PC. There are some problems with that, however, because of access times. And so as was stated earlier, the access time of this information is dependent upon the time of the type of search done, whether you're doing a straight match, a pattern expansion, a local or a global alignment. It depends on the type of CPU and the load on that CPU. So if you've got a large VAC's and you have 15 or 20 people on it all doing searches, you can grind it to a halt, and it depends on the structure of the data base. Is it indexed so that you can easily access the portion of the data base that you want to look at and it also depends on the storage medium. And in this case the CD ROM is a slightly slower mechanism of access of the data.

But if you want to do an overnight search or something like that, you can do that. But I would like to second Tim Hunkapillar's plea that we have more research done on things like parallel processing and also Mike Waterman's suggestion that we have more education in the area of computational molecular biology.

Now, what about the future directions of the data base? It needs to become more current. When we started the contract, the GenBank contract, in October of '87 the data -- we estimate the data for which GenBank is responsible, not for which European Molecular Biology Laboratory is responsible,

and GenBank's portion of the overall nucleotide data base represents about two-thirds of the total. GenBank was somewhere in the range of 80 percent of the data would be in within 12 to 14 months of publication. And this was totally inadequate. And we recognized that.

And there were steps taken by the NIH to improve this by providing more resources to allow us to become more current. And as of the last release, we believe that the data -- 80 percent of the data is getting in within five months. So there's been a drastic improvement in the currency of the data. And we expect that trend to continue and become much more up to date.

We want to organize this for more efficient use. We will be doing new indexing routines to allow the researchers to access it more rapidly. We want to develop links to other data bases. And in terms of this, a relational data base structure is being generated mostly through the group at Los Alamos. And they are working very closely also with the HGML data base, and so that the relational structure will allow tables and relational linkages to all of the human gene map.

Also, we are developing an author entry set of software. And as Elbert Branscomb was talking about, we need to have the same interface on most computers. And we are working towards that goal so that when the individual researcher

wants to enter his sequence and annotation information that you will be able to do it with the same format in every type of computer. This author entry software will be developed for both -- it will be on the PC first, and it will be available in the Spring of '89. Also, it will be developed on the Macintosh, on the Sun and the VAC systems as we can get that done.

And this also, in terms of data bases, is spawning specialized entities that are just coming into the four, and some of these are the cytogenetic map data base, the RFLP map data bases. We already have the Brookhaven Coordinate data base. VectorBank is a data base of all the plasmids. Roberts' Restriction Enzyme data base, and what we call a Signature Region data base.

In terms of analyzing sequences, as of March there were over 19,000 DNA sequences and over 6,000 protein sequences. And we ought to look at them in terms of their structural and their functional properties.

In terms of sequence patterns, we're looking at groups of bases or amino acids associated with a particular function or aspect. Examples of this, of course, are the tata box and consensus sequences for DNA or proteins.

Key Bank is a data base of known sequence patterns. And this has been generated by Intelligenetics as reported in the scientific literature. And currently in Key Bank we

have over 1300 entries of known structure function-type relationships. These entries are all fully referenced. And they're classified into categories and then indexed. And we currently have 16 categories of protein keys and 112 categories of nucleic acid keys.

Examples of these categories are amino acid modifications, signal sites, binding sites, enzyme active and allosteric sites, regulatory regions, binding sites and repeats and ends.

Key Bank was designed for use with the Quest program. Quest searches the sequence files for key words and sequence patterns. It searches the single sequence or many sequences. It retrieves sequences while searching. It allows use of ambiguities and boolean operators in the patterns and allows patterns to be combined to form complex patterns.

There's another program on the PC which does similar things for the protein. It's called Procyte. And here is an example of actually delineating. And currently we have something like 120 individual signatures or patterns that can be delineated on proteins. And this will be coming in the next version of PC Gene.

In summary, we think the Human Genome Project will generate vast numbers of large sequences, cosmids, YAC's and restriction maps. The data base growth will generate needs for better distribution, storage, access times and use. And we think that new computational tools as well as the existing

computational tools will be needed to analyze this data.

