Department of Energy
Office of Health and Environmental Research

SEQUENCING THE HUMAN GENOME

Summary Report of the Santa Fe Workshop
March 3-4, 1986

Los Alamos National Laboratory
Los Alamos, New Mexico 87545

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36.

DEPARTMENT OF ENERGY

OFFICE OF HEALTH AND ENVIRONMENTAL RESEARCH

SEQUENCING THE HUMAN GENOME

SUMMARY REPORT ON THE SANTA FE WORKSHOP

(MARCH 3-4, 1986)

Executive Summary. The following is a summary of the Santa Fe Workshop held on March 3 and 4, 1986. The workshop was sponsored by the Office of Health and Environmental Research (OHER) and Los Alamos National Laboratory (LANL) and dedicated to examining the feasibility, advisability, and approaches to sequencing the human genome. The workshop considered four principal topics:

I. Technologies to be employed.
II. Expected benefits.
III. Architecture of the enterprise.
IV. Participants and funding.

I. Technology

The participants of the workshop foresaw extraordinary and continuing progress in the efficiency and accuracy of mapping, ordering, and sequencing technologies. They suggested that a coordinated analysis of the human genome begin with the task of ordering overlapping recombinant DNA fragments obtained from purified human chromosomes that would provide an infrastructure for sequencing activity. At the same time, they support in-depth evaluation of current and developing strategies for sequencing including possible applications of automation and robotics that would minimize the time and cost of sequencing.

II. Benefits

The socio-political and health benefits, and the benefit:cost ratio were seen as highly favorable not only for human health, but in addition for the development of new diagnostic, preventative and therapeutic tools, jobs, and industries. A coordinated analysis of the human genome would also provide significant leverage for biotechnology activities in the private sector and academia. Overall, the human DNA sequencing initiative was seen to have significant financial and health benefits for the citizen and the nation. Furthermore, application of the evolving mapping and sequencing technologies in agriculture could have an even larger impact on world food supplied and economic development.

III. Architecture

A detailed organizational design for human genome sequencing activity will gradually evolve through the continuing efforts of

an OHER central steering committee reporting to Dr. Charles DeLisi, Director, OHER. An extensive network for data collection, organization, and distribution, based on but extending the models provided by the database efforts (e.g., Genbank and the Human Gene Mapping Library) currently serving the molecular genetics community. The participants were inclined toward a distributed laboratory effort to develop and implement technologies, sequencing strategies, ordering and mapping activities, and to facilitate the distribution of cloned and ordered DNA fragments. "Cottage industry" and single institute approaches were also discussed extensively.

IV. Participants and Funding

It is clear that the enterprise of sequencing the human genome should involve the participation of academia and the private sector, foundations, and federal agencies as well as the DOE National Laboratories. There are also great opportunities for international cooperation. In keeping with activities which are multi-institutional in scope, the funding structure should eventually reflect support from the spectrum of traditional funding sources including the DOE, NIH, NSF, private foundations, the commercial sector, and, possibly, international funding consortia coordinated by the World Health Organization.

In summary, the scope and significance of gene sequencing activities are vast and offer a superb benefit:cost ratio. A number of previous DOE efforts and ongoing OHER activities (e.g., the National Laboratory Gene Library Project, Genbank) and DOE National Laboratory capabilities complement the vast efforts sponsored by NIH, NSF, and other foundations and medical institutes. These and developing efforts on other continents should provide the infrastructure for a coordinated and effective approach.

REPORT

Introduction and Background

The Santa Fe Genome workshop was attended by 43 participants from the United States and Europe, of whom 18 were from DOE Laboratories and the remainder from academia and the private sector. The workshop participants enthusiastically discharged the obligation placed before them by Dr. Charles DeLisi, Director of OHER.

Dr. DeLisi requested that the participants evaluate the feasibility and potential utility of obtaining the complete sequence of the human genome, and, identify those initial steps which might be logically taken in effective pursuit of such a goal. Professor Frank Ruddle of Yale University served as Chairperson for the workshop and also provided a succinct historical perspective of the history of human gene mapping and sequencing.

The workshop participants examined four principal issues:

I. An evaluation of existing and emerging sequencing and ordering technologies.
II. Expected benefits which might accrue from gene sequencing and mapping activity.
III. Proper models for the institutional and administrative architecture and coordination of the sequencing enterprise.
IV. Participants and potential funding sources for genome sequencing activity.

Nota Bene: except where otherwise indicated, this report attempts to capture the consensus views apparent during workshop proceedings; in those instances where no consensus was apparent, the several competing views are discussed.

