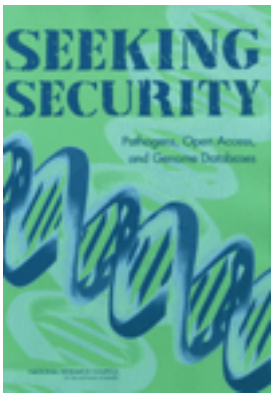

Seeking Security: Pathogens, Open Access, and Genome Databases

Committee on Genomics Databases for Bioterrorism Threat Agents, National Research Council

ISBN: 0-309-54544-7, 88 pages, 6 x 9, (2004)

SEEKING SECURITY

Pathogens, Open Access, and Genome Databases

Committee on Genomics Databases for Bioterrorism Threat Agents

Board on Life Sciences

Division on Earth and Life Studies

Policy and Global Affairs Division

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This study was supported by Contract No. DBI-0314614 between the National Academy of Sciences and the National Science Foundation. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

International Standard Book Number 0-309-09305-8 (Book)
International Standard Book Number 0-309-54544-7 (PDF)

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, NW, Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu.

Printed in the United States of America

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Bruce M. Alberts is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Wm. A. Wulf is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Bruce M. Alberts and Dr. Wm. A. Wulf are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

COMMITTEE ON GENOMICS DATABASES FOR BIOTERRORISM THREAT AGENTS

STANLEY FALKOW (Chair), Stanford University, Palo Alto, California
CORRIE BROWN, University of Georgia, Athens, Georgia
DAVID R. FRANZ, Midwest Research Institute, Frederick, Maryland
CLAIRE M. FRASER, The Institute for Genomic Research, Rockville, Maryland
PAUL KEIM, Northern Arizona University, Flagstaff, Arizona
TERENCE TAYLOR, International Institute for Strategic Studies, Washington, DC

Staff

KERRY BRENNER, Study Director
EILEEN CHOFFNES, Senior Program Officer
SETH STRONGIN, Senior Program Assistant
ROBERT TAYLOR, Writer
NORMAN GROSSBLATT, Senior Editor

BOARD ON LIFE SCIENCES

Preface

This committee was empaneled by the National Research Council to analyze the scientific issues that might accompany the release into the public domain of genome sequences for infectious agents with potential national security implications. Specifically, the committee was asked to consider the following: What are the categories of genome data that should be of greatest concern? What are the pros and cons of unlimited vs. restricted access to the scientific information? What are some options for making decisions about the release of this information into the public domain?

In an ideal world, it would be easy to advocate for a free and ready distribution of all genome information into the public domain. That would be in the spirit of free scientific inquiry as it would lead to the most scholarly and creative use of the information that is inherent (although not always obvious) in deciphering the genomic blueprint of any living thing. However, we live in a world where a small minority of individuals and, sadly, perhaps even some world governments might use pathogenic microbes as weapons. We have to ask to what extent genome information, particularly of microorganisms and their hosts, might help these misguided individuals.

Biologists, unlike physicists, haven't yet formulated many natural laws, and I am sorry to report that there is no sure pathway to making an effective vaccine, isolating new effective anti-infective compounds, or indeed understanding what makes a pathogen a pathogen. I have tried for 50 years and can attest to the fact that, even when the microbial genes

that are essential for pathogenicity are known, we more often than not don't understand their function, nor do we yet understand the underpinnings of susceptibility to or risk of infection of humans or other hosts. And we don't know why we so often fail to develop sterilizing immunity to infections such as HIV/AIDS or to all the other persistent infections that plague (no pun intended) humans.

Thus, with or without the availability of sophisticated biological research tools like genome sequences, we continue to face the potential for catastrophic epidemics due to naturally occurring organisms; and these, like the intentional release of infectious agents of bioterrorism, are probably not predictable or preventable. We need to push ahead to conquer global infectious diseases because they remain the greatest cause of worldwide death and suffering. Yet, unquestionably we face a dilemma that there will be a future time, and it is coming closer and closer, when in the wrong hands biotechnology making use of genome information could create a novel pathogen with unique properties.

