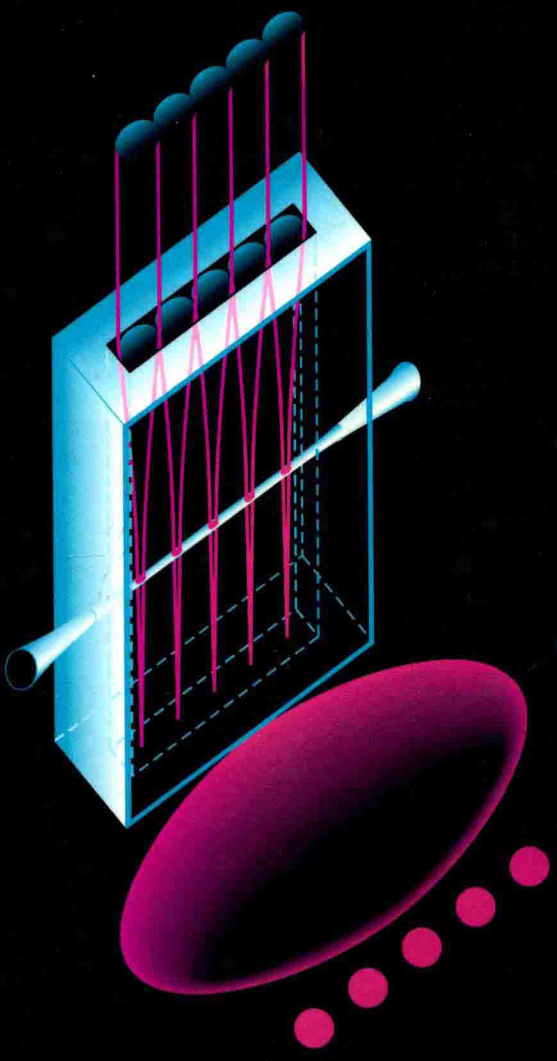
Automation Technologies for Genome Characterization



Edited by Tony J. Beugelsdijk

Automation Technologies for Genome Characterization

Edited by

TONY J. BEUGELSDIJK
Los Alamos National Laboratory
Los Alamos, New Mexico

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

This book is printed on acid-free paper. ♾

Published simultaneously in Canada.

Library of Congress Cataloging in Publication Data:

Automation technologies for genome characterization / edited by Tony J. Beugelsdijk.

p. cm. — (Wiley-Interscience series on laboratory automation)

"A Wiley-Interscience publication."

Includes index.

ISBN 0-471-12806-6 (cloth : alk. paper)

1. Gene mapping—Automation. 2. Gene mapping—Data processing. I. Beugelsdijk, Tony J., 1949- . II. Series.

QH445.2.A95 1997

572.8'633'0285—dc21 96-37604

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Automation Technologies for Genome Characterization

WILEY-INTERSCIENCE SERIES ON LABORATORY AUTOMATION

To the authors
for their contributions
and
to my wife Mary
for her love and support

Contributors

JOHN E. AGAPAKIS, PH.D., Acuity Imaging, Inc., 9 Townsend West, Nashua, NM 03063-1217

DAVID P. ALLISON, PH.D., Oak Ridge National Laboratory, P.O. Box 2008, MS-6123, Oak Ridge, Tennessee 37831-6123

DAVID R. BANCROFT, PH.D., Max-Planck Institute for Molecular Genetics, Abteilung Hans Lehrach, Ihnestrasse 73, D14195, Berlin-Dahlem, Germany

KENNETH L. BEATTIE, PH.D., Health Sciences Research Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6123

BART BEEMAN, Micotechnology Center Lawrence Livermore National Laboratory, 7000 East Avenue, Mailstop L-222, Livermore, California 94550

BILL BENETT, Micotechnology Center Lawerence Livermore National Laboratory, 7000 East Avenue, Mailstop L-222, Livermore, California 94550

CHAU-WEN CHOU, Department of Chemistry and Biochemistry, Arizona State University, Tempe, AZ 85287-1604