I. Technology Assessment. In the discussion of technologies which support ordering and sequencing the human genome, a broad range of subjects were addressed including the proliferation in sequencing techniques, a variety of approaches to the ordering of large and small DNA fragments, technologies for the "digestion" and physical separation and cloning and of human DNA fragments and the possible applications of automation and robotics to a variety of these processes.

A major emphasis concerned the rapidly changing current status of sequencing technology. While current costs may be in excess of one dollar per base pair, rapid progress toward more cost-effective sequencing (a penny or less per base pair) was envisioned simply as a consequence of the extension or combination of existing technologies and/or the development of new strategies. For this and related reasons it was emphasized that the most circumspect strategy was to focus initially on the development and application of ordering methods and technologies, especially as applied to the ordering of chromosome-specific lambda and cosmid libraries being developed under the auspices of OHER's National Laboratory Gene Library Project. The ordering of such libraries

would provide important immediate advantages for basic and clinical science. Furthermore, ordered overlapping clones would themselves provide an infrastructure and the raw materials for sequencing. An initial focus on ordering chromosome-specific clones would also allow time for careful assessment of a spectrum of emerging sequencing technologies as well as the development and application of automation and robotics to both ordering and sequencing methods. Thus, large-scale sequencing would initially be deferred until such time as improvement in technology and the application of automation and robotics would allow a parsimonious and technologically advanced assault on large scale human DNA sequencing.

There was strong agreement that integration of existing Mendelian, RFLP, and restriction maps with the ordering of large insert chromosome-specific libraries would be of immediate value for both clinical and basic researchers.

Considerable emphasis was given to existing DOE/OHER activities which have provided a valuable technological infrastructure for both physical mapping and future sequencing activities. These include the GenBank database (co-sponsored by the NIH), the automated sorting of specific chromosomes, and the National Laboratory Gene Library which has almost completed preparation of human, chromosome specific, small-insert recombinant DNA libraries and is beginning to address those technological problems associated with construction of overlapping, large insert libraries by both lambda and cosmid cloning.

The participants emphasized that much of the sequencing itself could be accomplished by trained technicians (and/or dedicated machines) and that important priorities should be given to evaluating which aspects of ordering and sequencing might best be automated and/or robotized.

While the principal emphasis of the workshop was focused on mapping and sequencing the human genome, there was also discussion of the value that would accrue from having additional data on the mouse genome and, in particular, a mapped mouse cDNA library accompanied by specific data with regard to the developmental timing and tissue dependence of specific gene expression. The development of heterozygous mouse strains with deletions, which in toto would cover the major portion of the mouse genome, was also a technical desideratum. These tools would provide important biological leverage for further understanding of chromosome organization and gene expression, especially during development. There was also interest expressed in trial "genomic" runs with the DNA of lower eukaryotes, such as yeast, in order to develop and evaluate a spectrum of ordering and sequencing technologies, while also accumulating valuable data in widely studied systems.

Information Resource

There was widespread appreciation of the fundamental importance of computational technology and the theoretical problems inherent in acquiring, storing, retrieving, and analyzing the enormous

amount of information contained in the approximately 3 x 10^9 base pairs, and >100,000 genes which comprise the haploid human genome. It was agreed that current computer capabilities are sufficient (even now) to deal with the number of memory bytes these sequence data would occupy. However, the more complex analytical tasks, especially regarding complex sequence comparisons would require considerable research into and development of appropriately optimized hardware and software. It was also pointed out that recent developments in massively parallel processing would prove well-suited to many of the more complicated aspects of sequence analysis.

Ordering and Mapping Technologies. There was considerable emphasis upon the immediate utility (both to clinical and basic researchers) of the development of maps which would provide the opportunity to integrate both physical and genetic information. Such ordered maps (generated with overlapping, chromosome specific, cloned fragments) would make sequencing more efficient and immediately meaningful. Ordering and mapping of chromosome specific DNA was seen as an initial priority which would pave the way for subsequent sequencing activities and also permit time for newer sequencing technologies to develop and mature. A variety of mapping technologies were mentioned including the techniques of chromosome jumping, the use of restriction site commonalities as mapping tools, and especially the ordering of clones utilizing the small and large insert libraries constructed through the Los Alamos/Livermore National Laboratory Gene Library Project. A recently developed and quite promising approach assigns signatures to DNA fragments by testing their hybridization with a small library (e.g., about 50) of synthetic random sequence oligonucleotide probes. It was clear that utilizing chromosome specific material could considerably simplify the entire process and possibly to resolve certain "closure" or "end game" problems associated with repetitive DNA sequences as well as sequences resistant to existing cloning methods and strategies(see below). Possible ways of reducing the complexity of the mapping problem were also discussed. Hydatiform mole was mentioned as a starting material since its DNA is not heterozygous. However, questions concerning possible abnormalities in mole DNA were potentially disconcerting even in moles which appear normal in chromosomal morphology.