It is useful to consider that on the day of our committee's workshop a parallel National Academies meeting was held dealing with the coronavirus that is the agent of SARS. SARS must be close to the perfect example of the dilemma that faces our committee, the scientific community, and anyone else concerned with policy and national security. The virus has a potential to cause a greater morbidity and mortality than the pandemic influenza A strain of the World War I era. The virus has been isolated. Its sequence was promptly published in the public domain, and dozens of companies and laboratories throughout the world are in the process of developing diagnostic reagents and proposed vaccines and are seeking to uncover the pathogenic mechanism with sophisticated contemporary research methods. It is not clear that a vaccine will work. It is not clear whether SARS will return. Sequence availability or not, we don't understand why the influenza A strain of the 1918 era was so virulent in young people, nor do we understand why the SARS virus caused such severe disease in people compared with the virus that supposedly came from a civet-like host in China.

How do we apply criteria to determine what is legitimate research or what is sensitive information or what can and cannot be published? It could be argued that the availability of the complete genome sequence of human isolates of SARS could be used by a very sophisticated bioterrorist as a pathway to synthesize a new version of the SARS virus. Which do we fear more, nature or bioterrorism? Would we gain anything by restricting access to the SARS viral sequence? Would there be any gain or loss in restricting release of later sequences? Should we restrict the sharing of information on the genes or motifs associated with host range or the induction of immunity? Is this information likely to be used by fanatics

who would employ the tactic of bioterrorism? Which answers best serve the global good?

That example is an encapsulation of the problem that confronted us as we began our study. There is no hiding the fact that ours is an immensely difficult task, and I suspect that some of the participants in our workshop took the challenge of speaking to us because it is an immensely difficult task. With this report, the committee has attempted to provide answers as well as we could.

Stanley Falkow,
Chair

Acknowledgments

This study was sponsored by the National Science Foundation, the National Institutes of Health, the Central Intelligence Agency, and the Department of Homeland Security.

The input of workshop speakers and participants was of particular value to the committee in its deliberations. The committee thanks them for taking the time to share their expertise. A list of those who attended the workshop can be found in Appendix C.

This report has been reviewed in draft form by persons chosen for their diverse perspectives and technical expertise in accordance with procedures approved by the National Research Council Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards of objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following for their review of this report:

Michael A. Apicella, University of Iowa, College of Medicine
Enriqueta C. Bond, Burroughs Wellcome Fund
Joe DeRisi, University of California, San Francisco
Gerald Epstein, Center for Strategic and International Studies
David Galas, Keck Graduate Institute
Gerald Mandell, University of Virginia

Randall S. Murch, Institute for Defense Analyses
Eugene W. Myers, University of California, Berkeley
David Relman, Stanford University
James Tiedje, Michigan State University

Although the reviewers listed above have provided constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations, nor did they see the final draft of the report before its release. The review of this report was overseen by Harold J. Fallon, University of Alabama School of Medicine (emeritus) and May Berenbaum, University of Illinois. Appointed by the National Research Council, they were responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the author committee and the institution.

Contents

Executive Summary

THE LIFE-SCIENCE REVOLUTION AND THE DUAL-USE DILEMMA

The life-science revolution that began with deciphering the genetic code has launched biological research into an unprecedented period of productivity. Parallel advances in computational techniques and the widespread use of global computer networks have contributed to the pace of biological research. Within less than 30 years, the entire genomes of many hundreds of organisms, from viruses to bacteria to humans, have been sequenced, and partial sequences from many thousands more organisms have been deposited into databases freely accessible to scientists around the world.

Modern biological research is a thriving international enterprise with enormous potential to benefit society. The synergy created by increasing knowledge and open exchange of ideas and information is accelerating the advance of medicine, industry, and agriculture. Emerging details about the interplay between pathogenic microorganisms and their hosts will allow scientists to continue to develop and deliver new and improved vaccines, stronger infection-fighting drugs, and more-precise diagnostic tools.