DR. NEUFELD: Thank you very much.

I think we'll go on to Dr. Marr who comes from Los Alamos Laboratories. And we'll continue this discussion on computational aspects.

DR. MARR: Unless I have to, I think I'll go ahead and stand here and use what few relevant slides I have left myself.

Not knowing who was going to be here or what was going to be discussed, I sort of naively put this set of overheads together. So I think it's clear that I shouldn't bother going over some of the things that have been discussed here already.

Let me just put up one thing that hasn't been clearly pointed out to the group, and that is some of the computational complexities involved with some of the problems that have been discussed here today.

If N represents the length of the sequence, and complexity is some index of -- well, in this case -- the time to compute, being proportional to -- in the case of similarity searching -- the length of the sequence squared. We start looking at some of the real interesting problems that confront us in going over the examining the DNA sequences.

We see that the problem gets complicated very quickly as we move from the relatively simple problem of similarity

searching, which is the thing that GenBank is used for on a daily basis. Somebody sequenced a piece of DNA, and they want to see what it looks like so they go through and apply some type of search Algorisms to GenBank, and it shows up to be similar to some other pattern that exists in the data base.

Well, as we start looking at predicting RNA secondary structure from primary sequence, we see that complexity or time to compute goes up as the cue of the length of the sequence and to the where the most interesting, in my opinion, problems is predicting tertiary structure from primary sequence data. And I don't know what that really scales at, but it's apparently some large number. Right now it takes one cray, one hour, to do this for a tetramer.

And if you think on the average -- what's that?

DR. SALSER: Tetramer of what?

DR. MARR: If you take primary sequence and you do a good calculation to come up with tertiary structure, it takes about one cray, one hour. There are simpler ways to do it on PC's, et cetera, but they're not believable structures they end up with.

This is using a Monte Carlo approach to energy minimization in the structure. So if you think -- if you want to do this over a larger piece of DNA, it gets up into many thousands of cray one hours very quickly.

So that's some more numerical view of the complexity

that we're confronted with.

It has come up that what the speakers are doing is viewing this problem from different backgrounds. My background is in evolutionary biology and system science. And the challenge that this is really -- or the new things that the DNA technology is opening up to evolutionary biologists, of course, is being able to test hypotheses that we've made glaring assumptions about for many years, that is evolution within populations.

Now we can measure gene frequencies, allelic frequencies, in populations by looking at distributions of Jeffry's-type probes in populations. We're not involved in doing that. We can apply the DNA technology to very practical wildlife management problems. So we can go in and take blood from whooping cranes and tell who is related to whom. And, therefore, if we're going to be doing artificial insemination studies, we shouldn't be pairing brothers and sisters in that situation.

So I think the technology in a broader sense opens up a whole range of new problems to other types of scientists. And that's what I think is exciting about potential spinoffs.

DR. NEUFELD: Thank you very much.

Our last speaker is going to address an issue which I think has been omitted from much of today's discussions; namely, where are we going to get the manpower to do all this computing and biology?

Dr. Steven Dahms is a Professor of Chemistry at San Diego State University, and he's also the Director of the California State University Program For Education Research in Biotechnology, which is a nice acronym of C-SUPERB.

DR. DAHMS: I'll try and keep my comments brief. This topic doesn't address how or with what the initiative will be conducted, but by whom. It relates directly to biotechnology training programs, both in predoctoral, postdoctoral levels as well as below.

My initial comments will temporarily disregard the mapping and sequencing initiative and will focus instead on current manpower demands and projections in California and training programs in the biotech arena.

I do this largely because the new and existing programs have not taken into account the potential initiative to any extent.

My initial comments will focus upon the NIH and California. NIH has recently addressed the problem. It was challenged in the Fall of '87 by the U.S. Senate largely to the action of Senator Chiles to assess what the NIH was doing and the biotech arena that involved biotech research and training.