There was unequivocal enthusiasm for the ordering of cloned fragments from one of the smaller chromosomes, as an initial project, to preceed sequencing and to test a spectrum of ordering strategies. There was significant interest in pulsed field gel electrophoresis, which offers the capability of physically separating extremely large DNA fragments ranging upwards from a half million to as many as five million base pairs. (The generation of such large fragments is accomplished with restriction enzymes such as Not I or SfiI, whose restriction recognition sites are relatively rare in genomic DNA.)

Two fundamental approaches to fragment ordering have been developed and were extensively discussed. The first (developed by Olson and also by Brenner) utilizes signatures which are gener-

ated from fragments by the use of a battery of restriction enzymes. Restriction signature mapping technology is labor intensive, systematic, and can reliably identify overlapping domains in collections of cloned random DNA fragments. A newer, stochastic, and more computational technique developed by Lehrach depends upon the hybridization of random sequence oligonucleotide probes to a battery of overlapping DNA fragments. Recent computational simulations of this technique (by Goad and also by Lehrach and Michaelis) are quite promising and have given rise to the estimate of an order of magnitude decrease in the time and labor requirements for ordering. More definitive empirical evaluations of the use of random oligonucleotide probes should emerge in the near future.

Sequencing Technology. As a prelude to the discussion of sequencing technology, the workshop participants placed emphasis on ordering and mapping ab initio with sequencing efforts focusing initially on identifying individual or combinations of technologies with the greatest promise.

The rapidly changing character of sequencing technology profoundly influences estimates of cost and time for whole genome sequencing. The participants began their discussions with current state-of-the-art methodologies with only minimal modifications.

1. Workshop estimates indicated that (1) it curently costs approximately $100,000/year to support an individual technical staff person, and (2) such an individual can sequence, using current technologies, up to 100,000 bp/year. Given the size of the human genome (3×10^9 bp), the application of current technology to sequencing the complete human genome would require approximately 30,000 man-years and $3,000,000,000. (Note that this does not include the cost of the ordering of fragment libraries that would presumably precede the sequencing or the cost of providing a computer-based information resource to manage the sequence data.)

2. With other modest improvements and enhanced application of currently envisioned technologies or adaptations of emerging technologies, cost is dramatically lowered. It seems a relatively simple matter (either with the Applied Biosystems device or the multiplex sequence approaches developed by Church) to generate a million base pairs per year per worker, reducing the time required to 3,000 man years and the cost per base pair to a dime. Under these conditions the entire enterprise would cost about 300 million dollars (over a 10-year period).

3. With rational and conservative extrapolation of current developments and modest expectations for growth and improvements in the technology, it should become possible for a properly equipped worker to generate 10 million base pairs per year at a cost of a penny per base pair and 300 man-years in effort. In this mode, the overall dollar figure is 30 million.

It thus becomes clear, that considerable reduction in costs is

not only possible but expected. This enterprise can properly be compared to the reduction in costs for computation (e.g., as manifested in the shrinkage in cost and explosion in quality of the hand held calculator).

There is clearly a profound need for a carefully planned phasing of the project with initial exploration of optimal sequencing, mapping, and ordering strategies and technologies with the appropriate initial effort directed toward ordering. Once again it was emphasized that the possibility of achieving costs as low as a penny per base pair was not seen as unreasonable within the next decade and it was acknowledged that even this estimate could actually turn out to be quite conservative.

The conferees agreed that permissible sequencing error rates would have to be quite low in order to efficiently detect such rare events as mutation and in order to properly evaluate the extent and clinical significance of human genetic heterogeneity. Strategies for sequence verification and error reduction are emerging and there is cause for optimism with regard to this parameter. For example, strategies involving focal replicate sequencing are seen as promising ways to address the issue of evaluating sequencing accuracy.

Discussion was also dedicated to the Applied Biosystems sequencing instrument which uses laser flourimetry and a proprietary modification of the Sanger method, involving primer extension and fluorochrome labeled dideoxynculeotide chain termination. The Applied Biosystems approach benefits from extensive automation and freedom from the use of isotopes and film emulsions for obtaining the data. There was also an interesting discussion of George Church's multiplexing system where restricted DNA fragments are transferred from sequencing gels to nylon membranes. This transfer permits successive probing of the DNA fragments with a variety of end-specific and strand-specific probes. It was also pointed out (by Ward) that substitution of peroxidase chemiluminescence in place of autoradiography would provide substantial increases in speed when added to the multiplexing approach.