However, with its promise, biological research presents a "dual-use" dilemma, in that its technologic advances could also be applied for destructive purposes in acts of bioterrorism or war. Results that have immediate implications for pathogen enhancement or weapons development have been called "contentious research" (Epstein, 2001) or are said to fall

into a gray zone where the benefits of publication may not outweigh the dangers. Any scientist working to develop new treatments for naturally occurring infectious diseases can tap the power of genomics and its globally accessible databases and analytic tools, but so could a malefactor trying to engineer enhanced pathogens for use as biological weapons. Hence, scientists and policy-makers are confronted with the challenging question of how to mitigate the risk of bioterrorism and still foster the research community's ability to counter current and future biological threats, whether naturally occurring or malevolently deployed.

GENOME RESEARCH IN AN AGE OF TERRORISM

The attacks on September 11, 2001, and the later deadly anthrax letters have focused increased national and international attention on the threat of terrorism. On October 8, 2003, the National Academies released a report, *Biotechnology Research in an Age of Terrorism* (NRC, 2003a), which examined the dual-use problem in life-science research. The author committee, chaired by Gerald Fink of the Whitehead Institute, offered recommendations on how to confront the potential for misuse of biological agents and technologies without unduly limiting progress in the life sciences. The report proposed modifications of the system of review of biological experiments and stressed the importance of addressing research in subjects of concern early and of educating scientists to be aware of the risks and benefits associated with their research and how to balance them responsibly. The committee recognized the importance of open communication in scientific research as a fundamental practice crucial to continued progress despite the fact that it might make the data accessible to those intent on misuse. A reliance "on self-governance by scientists and scientific journals to review publications for their potential national security risks" was recommended, and a number of major journals that publish life-science research have already committed to implementing such a review process (Atlas et al., 2003a,b,c).

Genome data, the focus of this report, occupy a unique position in the dual-use dilemma in that they are a source of raw material that, although not inherently dangerous, can be enabling for potentially destructive agendas. Furthermore, the culture of genomics is unique in its evolution into a global web of tools and information. The major Internet-based data repositories have policies that mandate free, unfettered, and anonymous access, and most scientific journals require that genome data be deposited into accessible databases as a prerequisite for publication. With the exception of rare cases in which information is classified for national security purposes, the U.S. government itself requires that data, including genome data, resulting from federally funded research be made publicly avail-

able. The committee did not address mechanisms used to determine whether or not information is classified.

CHARGE TO THE COMMITTEE

Members of the National Interagency Genomics Sciences Coordinating Committee (NIGSCC), which comprises representatives of several federal agencies that have an interest in genome research, had discussed the release to the public domain of genome data as they pertain to likely agents of bioterrorism. Given that complete genomes of more than 100 microbial pathogens—including those for smallpox, anthrax, Ebola hemorrhagic fever, botulism, and plague—are already in Internet-accessible databases freely open to all and that the genomes of hundreds more pathogens will be sequenced with the support of government funds in the next few years (Fraser, 2004), representatives of the member agencies discussed whether current policies regarding release of genome sequence data were appropriate. As a result of the discussions, some NIGSCC members decided to seek advice from the scientific community. The National Science Foundation, the National Institutes of Health, the Department of Homeland Security, and the Central Intelligence Agency funded the National Academies to convene a committee, to hold a workshop, and to produce a report about how biological scientists view the potential for misuse of genome sequence data and the policies governing access to databases containing these data.

At the first meeting of the Committee on Genomics Databases for Bioterrorism Threat Agents, the sponsors indicated that they hoped the report would present the perspective of working biological scientists, so that readers in the policy and intelligence communities could use the report when considering potential changes in policy regarding access to genome sequence data. It was understood that the security community would then take this scientific perspective and use it in combination with their own knowledge of security issues to make decisions. The sponsors specifically requested that the report capture input from workshop participants' presentations and discussions, identify general issues surrounding the release to the public domain of genome data for bioterrorism threat agents, develop a list of pros and cons associated with the release to the public domain of such data, and present recommendations for policy options and decision-making frameworks concerning release to the public domain of genome information.[1]

[1]The full charge to the committee, the statement of task, can be found in Appendix A.