It charged the NIH with supplying an internal analysis to the Senate by the end of January and also with establishing an advisory panel to establish the needs for

new training programs and to make recommendations therefore.

The report was submitted to Wyngaarden on February 1st, by Wyngaarden on February 1st to the Senate. And a panel was appointed and met at the end of March for several day. It focused upon specific areas to be developed in the biotech arena on all levels of instruction. I won't go into that. It's a matter of public record. There was a considerable discussion on manpower needs on the Ph.D., Masters, Bachelor's level and the Associate Arts' degree level as well from the 11 academic panel members and the six corporate members.

Considerable focus was on the trends of: What's happening in the next 10 years with regard to faculty replacements in the United States in the sciences?

A good example of the problem that exists now is in California is: What will be happening within the California State University system in the next 10 years where between 40 and 60 percent of all faculty will be turning over? With about 20,000 faculty in the system, this means that in the next 10 years about 750 faculty will be hired in chemistry and biology. At the same time the exponential hirings will be occurring in the corporate biotech arena.

Well, finally the report was submitted first to the Senate and eventually it was incorporated into Chiles and was a bill that was entitled "Biotech Competitiveness" and resulted in $2.7 in new money to NIH. It was supposed to be

$5 million, but NIH didn't separate it correctly. It's resulted in 150 new predoctoral trainingships in the biotech arena and a number of postdoctoral and some in the M.D. side as well.

A lot of the discussion focused on the predoctoral level. And I think in the next year the NIH might be reinterpreting the word predoctoral to mean pro pre level of training programs in the biotech arena below the Ph.D. granting area.

Moving onto California in particular. With about 20 percent of corporate U.S. biotechnology in San Diego our all Molecular Biology Institute and the State Biotech Program have conducted a survey in that area on the needs for the next five years. You might be surprised that these translate to 800 Bachelor's, 300 Master's and 200 Ph.D.'s in the next five years.

A recent committee composed of Salk/Scripps' corporate CEO's have in turn projected a need for at least 200 Associate of Arts degree level individuals for San Diego. And this is a new phenomenon on the West Coast, and it appeared to be popular on the East Coast. So I what I think you'll be seeing in California, more and more junior colleges moving into so-called biotechnology training positions to conduct some of these activities.

We have not completed the survey in the Bay Area,

but the initial figures agree with those in San Diego. It's clear that we're going to have difficulty meeting those manpower needs between UCSD and San Diego State, and major problems exist.

The impact of the sequencing initiative has not been addressed in any surveys. It's important to realize that not only are manpower needs to be considered for the Ph.D and Masters' and Bachelor's level and below for the initiative but for life thereafter in the life sciences, that is thereafter the sequence initiative is completed.

I think once Dr. Salser gave projections of 4500 man years. If you take the suggestion by Norman Whiteley earlier today that assuming a dollar were a base pair, and three billion base pairs, and 8 to 10 man years per $1 million expense, this translates to about 30,000 man years. If you want to do this over a five-year period, that means 6,000 personnel, and that's at 3 billion base pairs, not a 100 billion base pairs, the panel has not evaluated any data of this type to consider where it should be going. All NIH has done so far is to project a program within five years that will have 1500 Ph.D. trainees in the works.

Right now this first year there will be 150. Elementation will take place next year for another 150, plus 300 more. So it will be building.

The 6,000 personnel needed nationally, very conservative

estimates, assuming 10 percent goes to California, that means roughly a 25 percent increase in demand projected in the next five years. What the real figure is nobody knows. It could be 10-fold higher due to spinoff technologies and related endeavors. What this translates to in terms of the number of Ph.D.'s and non-Ph.D.'s isn't clear; but, generally, it's about a 5 to 1 ratio of non-Ph.D.'s to Ph.D.'s in this area.

According to the evaluation, the impact upon the manpower pool needs to be conducted in California with about 360,000 students in the California State University system and 140 in the U.C., together there's about a 20 percent coverage of higher education in the United States. It might be wise for a joint CSU/UC committee to evaluate manpower needs in defining new facilities and new training programs.