There was repeated emphasis on the need for integration of the ordering of cloned fragments, the mapping of genes, the mapping of restriction fragment length polymorphisms, and the overall sequencing effort.

Participants estimated that as much as 99% of the genome might be stably carried in current cloning vectors. Nevertheless, an intriguing caveat was expressed concerning the possibility of unforeseen problems in sequencing "closure" otherwise described as the "end game." For example, the possibility that parts of a chromosome might be resistant to sequence analysis because they include large stretches of repetitive DNA deficient in restriction enzyme sites, or "refractory" sequences which are otherwise ill-suited for propagation in current cloning vectors. Rebuttal of this concern was based on the use of chromosome-specific DNA (to address repeat families extending over more than one chromo-

some) as starting material and on the idea that with sufficient numbers of overlapping fragments (to address regions of tandem repeats) even refractory domains would eventually succumb.

Other concerns surrounded the repetitious and somewhat unexciting nature of the sequencing activity itself, which might result in the development of sequencing information only for the most "clinically interesting" domains. It was acknowledged that the problem could, to a significant extent, be addressed by the application of robotic and automation methodologies wherever feasible in the entire operation, starting with DNA preparation and extending to the automated transfer of data from the sequencing apparatus to the data base. Moreover, the very important possibility was considered that nontranscribed domains may contain vital information with regard to either DNA packaging, chromosome duplication and/or gene expression. Such considerations appeared to increase the likelihood that long range sequencing of intergenic regions (beyond and between "transcribed" DNA domains) would eventually be intensively pursued rather than ignored. While "geared-up" sequencing activities would not be emphasized initially, some modest level of sequencing was encouraged even from the start, if only as a means for evaluating competing technologies.

Once again the central coordination and integration of genomic sequencing especially with the DNA database was seen as an important feature which could minimize duplication, optimize accuracy, and ensure technical excellence. Moreover, the advantages which would accrue from judicious application of automation and robotics technologies and supporting computational technologies were subjects which received repeated emphasis.

II. Expected Benefits of Sequencing Activity. Substantial benefits were identified and repeatedly emphasized by the participants in the Workshop. It was clear however that the perceived benefits varied in a significant way as a function of the research interests of the participants. Generally, it was agreed that detailed knowledge of the full organization and molecular structure of the human genome will have a powerful impact on our understanding of such basic areas as embryogenesis, the molecular mechanisms that regulate gene expression, and the molecular basis for inherited diseases.

One of the most striking points to emerge at the workshop was the high benefit:cost ratio, in spite of the great variability in estimates of cost. There was widespread expectation of rapid shrinkage both in the costs for sequencing and the time requirements in terms of man-years of effort to complete the task. This guarantees an immeasurable gain in our understanding of human biology and medicine for a relatively modest outlay of funds.

The benefits seen were clearly both of a basic and of a clinical nature. It was emphasized that extended sequencing information would have a very profound effect on our understanding of the regulation of gene expression. Many expressed the view that so called "uninteresting" domains (e.g., nontranscribed and/or

repetitive sequences) may well have a profound influence on both the condensation and packaging of DNA prior to mitosis and the deployment of expressed domains in the interphase nucleus. The remarkable number, 2.4 meters, is the current best estimate for the total length of the DNA in a single human cell nucleus. This underlines the extraordinary complexity and sophistication in packaging which is needed to coil this length of polymer into a readily deployed template which is routinely probed for information and gene expression within the very limiting confines of a 7 micron cell nucleus.

There was strong expectation that extensive sequencing would shed significant light on basic information with regard to the processes of aging and cancer as well as other disorders with a genetic predisposition. In this view, aging itself may, in large part, be genetically preprogrammed. It was also emphasized that many as now unrecognized genes would be likely to exhibit profound therapeutic affects both in the form of therapeutic DNA inserted by genetic engineering and/or somatic gene replacement. An important role was also seen for human gene products which could be mass-produced in micro-organisms and administered to man as replacement therapy. It was also pointed out that one of the principal concerns with genetic engineering is not merely what coding region to insert but where and how to localize the insertion. (The insertion locus has a profound influence on the level of expression of the newly introduced genetic information.) An understanding leading to the optimization of these loci would be one of the most profound basic and clinical consequences of long-range sequencing.