The National Academies committee organized a 1-day workshop on the public release of genome data on bioterrorism-threat agents, which was held in Washington, DC, on October 1, 2003. About 40 invited scientists and policy experts who work in government, private industry, and academic laboratories attended. Workshop participants were asked to address three questions concerning genome data for possible biological weapons agents:

- What categories of genome data present the greatest concern?
- What are the pros and cons of unlimited vs. restricted access to such data, including threats posed to the scientific community or to national security?
- What are some options for making decisions about release to the public domain?

The workshop agenda and a list of the participants are appended to this report. Although the questions posed to the committee were limited to consideration of genome sequences of bioterrorism-threat agents, these were by no means the only kind of data that workshop participants discussed. The broader context is complex, and there is no clear demarcation between bioterror-agent genome sequences and other genome data, gene-expression data, protein structures, and other kinds of research results. The key advances in modern life science are not readily apparent in any particular piece of genome data. Instead, the growing set of full-length sequences of many organisms can be thought of as "raw material" for modern biological research or as the platform from which research can be launched. Data on one organism often prove to be invaluable for building a better understanding of other organisms, and data from many organisms taken together and compared, analyzed, and applied to new questions will allow new and fundamental insights into biological processes.

GENOME DATABASES TODAY

At the workshop, presentations described genome databases and how they are used to advance research in the life sciences. This report describes two recent success stories—the rapid international response to the 2003 outbreak of severe acute respiratory syndrome (SARS) and the creation of meningococcus B vaccine candidates—that illustrate the power of genomics and openly accessible databases to help improve our understanding of and aid in the development of countermeasures for infectious diseases. The report also considers how genome data and related technologies might be misused for the development of genetically enhanced biological weapons, and it discusses potential malefactors. As access to the knowl-

edge and resources necessary to engineer microorganisms grows, the ability to manipulate pathogen genomes will be far more widespread and accessible than it is today. Such techniques could be used to produce advanced biological agents that are more dangerous, or easier to use, than naturally occurring agents. Although the technical hurdles that would confront a bioterrorist intending to deploy a naturally occurring agent to cause large numbers of casualties are substantial, they are much lower than those associated with enhancing the virulence of a known pathogenic species with genetic manipulation. Thus, an attack with a natural pathogen is more likely; however, given the developments in biotechnology described in this report, a more sophisticated attack with an engineered pathogen is a serious concern.

ISSUES IN THE CONTROL OF GENOME INFORMATION

The committee members and workshop participants discussed a variety of issues as they asked whether and to what degree access to pathogen-related genome data should be restricted. They identified the major domestic stakeholders as the scientific community, the security community, and the general public, and they considered the interests and positions of these groups carefully. The effectiveness of any policy depends on international consensus because databases are globally accessible. The position of the international community and the potential political implications of restrictions imposed by the United States were also taken into account.

The committee was charged with determining which types of pathogen-related genome data present the most concern. Biological agents discussed at the workshop included those on national "select agent" lists and those which could become plausible threats in the future. Genome data from sources other than pathogenic microorganisms were also included, inasmuch as insights about infection processes can come from studying a pathogen's hosts or nonpathogenic relatives. With the input of workshop participants, committee members discussed whether it was possible to categorize data usefully on the basis of whether they might be misused for bioterrorist purposes. Categories of data that were discussed include primary genome sequences, annotated and analyzed sequences, sequences from select agents, and sequences from engineered microorganisms. In further discussions after the workshop, the committee concluded that assigning data to one of those categories would not be a significant help in determining risks. They reached that conclusion in large part because of the ways that information from one category can inform studies in another category, such as when comparisons are made between closely related organisms. Data on all organisms present some level of

concern; although some organisms are inherently more dangerous, it does not necessarily follow that the genome sequences of the organisms are more dangerous. The organisms themselves are beyond the scope of this study, and many organisms relevant here are governed by the select-agent rules.

Workshop participants also discussed the idea of a gray zone, or a field of contentious research. In a 2001 publication, Gerald Epstein (2001) described contentious research as containing "fundamental biological or biomedical investigations that produce organisms or knowledge that could have immediate weapons implications and that therefore raise questions concerning whether and how that research should be conducted and disseminated." The conduct of such contentious research is beyond the charge to this committee, but the dissemination of the results falls within our purview.