CSU is contributing quite significantly to meet these manpower needs through a modest number of Ph.D.'s that are produced, but in particular to a large number of Masters' and Bachelor's level people that are brought through strong research programs. There are a number of novel alternate training programs, non-degree, that in specific biotechnologies have been established. They involved postdoc's, technicians and other individuals.

Biotech training programs at UCSU are in the exponential

phase resulting in new research programs being established and training programs, but these, I emphasize, have been conducted without new state monies and have impacted existing programs and facilities as much as it has clearly within the UC.

I guess there are several take-home lessons. In California we're having severe difficulty meeting manpower needs now. What will be able to do with the initiative? Probably not much unless we take action soon.

Secondly, NIH has not considered the sequencing initiative. It needs to. And it has been recommended to the advisory panel that this be discussed as soon as possible.

Thirdly, a number of other federal agencies are initiating new biotechnology training programs. And it's clear that there's not much communication taking place between the agencies, and this should be taking place.

Lastly, there was a quotation earlier today, something to the effect that "I hope our aspirations don't exceed our abilities." I'd modify that somewhat to say, "I hope our aspirations in the manpower arena are not greater than our abilities to provide it."

Thank you.

DR. NEUFELD: Are there questions from the audience?

DR. SALSER: I have a question for Dr. Kelly, and that is: Are you planning to support the next computer or

will its unit system be compatible with some software package you're already supplying?

DR. KELLY: We are not currently planning on supporting the next computer. We are waiting to see when it will be viable as a system.

DR. SALSER: Do you have reservations about that?

DR. KELLY: Yes, only since we've only seen the initial stages of it. And we're waiting to see whether it will be a viable system.

DR. BOYER: I have a question for Dr. Dahms. If UCLA -- one of the depressing statistics is the fraction of the entering freshman class that indicate an interest in science and chemistry and biology as a major. This is down dramatically from what it was a decade ago. There's only, I think, something about half or 60 percent or as many indicating this interest.

I wonder if that's the same trend you have in the state college system. A much smaller fraction of our entering class of freshman is indicating an interest in science as a major.

DR. DAHMS: I think it's a national phenomenon. It's clearly been a discussion on the NIH advisory panel level. There's even been discussion of NIH putting money into high school instructor training programs so that they can affect the quality and numbers of students who elect careers in

science. It's a subject of discussion among a number of professional societies. I think it's a national phenomenon that the NSF has not sufficiently addressed. It is the agency that should do it.

DR. AVDALOVIC: I have a question and a suggestion which I didn't have a chance to make during my presentation.

We all realize that there is a big problem --

UNIDENTIFIED SPEAKER: You probably should move that up somewhat.

DR. AVDALOVIC: I didn't want to obstruct you with the first question.

-- is to promote some postdoctoral projects with the interdisciplinary education with emphasis on computer applications because we all see the need for a cross hybridization between molecular biology and computer sciences.

What we have here is more a lack of good and fresh ideas and approaches than a lack of computational speed. So that could be perhaps one of the proposals for interdisciplinary postdoctoral programs.

The question which I had is for Dr. Kelly. Last month at the human genome meeting in Valencia, Rich Roberts from Cold Spring Harbor raised a very interesting question asking: Are there any efforts in the arena of literature on computer? In other words, all these efforts in human genome might be perhaps faster expedited if we had the access to the

real [inaudible] through the computer to access to the data like structural analysis, crystalography data and so on. Are there any efforts in that area? Do you think that is something feasible?

DR. KELLY: Could you rephrase that question? I'm not quite sure that I understand.

DR. AVDALOVIC: Access to the real articles. When you retrieve the sequence you get a sequence, but you would like to know what that sequence does. This is translated into protein somewhere.

DR. KELLY: So the complete article in an electronic form?

DR. AVDALOVIC: Yes.