Mention was also made of profound increases in our understanding of radiation and chemical damage and its repair in the genome. There were very sanguine expectations for unprecedented growth in our understanding of both radiation and chemical genetic toxicology. There was also enthusiasm for sequencing as the ultimate tool for understanding and defining the extent of human genetic diversity. Some participants felt that it would be necessary to have sequence information on not one but many individuals especially for "clinically relevant" genetic domains. Such portions of the genome either influence inherited diseases or contain or regulate the expression of genes, which profoundly influence health. The expected routine use of sequence information in clinical practice would obviously benefit by a greater wealth of sequence information.

A very large impact was expected on our understanding of the huge collection of human neoplastic and malignant disorders, as well as benefits for both metabolic and cardiovascular diseases. Substantial gains would also accrue for a spectrum of pulmonary fibrosing disorders, genetically determined nephropathies, the whole range of human arthritidies and also auto-immune and inherited immune-deficiency diseases. The complete genomic sequence will of course have a profound impact on our understanding of the rapidly growing collection of RFLPs which are linked to particular inherited diseases. There was also a strong expectation of

benefits with regard to behavioral and psychiatric disorders many of which have a readily demonstrable familial/genetic basis.

In terms of sociological and economic benefits, the gain in knowledge for human health will markedly improved the quality of human life. The participants empahsized the potential to substantially reduce loss of productivity associated either with acute catastrophic illnesses or chronic debilitating diseases such as pulmonary fibrosing disorders and both acute and chronic arthritidies. The Department of Health and Human Services estimates that the United States public will have spent over 400 billion dollars in health care in 1986. (This is exclusive of any investment in biomedical research!) The health advantages arising from the availability of complete genomic sequences will eventually have a very large impact on such expenditures, sufficient to more than offset the cost of the entire enterprise. Even a 1% reduction in current health care costs would save far more than even the most pessimistically estimated cost of ordering and sequencing the entire human genome! Enthusiasm was also expressed with regard to benefits for agriculture from the development of gene sequencing and mapping technologies. It was pointed out that the portion of gross national income dedicated to health care was <u>far less</u> than our expenditures on food! The use of sequence information for the development of grain cultivars with increased yields livestock improvement, and pest control would have a profound and indeed an immeasurable impact on world agricultural practices and economics.

Substantial fiscal benefits are also expected from the biomedical, computational, and robotic products and jobs spawned by the sequencing effort. In addition, important opportunities for international cooperation were cited as potential political benefits. There are, even now, substantial efforts in Japan, Europe, and the Soviet Union with regard to developing a more detailed understanding of human DNA. Finally, the participants acknowledged that information arising from the proposed genomic sequencing activity might be misunderstood and misused, as an invasion of privacy and encroachment upon individual rights. However, they were unanimous in their view that the expected benefits accruing to each individual both in terms of health and economic status would far outweigh the potential for abuse of this information.

III. <u>Architecture and Model of the Enterprise</u>. A lengthy and impassioned discussion of the architecture, coordination, and organization of genomic sequencing activity generated a very wide spectrum of opinion. There was, however, strong support for a well conceived central coordinating body, which would perennially identify and evaluate promising technological approaches and identify the most appropriate participants both at the investigator and institutional level. The participants strongly supported the idea of appointing, under the aegis of Dr. DeLisi, a distin-

guished OHER advisory committee* which would continue to guide the development of strategies, review grant proposals and evaluate progress as genome sequencing activities evolved. The advisory committee would facilitate implementation of especially promising research lines, the development of an integrated funding structure, and the identification of more powerful technologies. The idea of rapidly identifying a distinguished steering committee which would facilitate the evolution of an architecture for genomic sequencing continued to receive emphasis and strong agreement. It was also clear that a computational base residing in a dedicated information resource would provide an important coordinating framework.

This central resource could provide important safeguards against inadvertant duplication of effort, and a useful form of interaction among the participants. The central coordinating capability, in concert with the steering committee, would be needed to map out strategies for the development of such laboratory operations as cloning, mapping, and sequencing and for the management and distribution of data as it flowed into a central data collection facility. A central facility would also coordinate the distribution of critical materials (e.g., chromosome-specific DNA clones) which would support ordering and sequencing activities and provide an invaluable resource for the research community. The participants expressed concern that a centralized and monolithic sequencing institute (advocated by some) might provide greater focus and technical consistency but it might also stifle innovation and thus retard progress. On the other hand, decentralization suffered from the risks of diffuseness and problems of quality control. The most promising approach appeared to combine a strong central administrative/coordination facility with a carefully distributed research and development effort.