BOBI K. DEN HARTOG, Los Alamos National Laboratory, P.O. Box 1663, Mailstop J580, Los Alamos, NM 87545

MITCHEL J. DOKTYCZ, PH.D., Health Sciences Research Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6123

NORMAN J. DOVICHI, PH.D., Department of Chemistry, University of Alberta, Edmonton, Canada T6G 2G2

CHRIS FIELDS, PH.D., Molecular Informatics, Inc., 1800 Old Pecos Trail, Suite M, Santa Fe, NM 87505

H.R. (SKIP) GARNER, PH.D., University of Texas, Southwestern Medical Center at Dallas, 5323 Harry Hines Blvd., Dallas, Texas, 753-235-8591

DEAN HADLEY, Micotechnology Center Lawerence Livermore National Laboratory, 7000 East Avenue, Mailstop L-222, Livermore, California 94550

SCOTT P. HUNICKE-SMITH, PH.D., Stanford DNA Sequence and Technology Center and Stanford Department of Mechanical Engineering, 85 California Avenue, Palo Alto, CA 94304

SARATH KRISHNASWAMY, PH.D., Acuity Imaging, Inc., 9 Townsend West, Nashua, NM 03063-1217

PHOEBE LANDRE, Micotechnology Center Lawerence Livermore National Laboratory, 7000 East Avenue, Mailstop L-22, Livermore California 94550

HANS LEHRACH, PH.D., Director, Max-Planck Institute for Molecular Genetics, Abteilung Hans Lehrach, Ihnestrasse 73, D14195, Berlin-Dahlem, Germany

STACY LEHEW, Micotechnology Center Lawrence Livermore National Laboratory, 7000 East Avenue, Mailstop L-22, Livermore, California 94550

ELMAR MAIER, PH.D., Max-Planck Institute for Molecular Genetics, Abteilung Hans Lehrach, Ihnestrasse 73, D14195, Berlin-Dahlem, Germany

PATRICIA A. MEDVICK, PH.D., Los Alamos National Laboratory, P.O. Box 1663, Mailstop B295, Los Alamos, NM 87545

DEIRDRE MELDRUM, PH.D., Department of Electrical Engineering, University of Washington, Seattle, WA 98195

M. ALLEN NORTHRUP, PH.D., Lawrence Livermore National Laboratory, 7000 East Avenue, Mailstop L-222, Livermore, California 94550

MARTIN J. POLLARD, PH.D., Human Genome Center, Lawrence Berkeley National Laboratory, University of California, 1 Cyclotron Road, M/S 74-157, Berkeley, CA 94720

DAVID M. SCHIELTZ, PH.D., Department of Molecular Biotechnology, University of Washington, Seattle, WA 98195

THOMAS G. THUNDAT, PH.D., Oak Ridge National Laboratory, P.O. Box 2008, MS-6123, Oak Ridge, Tennessee 37831-6123

EMERSON TONGCO, Department of Electrical Engineering, University of Washington, Seattle, WA 98195

ROBERT J. WARMACK, PH.D., Oak Ridge National Laboratory, P.O. Box 2008, MS-6123, Oak Ridge, Tennessee 37831-6123

PETER WILLIAMS, PH.D., Department of Chemistry and Biochemistry, Arizona State University, Tempe, AZ 85287-1604

Preface

The discovery of human disease genes has historically been an arduous undertaking. Extensive and exhaustive studies of genetic inheritance and pedigrees in generations of families led to the discovery of the color blindness gene on chromosome Y in the early 1990s. As more biological tools became available, the pace of gene discovery increased. However, as recently as 1983, when the locus of the Huntington's disease gene was reported and the 1987 discovery of the muscular dystrophy gene, much of the biological laboratory practices were still rooted in intensively manual procedures. By October 1987, only about 1200 genes had been mapped to specific chromosomes or regions of chromosomes. Many of them like Huntington's and muscular dystrophy merited headline attention.