DR. KELLY: I think there are some indications that people want to put journals on CU ROM. The access to that information only would occur through programs such as Quest or other programs such as that where you do key word searches currently.

There needs to be more research in that area. And I don't think there is a strong effort to do that. On Bionet we are not starting to put the abstracts of articles on the system. So if individual journals want to have an electronic abstract system, Bionet will be a vehicle for them to look at the abstracts electronically.

DR. MARR: I might be able to say something about

that as well.

We've been talking to Dennis Benson at the National Library of Medicine. Right now we can take GenBank accession numbers and pull up abstracts out of MedLine from the medical literature. It's just a question of dollars. If somebody puts significant -- a sufficient amount of money into the National Library of Medicine for them to digitize the primary literature, then the technology is there to do it. It may be a little slow right now, but certainly within the next few years high-speed communications networks will be available to support full text retrieval in pointers across network data bases.

So it's just to point out to funding agencies the requirement for that.

DR. KELLY: Along with that, there is an experimental program that NLM has generated called IRX that will do a sophisticated type of key word search.

DR. DAHMS: Could I comment on your statement about the postdoctoral training programs?

The new program that has just come on line within the past month, the first deadline for application, I think, is January 1. It would cover that clearly. It's looking for industry interactions in the predoctoral and postdoctoral level.

And if I could provide another comment to Dr. Boyer.

The most recent publications in the National Academy of Science indicate that students make their decision on entering a career in science usually at the latest by the 10th grade. So there is a concerted effort that has to be made to influence students to elect science as a career. And there are some activities within the state now in the molecular life science arena that are doing this.

There's a collaborative program between San Francisco State and Santa Cruz that is bringing high school students through DNA training programs in the summer in an attempt to educate them and affect them to the point of increasing the number of students electing science as a career.

DR. NEUFELD: Any other comments?

Larry.

DR. SIMPSON: I'd like to make a comment about commercially available software for analysis of sequences. I realize that by the time it gets to be commercially available, it's certainly not state-of-the-art, but it's what is available to most people out there in the field.

And I'd just like to say that most of the packages are foreign sequence unfriendly. In other words, I guess I understand why, but they try to be self-sufficient and self-contained whereas the user out there, the noncomputer expert who just wants to use and understand the sequences, has to use the best things of this package, the best parts of

that package, and the best parts of that package. So should be easily able to switch from package to package. And part of this problem is simple sequence format problems.

I'd just like to emphasize that this is a major problem for the user.

DR. NEUFELD: Thank you very much.

DR. SALSER: Well, I thought that the best way to wrap this up would be for me to twist Charles Cantor's arm to give us his overview. And he has graciously agreed with little notice to do that. And so I'm going to turn this over to him.

DR. CANTOR: As if you haven't heard enough from me already. What I thought I would do is spend, I hope, no more than five minutes just summing up what I think are the main things that were said today.

The title is the Human Genome Projects. And I would contend that one of the things that you heard today, at least within the United States, is that there is only one project, that DOE and NIH seem very likely to be sufficiently coordinated so that there really is no discord between them. It remains to be seen whether this coordination can be spread internationally. But I'm optimistic about what's happening within the U.S.

Issues. I think there really are two major issues that we have to keep in mind. The first, and I think the

major concern with the biological community that led originally to a lot of opposition to the project was the fear that it would sap resources on traditional biological research. I think that this simply hasn't happened.

And it hasn't happened for two reasons. One, thus far we've been successful at raising new money for the project, roughly $46 million dollars in totally new money in fiscal '89. We have to make sure this continues.

Two, the project has been broadened in its definition and scope so that it actually encompasses quite a bit of what was traditional, albeit, fairly high-tech biology. And so it probably is actually providing indirectly additional funding from mainstream biology labs and sapping it. I think it's important that we make sure that that be the trend.

The second issue that I want to touch on briefly is the ethical issue. I think the ethical consequences of this project are quite considerable. But from all I've heard, they are quantitative and qualitative. There are no issues being raised by the Human Genome Project, just the existing issues are being amplified. And I think it's important that we face those issues early in the game and up front and deal with them.