Workshop participants and committee members also considered possible mechanisms for controlling access to data. Data could be designated as *classified* so that they would be withheld from people who do not possess a government-issued security clearance. Alternatively, data could be withheld from widespread public release by another mechanism, such as a new screening process that would provide access to those deemed authorized. A third possibility would be to require registration for database access but not to impose any restrictions on who could register; this alternative could provide an opportunity to track database users. Finally, current policies of free access without a requirement for registration could be maintained.

ADVANTAGES AND DISADVANTAGES OF RESTRICTING ACCESS

The committee and workshop participants weighed the possible advantages and disadvantages of the various ways of restricting access to genome data. They considered the potential to thwart efforts to develop genetically engineered bioweapons but recognized that the genome data most likely to be restricted are also the data most relevant to the development of countermeasures and treatments for naturally occurring or engineered pathogens. They noted that restricting access might ease public concern and increase public confidence in the scientific community's willingness to confront the dual-use dilemma responsibly. However, an open-access policy also has great benefit in that it allows all scientists the opportunity to collaborate and to use all possible information to scrutinize and verify results and conclusions. Given the numerous interconnections between different topics related to life-science research, it is not possible to predict which scientists will benefit from access to which data; this makes restricting access all the more tricky to implement. The group also discussed practical issues that would surround the development and imple-

mentation of a policy to restrict access to genome data, such as who would decide which scientists would have access to which information and whether there are realistic ways to contain digital data.

CONCLUSIONS AND RECOMMENDATIONS

The presentations and discussions at the workshop and their own research has led the committee members to make the following recommendations. In preparing these recommendations, the committee considered the reality that advances and technologies of life-science research could potentially be misused by individuals, groups, or nations to create agents capable of causing great harm. However, given that society has reason to fear natural outbreaks and intentional attacks, the committee concludes that biosecurity would be better served by policies that facilitate, not restrict, scientists' ability to understand infectious disease and to develop countermeasures to both naturally occurring pathogens and biodefense threats.

Recommendation 1: Policies with regard to release of genome data on microbial pathogens should not change. Rapid, unrestricted public access to primary genome sequence data, annotations of genome data, genome databases, and Internet-based tools for genome analysis should be encouraged.

With a growing understanding of microbial pathogens and their interactions with the hosts they infect, national governments, subnational groups, or single individuals could attempt to apply such knowledge to destructive purposes and with potentially grave consequences. However, after careful deliberation, the committee concluded that preserving open access to genome data and free exchange of knowledge and ideas that flow from the data will facilitate scientific and medical advances that will improve health and society's ability to react to biological threats. That conclusion is supported by the following arguments.

Current Policies Are Effective

Unfettered, free access to the results of life-science research is the historic norm and has served science and society remarkably well. Open access allows life scientists everywhere to evaluate, interpret, adapt, and extend results from many fields of inquiry for use in their own work and thereby accelerates research and speeds the delivery of life-saving benefits that biological and medical research are so rapidly creating. Current policies allow for the most rapid and effective scientific response possible during an infectious-disease crisis, such as the SARS outbreak of 2003. At

such times, when scientific and public-health resources must be rapidly mobilized to combat a poorly understood emerging disease, free and rapid exchange of data, results, and ideas is essential to allow scientists to communicate effectively and to build on one another's findings.

Effective Restriction of Genome Data Is Not Practical

As a practical matter, restricting access to genome data would be difficult, expensive, and probably counterproductive. It is notoriously difficult to control access to digital data, and files that contain entire genomes are not particularly large and therefore are easily stored, transferred, and exchanged. Also, in the absence of a uniform international agreement to impose similar control measures worldwide, potential users who are denied access because of U.S. policy could direct their Internet browsers to genomics sites in other countries that have the same kind of data. In addition, any policy stringent enough to reduce the chance that a malefactor would access data would probably also impede legitimate scientists' use of the data and would therefore slow discovery and limit the vitality of the life sciences.