Another important function in addition to ordering, sequencing, computational analysis, and material distribution was education. It was felt that for the project to be adequately understood and supported some effort was needed to educate the public, other scientists, academicians and the corporate sector to the very favorable benefit cost ratio and the biomedical, basic, and economic significance of the proposed undertaking.

There were continual reminders that the enterprise must begin in a very flexible mode in view of the current dynamically evolving status of both theory and technology. Thus the advisory committee would be expected to develop a number of competing objectives whose relative merit and value to the overall effort would be periodically evaluated in the context of ongoing experimentation. There was also strong support for a computer-based information resource for data management. This could perhaps be based on models such as GenBank whose staff interact with both the genetics and other nucleotide sequencing databases. The Genbank staff has already addressed many of the issues of data management that

*This committee, which is called the Human Genome Steering Committee, is now in existence and its membership is given in Appendix I.

would be shared by an information resource dedicated to the human genome project

It was clear that preliminary decisions of how to proceed, and selection of near- and medium-term goals would also have an impact on the architecture. For example, the approach for focusing, mapping, and sequencing activities upon "clinically interesting" loci would differ from that intended to develop a database and infrastructure for the total linear sequence of a single intact chromosome. It was also felt that if the genomic sequencing capability achieved a level of productivity approaching its expected capacity, it would also provide extraordinary research value in sequencing non-human DNA including mouse, yeast, and also plants. The basic and biomedical benefits from such activities could ensure continued utility of the developed administrative/organizational structure and its research capabilities.

Emphasis was given to involvement of the private sector by use of the contract mechanism or direct funding and the involvement of academia, again by use of a consultant mechanism or direct grant funding. It was also felt that both the private sector and academia, as well as other participants might communicate effectively by utilizing the planned computational network for submitting and retrieving mapping sequence and other experimental data.

As noted above, there was considerable debate concerning where the actual genomic sequencing should be undertaken and by whom. The spectrum of opinion here varied from the establishment of a single centralized sequencing institute versus a more distributed effort. In the latter, perhaps 10 centers would all be coordinated by a strong central administrative body (i.e. the Human Genome Steering Committee) and would interact with very strong regional basic and clinical research activities, either in academia, the private sector, or the National Laboratories. Selected centers might be especially qualified for particular sequencing responsibilities. Sequencing could proceed apace once ordering of overlapping chromosome-specific, cloned DNA fragments had been accomplished. Thus, promising interagency coordination and cooperation was seen as a vital function of the steering committee.

There was emphasis on the distinction between a focused but distributed versus a diffuse effort. A distributed effort could remain focused and retain both quality control and strong coordination with a central body if each center was selected and staffed with great care. There was also considerable interest in the possibility of international cooperation and collaboration. It was agreed that the detailed structure of a multinational effort would need significant thought and multilateral agreements. Data sharing among the various data bases is already an accomplished fact. While no consensus was achieved in the discussion of distributed versus central architectures, the advantages of each were discussed. The wider participation and scientific benefits of the distributed model attracted much support.

In summary, the architecture of the sequencing enterprise would of necessity reflect the priorities chosen by the advisory committee. The architecture is expected to be a self-evolving structure which retains flexibility and responsiveness to the needs of the enterprise and which can readily adopt the required intra-course corrections.

IV. Funding and Participants. One approach to dealing with the overall funding of the genome sequencing project would be to structure, ab initio, a multi-institutional or multi-national support system. Clearly the success of such an enterprise will depend upon strong interaction among the potentially interested institutions and it may be necessary here to take a "wait and see" attitude. There are already ongoing mapping and sequencing efforts in the US, Europe, Japan, and the Soviet Union, and the possibility of developing an international effort for the design and funding of the project is quite attractive.

Role of DOE/OHER

The DOE's Office of Health and Environmental Research has supported a number of seminal contributions in the area of molecular biology. In enumerating these contributions, one must begin with the historical interest of the DOE in the areas of genetic toxicology, mutagenesis, carcinogenesis, and physical/biological studies of chromatin structure. More recently OHER has provided very useful contributions in its support of the National Laboratory Gene Library Project and (in combination with the NIH) GenBank activities as well as its support of the development of chromosome sorting by flow cytometry. The National Laboratory Gene Library Project now enters a second (large insert) phase as a collaborative effort between Lawrence Livermore Laboratory and Los Alamos National Laboratory. The cloned libraries should provide important starting materials for ordering, mapping, and subsequent sequencing activities.

In addition to OHER there has been significant interest expressed in this project by the Howard Hughes Medical Institute, which has given much previous support to gene mapping activities. There is acknowledged interest on the part of the National Institutes of Health in the overall area of sequencing activities, especially as they relate to problems in aging, cancer, and activities in genetic engineering. NSF has also expressed interest. The World Health Organization might also constitute one of a number of possible avenues for the development of international cooperation and potentially a consortium of funding instruments.