By the late 1980s genes were being discovered at a rate to merit only scant mention in the popular press. Yet, considering that the human genome has somewhere between 50,000 and 100,000 genes, even this "accelerated" pace is hardly breathtaking. It is especially disconcerting to see the enormous resources being applied to hundreds of individual gene discovery efforts when a concerted, large-scale effort at mapping and sequencing the human genome could efficiently address the gene discovery problem and free those resources for the study of gene function and expression. It was with this realization and the temporal juxtaposition of requisite technologies, that the Human Genome Project (HGP) was deemed possible and launched officially in 1990.

The HGP is now well under way. It is the largest, concerted effort in human genetics. The HGP traces its beginnings to the mid-1980s with a series of exploratory meetings on the feasibility of determining the physical map and ultimately the genetic sequence of the entire human genome. These meetings were attended with much excitement and high levels of energy as it was recognized that after decades of key advances in molecular biology, laboratory procedures were in place and well enough understood to begin a systematic and directed effort at this enormous undertaking. However, the scale of the project knew no precedent and far exceeded the then current experience of laboratorians.

With the Human Genome Project, the work of molecular biology expanded beyond the field in a major way. Scientists and engineers in other disciplines began to apply their crafts to the technological challenges presented by the HGP. Human genome centers were formed, funded by the United States Department of Energy and the National Institutes of Health. These centers were chartered with different tasks centered around goals of the Human Genome Project and the interests of their members.

However, the centers that built solid interdisciplinary programs have contributed the most toward the goals of the HGP. A strong presence in molecular biology with the fostering of a culture of interdisciplinary teaming marked their success. Moreover these centers were guided by a strategic vision of the leverage made possible by automation technologies.

Under the sponsorship of these centers and in other laboratories around the world, the latest advances in informatics, optical techniques, robotics, microfabrication technologies, and laboratory information management systems were applied to the challenges of the HGP. Many of the early critics of the genome factory approach have been silenced by the advances made by those organizations that effectively developed and deployed automated systems. The predominant single investigator model of biological research made room for the focused interdisciplinary team model. Successes such as those experienced at Centre de Etude Polymorphisme Humain (CEPH) in France and Human Genome Sciences (HGS) in the United States proved the value of large-scale automated assaults on the human genome. These organizations have made significant advances in mapping and sequencing knowledge, with heavy reliance on automation technologies, and have validated the automation paradigm. In addition to dramatically increasing the pace of gene discovery and furthering genomic knowledge, significant economic value has also been generated by these organizations. Contracts totaling hundreds of millions of dollars have been signed, guaranteeing access to their technology and databases in the highly competitive races for new drug discovery. Several bench scientists have become multimillionaires.

Much has been written about the science of the Human Genome Project. The science is enabled and accelerated by the technology that moves the project toward its goal. This book tells that story and brings together a crosscut of the various technologies in current use or being developed. These technologies were either not present or not extensively applied to DNA characterization at the start of the HGP. The technologies and methods were developed through the stimulus provided by the project. While many systems were initially met with skepticism, it is unthinkable for laboratories today to not rely on these machines, many of which are now commercially available. The authors' work has not only transformed the science of molecular biology but also transformed the "PI-centric" culture of biological science and validated the interdisciplinary team model of investigation.

It is impossible to present in this volume all of the technologies currently being developed or being used in the HGP. Such an undertaking is, necessarily and happily, always incomplete. Exciting new ideas are generated daily as the skills of other disciplines are brought to bear on the challenges of the HGP. Many of the work horse technologies that could have been featured in this book such as fluorescence-based sequencing and flow cytometry, to name two, have already been well documented elsewhere. A review of these techniques would have contributed little new information to an already large body of knowledge. Indeed, this suite of existing technologies form collectively the "shoulders of giants" that the HGP stands on. Omissions of such technologies from this work are not intended to diminish their importance or impact.