At the outset of the workshop, the concept of requiring all users of genome databases to register to gain access seemed to many participants to be a reasonable policy compromise. Under such a policy, anyone could gain access but only after stating a name, address, and institutional affiliation. After additional discussion, however, the committee concluded that a registration requirement of this kind would not be an effective way of protecting society from bioterrorism. Registration would not prevent a determined malefactor from accessing genome databases. Although registration might deter a less determined malefactor or provide a mechanism for tracing his or her activities, it would also raise many troubling questions about who could use registration information and under what circumstances. In addition, the lack of an international consensus that registration should be required would render such measures futile. It seems unlikely that a uniform agreement could be generated between all public and private database managers and others who generate genome data, which would be necessary to track those with access to genome sequence. Downloading by pharmaceutical companies, large research centers, and others of the available data onto their own networks so that they can be used privately would hinder the usefulness of attempts to track discrete queries to databases. Many of the data have been in the public domain for years and may well be stored in dozens or even hundreds of locations around the world. Given the international availability of the data, many people could access sequence information without relying on a database that requires registration. For all the above reasons, the committee feels

that it is not appropriate to implement a system of registration for the use of genome databases.

Pathogen Genome Sequences Are Not Uniquely Dangerous

Primary sequence data on pathogens become dangerous only if the user has a sophisticated ability to exploit them and a malevolent goal. Mere possession of the sequence of a pathogen does not confer the ability to enhance the virulence of the organism to which it pertains, nor would it help to solve the demanding technical problems associated with conducting a terrorist attack. Although a potential malefactor might be able to adapt published research results that reveal genetic manipulations that would enhance the virulence of a pathogen, discovering which genetic change would enhance virulence is difficult and would require a substantial and sophisticated effort.

The workshop participants considered what categories of genome data present the greatest concern, these categories are described in Chapter 3. The committee did not see evidence that identifying data as belonging to any one of these category would necessarily make them a greater threat. It is important to remember that the focus here is on access to *data* pertaining to organisms, not on access to the organisms themselves; for example, U.S. government regulations on select agents apply to the possession of the *organisms* and not to their genome sequences.

There are many reasons why it is difficult to categorize genome data by risk. First, the study of nonpathogenic microorganisms is often closely related to the study of pathogenic species. The ubiquitous soil bacterium *Bacillus cereus*, for example, is closely related to *Bacillus anthracis*, the bacterium that causes anthrax; insights gained from the genome of one have been directly applicable to the other (Parkhill and Berry, 2003). Second, biological-weapons developers and those studying ways to counter biological weapons both use model strains to simulate real agents so that they can do development work and trials more safely. One classical model of anthrax is the insect pathogen *Bacillus thuringiensis*, which is widely used as a microbial pesticide. It could be argued that knowledge of its genome would be beneficial to a malefactor hoping to genetically enhance *B. anthracis*. Third, data derived from a single microbial species are not the only data relevant to understanding it. Instead, the ability to compare genes, genetic control mechanisms, and protein function among the entire growing and diverse catalog of completely sequenced microbial genomes is what drives many current research efforts (Frazer et al., 2003; Kanehisa and Bork, 2003). Such comparisons among species have already proved to be a productive approach to deciphering how pathogenic and nonpathogenic species function as complex biological systems. Fourth, genome data

that help scientists to clarify how pathogenic microorganisms cause disease are by no means limited to microorganisms. Human gene sequences and sequences from other "host" species are crucial data for those seeking to understand the intricacies of the interactions between the immune system and microbial pathogens, including specific immune mechanisms and vulnerabilities. The gene sequences of humans and other host species and the insights derived from them therefore would be crucial "enabling data" both for those who would work to find new ways to defeat pathogens and for those who might hope to modify pathogens to exploit immune vulnerabilities and create pathogens with unusual or particularly destructive properties.

The committee was charged with determining which types of pathogen-related genome data present the most concern. As described in the report, it is possible to identify categories of data, but it is not clear that types of data can be correlated with a specific risk of misuse for bioterrorist purposes. Data on all organisms present some level of concern but, although some organisms are inherently more dangerous, it does not necessarily follow that their genome sequences are more dangerous. The organisms themselves are beyond the scope of this study, and research on many of the organisms relevant here is governed by the select agent rules.