A major concern of the Santa Fe workshop participants was to clearly establish, for the biomedical research community, that the funding sought by OHER for this work would be incremental and would not impact on NIH R01 funding effort or on OHER's existing research programs. The importance of making this distinction, and of encouraging the enthusiastic participation of the academic and private sectors, was continually emphasized. Concern was expressed that we avoid the misunderstanding of focusing upon the more pessimistic cost estimates and creating the impression

that several billions of dollars would be subtracted from NIH budgets in order to proceed immediately with a massive attack on human genomic sequencing. This was clearly not the intent nor is it a pragmatic approach. DOE was seen as a favorable locus for initiating and coordinating a plan leading toward comprehensive sequencing activities and involving many in the molecular biology community. The DOE budget was seen as discrete from NIH and foundation budgets and the work was seen as a practical extension of several existing programs supported by OHER with important contributions from the NIH. Emphasis was also given to the timeliness of developing a discrete mechanism for education of the scientific community via workshops and of the general public through the communications media so that the beneficial impacts of the planned sequencing activities upon the health and economic well being of the average citizen would be more thoroughly understood and more widely appreciated.

In conclusion, the most critical points emphasized were (1) the extraordinary benefit:cost ratio of the genomic mapping and sequencing activity and (2) the idea that the activity deserves and requires multi-institutional support and cooperation, as well as a multi-national constituency.

Acknowledgments:

I wish to thank the members of the Los Alamos "Genome Team" for their invaluable contributions to this report. They are: Dr. C. E. Hildebrand, Dr. L. Deaven, Dr. B. Moyzis, Dr. B. Shera, Dr. C. Burks, Dr. W. Goad, Dr. J. Fickett, Dr. D. Knobeloch, Dr. T. Beugelsdijk, and Dr. R. Walters. Also, I acknowledge the valuable post-workshop comments contained in letters from the participants. They include: Dr. G. Bell; Dr. C. Burks; Dr. C. Cantor; Dr. D. Comings (who also kept detailed notes of the workshop proceedings); Dr. R. Gelinas; Dr. J. Hearst; Dr. W. Goad; Dr. E. Heron; Dr. M. Kashdan; Dr. S. Brenner; Dr. S. Mitra; Dr. T. Puck; Dr. F. Ruddle; Dr. H. Smith; Dr. G. Stein; Dr. J. Stein; Dr. H. Sutton; Dr. C. Ward; and Dr. S. Weissman.

I also acknowledge the invaluable contributions of all the attendees at this workshop. (They are listed in Appendix II)

I am responsible for any factual errors or misplaced emphases which may be found in this report. Very special acknowledgement is due Christian Burks of the Theoretical Division and Ed Hildebrand of the Life Sciences Division whose detailed and perceptive reviews contributed in a striking way to accuracy and objectively of this report. Finally, I thank Ms Martha Waters and Ms Irma Lujan for their very proficient preparation of the manuscript.

M. W. Bitensky M.D.

M. W. Bitensky, M. D.
Senior Fellow
Los Alamos National Laboratory
December 3, 1986

GENOME SEQUENCING WORKSHOP
MARCH 3-4, 1986
PARTICIPANTS

Norman Anderson	Proteus Technology Rockville, MD
Ben Barnhart	OHER Washington, D.C.
George Bell	Los Alamos National Laboratory Los Alamos, NM
Mark Bitensky	Los Alamos National Laboratory Los Alamos, NM
Fred Blattner	University of Wisconsin Madison, WI
Albert Branscomb	Lawrence Livermore National Lab Livermore, CA
Sidney Brenner	Medical Research Council Cambridge, England
Christian Burks	Los Alamos National Laboratory Los Alamos, NM
Charles Cantor	Columbia University New York, NY
Anthony Carrano	Lawrence Livermore National Lab Livermore, CA
Thomas Caskey	Baylor College Houston, TX
George Church	University of California SF San Francisco, CA
David Comings	City of Hope Medical Center Duart, CA
Scott Cram	Los Alamos National Laboratory Los Alamos, NM
Joseph D'Anna	Los Alamos National Laboratory Los Alamos, NM
Larry Deaven	Los Alamos National Laboratory Los Alamos, NM