One of the early goals of the HGP was to develop detailed physical and genetic maps of each of the human chromosomes. Unlike DNA sequencing, there was a paucity

of commercially available technology to meet the mapping task. The market for such technology is specialized and largely restricted to the large human genome centers, so no technology products were even under development. This situation presented an opportunity for the larger centers and laboratories to staff instrumentation development groups to build the automated systems required. While some of these early systems had limited success, many of them were later refined with the feedback of experimentalists and became very important in the successes of their sponsoring centers. Some have even achieved commercial status.

This book is organized into four sections. The first section describes laboratory automation activities at several major human genome centers and research laboratories. A consistent philosophy emerges for successful implementation of new automation technologies based on the experience of these authors. This section also includes a chapter on imaging technologies and the major role played by them in the image-rich biology laboratory. The second section discusses control system approaches. Both chapters in this section address the problem of integration of dissimilar systems. General features for control systems are distilled from the experience of these authors. The third section of this book describes some advanced and nontraditional technologies being applied to DNA characterization. Finally, this book covers some of the analysis, modeling, and database issues accompanying the characterization of genomes.

Dr. H. R. (Skip) Garner describes his experiences and success of the University of Texas Southwestern Medical Center's laboratory in their automation efforts. He outlines the requirements for assembling an interdisciplinary team and a successful strategy in approaching automation projects, and describes the cultural and nontechnical issues surrounding the acceptance of a new technology. The extremes wherein home-grown automation projects arise are those between "little utility and commercially valuable." While this gives birth to opportunity, it is also a limiting factor in widespread technology transfer. He underscores a common experience that automation is not a product but a continuous process of which only a portion is technological.

Dr. Martin J. Pollard of Lawrence Berkeley National Laboratory (LBNL) describes his labotatory's approach to automated systems design. The bottom-up approach used at LBNL was dictated in the early stages of the HGP by the need to establish new working relationships with experimentalists and the low funding levels for automation development. Building on early successes, LBNL has built several systems that have been of great utility and commercial interest.

Drs. Elmar Maier and David R. Bancroft of the Max-Planck Institute for Molecular Genetics present their organization's hybridization approach to the construction to physical maps and their development of successive generations of robots to meet the demand generated by their reference library database. They have developed and describe their systems for clone picking, high-density filter array generation, image analysis, and high-throughput polymerase chain reaction (PCR). Several commercial products trace their origins to their very innovative laboratory.

One of the key technologies incorporated into the above systems is machine vision. Drs. Sarath Krishnaswamy and John E. Agapakis of Acuity Imaging, Inc. describe machine vision and give an overview of the contributions imaging has made to machine guidance as well as data collection. They discuss the development of a system that uses

machine vision in conjunction with a motion control device that enhances a previously "blind" process—that of colony picking. A once tedious and error-prone operation has become very accurate, efficient, and highly automated using vision technology.

One of the pervasive difficulties in building automated systems is the lack of interconnect standards and interfaces for both hardware and software modules. Each system thus becomes a unique design, and except for systems developed at a particular institution, exchange of systems or components is not possible. While standardization activities are underway, progress has been very slow. The two chapters in Part II of this book are devoted to attempts to develop a standardized software interface. Dr. Pat A. Medvick and Bobi K. Den Hartog of Los Alamos National Laboratory (LANL) describe a high-level software controller developed at LANL. Building upon and extending an existing automation effort at LANL, they describe the use of "Services" which, when linked together, become a program called the script of a method. Taking another approach, Scott Hunicke-Smith has developed the Graphical User Interface to Laboratory Equipment (GUILE) technology. GUILE was created to present a consistent, extensible, yet simple interface to devices and protocols. Both developments recognize the fundamental interconnection and integration problem facing developers of new automation systems and propose models for standardization. While still early to predict widespread commercial acceptance of these technologies, the ground-breaking work of these authors generates a certain momentum and features of their work may eventually emerge in future powerful development and integration tools.

Part III of this book treats several emerging and advanced technologies that show great promise.