For the most part, the issues are the same for genetically engineered organisms as for naturally occurring organisms; information on the altered sequences and the resulting phenotypes can provide insight into basic biology, and most alterations are not particularly useful to a potential bioterrorist. However, sequence data from some genetically engineered organisms could be very useful for a potential bioterrorist attempting to create a more dangerous pathogen. Regulations on the actual conduct of the experiments that might generate such an engineered organism are beyond the charge to this committee, although it is certainly an important issue. Decisions on the appropriateness of conducting particular experiments should ideally be made before the experiments are begun. Local institutional review boards (IRBs) play a large role in that process, and the newly announced National Science Advisory Board for Biosecurity (NSABB), discussed in detail in Recommendation 3, will play a growing role. The guidelines for IRBs and codes of conduct for individual scientists that the NSABB envisions should help to ensure that appropriate consideration is given to the potential implications of research approaches before they are begun. In addition, journal editors have a responsibility to consider carefully the national-security implications of the papers they publish. Given all those caveats, if an experiment is published, the accompanying genome data should not be restricted by regulations. The data are essential for others to understand the significance of the research and may be crucial to future experiments that could help protect us from dis-

ease. In addition, there is some concern that restricting access to this information might lead to a situation in which the mainstream scientific community is unaware of the potential dangers that may threaten us, and some have proposed that observing changes in the frequency of publications (and conference presentations) of potential malefactors can provide useful clues as to whether they are conducting secret experiments.

For all those reasons, the committee concluded that maintaining the current standard of free access to all genome data is the best policy choice. The problem with which the committee has been charged is not to strike the correct balance between security and openness; that is a false dichotomy—openness has enhanced security in the past and is the best way to ensure security in the future. Instead, the most important task is to be as well prepared as possible to cope with the serious infectious-disease threats that society is sure to face in the coming century, both natural and human-made. The committee believes firmly that the policies currently in place for genome data—immediate release and free access—are correct because openness is essential to maintain the progress needed to stay ahead of those who would attempt to cause harm.

Recommendation 2: Genomics and genome sequence data should be exploited fully to improve our ability to defend against infectious agents of all types, including those which contribute to epidemic diseases and infant mortality and the naturally occurring or genetically enhanced organisms that could be used in a bioterrorist attack.

Since the terrorist attacks of 2001, federal spending intended to improve defenses against bioterrorism and natural infectious-disease outbreaks has increased markedly. Indeed, many of the pathogen whole-genome sequencing efforts that have been recently completed or begun have been funded with money earmarked for biodefense. Research exploiting the revolution in genomics has an important role to play in increasing our ability to defend against infectious agents of importance to biodefense and in global infectious disease. Indeed, research on many of the currently important societal infectious threats, such as antibiotic-resistant bacterial pneumonia and antibiotic-resistant staphylococcal disease, will benefit enormously from the genome revolution. Extensive sequence comparisons between pathogenic and nonpathogenic organisms, studies of changes in the pattern of gene expression in pathogens and their hosts as they interact, and sequencing of multiple strains of specific pathogens will all contribute to the development of new diagnostics, vaccines, and therapeutics for disease-causing organisms, including those which might be used in a bioterror attack. Infectious agents that plague agricultural crops and livestock are of critical importance for our economy and our national secu-

rity. The biodefense effort should include both human pathogens and those which might be deployed against agricultural interests.[2]

Recommendation 3: Future advances in genome science should be regularly reviewed to keep all relevant government departments and agencies apprised of new developments that may affect national security. Regular meetings of scientific and security experts should be held to discuss the implications of new developments and to develop coherent responses. The newly formed National Science Advisory Board for Biosecurity or another appropriate entity with the ability to connect with diverse federal agencies would be a suitable home for that function.