James Fickett	Los Alamos National Laboratory Los Alamos, NM
Richard Gelinas	Fred Hutchinson Cancer Res. Ctr. Seattle, WA
Walter Gilbert	Harvard University Cambridge, MA
Walter Goad	Los Alamos National Laboratory Los Alamos, NM
Gerald Guralnik	Los Alamos National Laboratory Los Alamos, NM
Joyce Hamlin	University of Virginia Charlottesville, VA
John Hearst	University of California Berkeley Berkeley, CA
Elaine Heron	Applied Biosytems Foster City, CA
Ed Hildebrand	Los Alamos National Laboratory Los Alamos, NM
Maurice Kashdan	Dupont, New England Nuclear Res. Boston, MA
Hans Lehrach	European Molecular Biology Lab Heidelberg, W. Germany
Sankar Mitra	Oak Ridge National Laboratory Oak Ridge, TN
Jane Moores	University of California SD La Jolla, CA
Robert Moyzis	Los Alamos National Laboratory Los Alamos, NM
Ted Puck	University of Colorado Denver, CO
Richard Roberts	Cold Spring Harbor Long Island, NY
Francis Ruddle	Yale University New Haven, CT
David Smith	OHER Washington, D.C.

Hamilton Smith	Johns Hopkins University Baltimore, MD
Gary Stein	University of Florida Gainesville, FL
Janet Stein	University of Florida Gainesville, FL
F. William Studier	Brookhaven National Laboratory Upton, NY
Eldon Sutton	University of Texas Austin, TX
Robert Wagner	Los Alamos National Laboratory Los Alamos, NM
David Ward	Yale University New Haven, CT
Ronald Walters	Los Alamos National Laboratory Los Alamos, NM
Sherman Weissman	Yale University New Haven, CT

APPENDIX II

Meeting of the

U. S. Department of Energy

HEALTH AND ENVIRONMENTAL RESEARCH ADVISORY COMMITTEE

Subcommittee on Human Genome

8:30 a.m., November 6, 1986

Attendees

Chairman

Dr. Ignacio Tinoco
Department of Chemistry and Biochemistry
University of Colorado
Boulder, Colorado 90309

Subcommittee Member

Dr. George Cahill
Howard Hughes Medical Institute
7984 Old Georgetown Road
Bethesda, Maryland 20814

Dr. Dean L. Engelhardt
Vice President for Research
Enzo Biochemicals, Inc.
325 Hudson Street
New York, New York 10013

Dr. Leroy Hood
Chairman
Division of Biology
California Institute of Technology
Pasadena, California 91125

Dr. Leonard S. Lerman
Genetics Institute
87 Cambridge Park Drive
Cambridge, Massachusetts 02140

Dr. Mortimer L. Mendelsohn
Associate Director for Biomedical
and Environmental Research
Lawrence Livermore National Laboratory
P. O. Box 808
Livermore, California 94550

Dr. Temple Smith
Dana/Farber Cancer Institute
Jimmy Fund Building, Room 315
44 Binney Street
Boston, Massachusetts 02115

Dr. Gary Stormo
Department of Molecular, Cellular
and Developmental Biology
University of Colorado
Boulder, Colorado 80309

Dr. Ray White
Howard Hughes Medical Institute
600 Wintrobe Building
University of Utah Medical Center
Salt Lake City, Utah 84132

SUBCOMMITTEE MEMBERS UNABLE TO ATTEND

Dr. Charles Cantor
Chairman
Department of Human Biology
Columbia University
New York, New York 10027

Dr. Thomas Caskey
Department of Medicine
Baylor College of Medicine
One Baylor Plaza
Houston, Texas 77030

Dr. Renato Dulbecco
Salk Institute
P. O. Box 85800
San Diego, California 92138

Dr. Robert L. Sinsheimer
Chancellor
University of California
Santa Cruz, California 95064

U. S. DEPARTMENT OF ENERGY STAFF

Dr. George D. Duda
Health Effects Research Division
Office of Health & Environmental Research, ER-72
U. S. Department of Energy
Office of Energy Research
Washington, D. C. 20545

Ms. Jean M. Hummer
Management Assistant
Office of Health & Environmental Research, ER-70
Office of Energy Research
U. S. Department of Energy
Washington, D. C. 20545

Dr. David A. Smith
Health Effects Research Division
Office of Health & Environmental Research, ER-72
Office of Energy Research
U. S. Department of Energy
Washington, D. C. 20545

OTHERS PRESENT

Dr. Edgar Hildebrand
LS-3 Division, MS-M886
Los Alamos National Laboratory
P. O. Box 1663
Los Alamos, New Mexico 87545

Dr. Walter B. Goad
T-10 Division, MS-K710
Los Alamos National Laboratory
P. O. Box 1663
Los Alamos, New Mexico 87545