I think Jim Watson mentioned this morning that he plans to use some NIH funds for education. What I've been struck by thus far is that the agencies have made -- especially

DOE and also NIH -- small efforts to educate the press. And if you have very few bucks to spend on education, it seems to me this is probably the best place you could put it right now because if we can get more accurate reporting on what's going on, we will educate a large number of people. I think those are the key issues.

Goals. It strikes me that there's a remarkable consensus at this point as to what the short-term goals of the project ought to be. And considering how new the project is, this is very satisfying. I think we will be mapping. We will be developing techniques. We all agree that at this stage a major hurdle is to try and organize the informanics and computing aspects of the project. That's clearly the arena where at least nationally, and probably internationally, efforts will be made to try to orchestrate things properly.

I think we also all agree, although it came very late in today's discussion, that the project today is people limited, and it's likely to be so for quite a considerable time. So it's critical to build a training component into the Human Genome Project somehow. That's what I wanted to say about goals.

One other thing about goals, you heard an enormous spectrum, especially in this afternoon's session, of optimism and pessimism. I think that's appropriate for people who are actually trying to do the work with their own hands. And I

think it's simply a natural part of human nature and science. And as always, I'm in the middle of the road. I don't think the optimists are right. I don't think the pessimists are right. This project is going, and it's going to continue going.

I want to talk about California for a second. I think the simplest thing I can say is that I don't think any other state could put together such a program with only one exception, only people whose home is or will be in this state. Since the project is people limited, it's destined that California will play a major role in the project so long as it holds all these people.

The only bad note is that it's quite clear from what was said this morning that you have to have two jobs to somehow succeed. That's a little taxing to say the least.

I think that one other way that I could just in the process of arranging for a meeting, which will be held in October in San Diego on the Human Genome Project, co-sponsored by Science Magazine and by the Human Genome Organization, and in organizing a meeting, you usually worry about your travel budget. And I was very surprised as we began to invite all the people we wanted that we don't have any trouble with our travel budget because so many of them come from California that it's not going to be so bad.

Finally, let me turn to what I still consider to be

the most difficult challenge with this project, and that is organization. We as scientists are simply not used to being told what to do, and justifiably we resist it. And we're also used to a system in which people sort of individually compete in many research areas, and the person who gets there first almost beautifully gets the awards.

Those concepts have to change a bit if you're talking about live science. I mean the physicists have gotten used to the fact that they have 50 or 60 people working on a physics project at a major accelerator center. And somehow, even though someone was 50th over, they get a Ph.D. And the biologists haven't dealt with this problem yet of how you coordinate very large efforts and share the credit in an appropriate way.

And I think a model to this has its strains but also it's successes is the search for the Huntington's disease gene coordinated by the Hereditary Disease Foundation represented by Nancy Wexler who is in the back of the room. That project has been going for four years, and it is coordinating internationally, seven or eight research groups, which have their cooperation and their intentions.

But I think it serves as a pretty good local -- since it's California -- model for the fact that it's possible for biologists to carry out relatively large collaborative research effort and still remain friends.

So I think that's what I wanted to say. I look forward to another meeting of this type, hopefully, with even more optimism.

DR. SALSER: Thank you, Dr. Cantor. And thank you all for participating. And I hope we see you again as this project goes forward.

[Thereupon the hearing was concluded at 5:46 p.m.]

CERTIFICATION

STATE OF CALIFORNIA)
) ss.
COUNTY OF VENTURA)

I, ELLEN LYNN PHILHOWER, do hereby certify that the foregoing pages 1 through 235, inclusive, comprise a true and correct verbatim transcript of the proceedings as reported by me.

Witness my hand this 4th day of January , 1989, at Ventura, California.

ELLEN LYNN PHILHOWER

PRISCILLA PIKE COURT REPORTING SERVICES
3639 E. Harbor Blvd. Suite 203-A
Ventura, CA 93001
805-658-7770