In Chapter 7 Dr. Norm J. Dovichi discusses recent developments in capillary gel electrophoresis. The last decade has witnessed major improvements in gel technology, speed, and efficiency of separation, miniaturization, and detection. The early manual Sanger and Maxam-Gilbert sequencing protocols yielded little more than a few hundred bases of raw sequence per day. The use of fluorescent labels, parallel capillary arrays, and laser-based detection has led the way to dramatic increases in sequencing rates. Modern technologies have given an old separation technique a significant new life: Rates of several million bases of raw sequence per day are now imminent.

Drs. Dave P. Allison, Thomas G. Thundat, and Robert J. Warmack at Oak Ridge National Laboratory (ORNL) review the application of scanning microscopy techniques to mapping and sequencing DNA. Both atomic force microscopy (AFM) and scanning tunneling microscopy (STM) have been used successfully to image DNA strands. Current research focuses on increasing the image quality and developing techniques to read actual nucleotide sequences.

Miniaturization presents some of the most exciting possibilities for genome characterization. The marriage of semiconductor microfabrication manufacturing techniques with microchannel fluid devices and chemistry on solid supports brings the possibilities massively parallel and dramatic speeds we have come to associate with computers to DNA analysis. Dr. M. Allen Northrup and co-workers of Lawrence Livermore National Laboratory work at this molecular biology–semiconductor technology interface.

Drs. Mitchel J. Doktycz and Kenneth L. Beattie present another implementation of semiconductor technology in their work with "DNA-chips." These devices, or "genosensors," consist of individually addressable hybridization targets covalently linked to silicon wafers in high-density arrays. Massively parallel experimentation is achievable with these systems. Their application to DNA sequencing and DNA diagnostics is discussed.

Dr. Peter Williams and coworkers at Arizona State University give an overview of the application of mass spectrometry to DNA characterization. The recent development of electrospray ionization (ESI) and matrix-assisted laser desorption-ionization (MALDI) techniques have permitted the study of large biomolecules without excessive fragmentation. Mass spectrometry has an inherent speed advantage over gels and can be used where short, highly accurate sequences need to be determined. The techniques they describe form an important emerging suite of technologies complementary to gels for sequence determination.

The final section of this book describes work in the areas of simulation and data management. Simulation studies are very valuable in designing a system. Significant improvements can be made at this stage to process, system, and machine design. No automation effort is complete without consideration and design of the databases and the uses they serve. Many are designed after the fact and, if poorly designed, can negate the entire benefit of automated systems.

Dr. Dierdre Meldrum and Emerson Tongco of the University of Washington describe the use of Petri nets in simulation of the "genome factory." The analogy of agents, parts, operations by the agents on parts, and process conditions or states closely mirrors industrial manufacturing operations. The graphical representation of the Petri net model makes it relatively easy to simulate on a computer, to study the performance of the system without first reducing it to hardware, and also to translate the model to control code for the "genome factory."

Dr. Chris Fields of the National Center for Genome Resources addresses the information management issues of the HGP. The traditional style of information management typical of small laboratories is entirely inappropriate to production laboratories in which high throughput, efficient division of labor, and cost effectiveness are central concerns. Published results from genome laboratories are not complete works; they are starting points for further work to be carried out by others. The working and intermediate nature of these scientific results has raised the data management issues to the point of funding agency policy. Both the highly automated generation of large data sets and public expectations of the data impose demands for information systems beyond those of the small-scale biology laboratory. Dr. Fields presents a model for database design and engineering and argues eloquently for these systems to be designed during the conceptual phase of constructing genome factories.

These authors who have contributed to this book have formulated a vision of the future. This vision is based on weaving together the best technologies other discplines have to offer and the enormous advances these new technologies and methods have to offer to the science of biology. Through their creative energies they have made that vision a reality.

It is with great pleasure that I have worked with the authors who describe their work in this book. I have learned much from their contributions and anticipate that my experience will be shared by others.

TONY J. BEUGELSDIJK

Automation Technologies for Genome Characterization

Contents