The pace of scientific progress creates a need for continuous and thorough evaluation of scientific technology as it affects national security and the health and welfare of all the inhabitants of this planet. Decisions about pathogen genomes cannot be properly made unless they are considered in the context of other scientific advances. New developments in law enforcement, forensics, and public health based on continued research may provide better approaches to improving biosecurity than attempts to restrict access to genome data. However, a mechanism is needed to ensure adequate communication between the scientific and security communities. A well-informed body with both scientific and security expertise should review advances in genome science in case future developments warrant the creation of additional monitoring of or restrictions on access to genome data. Review should be scientifically broad because the effect of genomics on biosecurity goes far beyond the biology of biothreat organisms and includes both biomedical topics, such as drug and vaccine development, and topics pertaining to forensics, intelligence, agriculture, and the environment. Limiting the evaluation to direct studies on genomes of pathogens would not adequately address threats to biosecurity.

Knowledge of the genomes of infectious agents that might be used as weapons of bioterror is obviously important, but the genomes of potential hosts (humans, other animals, and plants) also offer opportunities for manipulation. Over the next 10 years, scientists may learn at least as much about the molecular basis of genetic resistance and susceptibility to infection as about specific microbial virulence factors and their function. The perspective of those involved in basic research related to humans, plants, animals, and microorganisms is essential for staying on top of new developments that may affect biosecurity. Continuing review of new technology

[2]For more information about biodefense and agriculture, see the 2003 National Research Council report *Countering Agricultural Bioterrorism*.

could include the use of functional genomics as it pertains to understanding microbial virulence; host susceptibility and resistance to infectious diseases of plants, domestic animals, and humans; and relevant aspects of the development of new drugs, vaccines, and anti-infective therapies.

To be well informed, the reviewing body must be part of a network for information exchange among academe, industry, international actors, and U.S. government agencies, including those in the intelligence and security community. Coordination of efforts in all arenas, including the international community and those involved in assessing and responding to threats, would provide a means of evaluating the significance of advances in genome research in terms of both increased threats to security and improvements in understanding of the environment and of human health and disease. As an additional benefit, providing a network for information exchange would help to further research in disease diagnosis and epidemiologic surveillance on a national and global basis and facilitate communication of information required for the unambiguous identification and attribution of pathogens in forensics.

There are many factors to balance in determining where the proposed reviewing function should be based. One option is the newly proposed NSABB announced by the Department of Health and Human Services on March 4, 2004. The NSABB is asked to "advise all Federal departments and agencies that conduct or support life sciences research that could fall into the dual use category (www.biosecurityboard.gov)." However, it may not be feasible for that group to manage the necessary continuing review of genome information while acting on establishing guidelines for the oversight of biological research. In any event, the partnership and full participation of each of the relevant agencies is crucial to ensure that all the available information and insight are used. The entity that becomes responsible for reviewing scientific advances in genome science for their potential effect on national security must be scientifically respected, have the ability to integrate information from diverse sources, and have a clear ability to influence discussions in numerous federal departments and agencies.

Recommendation 4: The committee endorses Recommendation 7 of *Biotechnology Research in an Age of Terrorism*, which calls for an international forum to unify the discussion on the effect of genomics on biosecurity.

Life-science research is global, and no single nation can successfully implement policy concerning access to and release of life-science data and results without reference to the rest of the international community. For that reason, it is of the utmost importance that the international community establish a common understanding of security concerns and shared

resources in order to make the most efficient and safest use of genome data and experimental results, some of which might suggest how pathogens could be successfully enhanced. The committee therefore strongly endorses Recommendation 7 of *Biotechnology Research in an Age of Terrorism,* which calls for "the international policymaking and scientific communities [to] create an International Forum on Biosecurity to develop and promote harmonized national, regional, and international measures that will provide a counterpart to the system [recommended] for the United States." An international forum to discuss the potential for the misapplication of life-science research should be convened in the near future to serve as a first step toward achieving harmonized international oversight. The forum should include broad representation of all interested countries. If conducted openly and in the proper spirit, the process of discussing these issues might actually build understanding, and some trust, among the nations involved and, eventually, help establish an international norm against misuse of genetic information.

Recommendation 5: The committee endorses Recommendation 1 of *Biotechnology Research in an Age of Terrorism,* which calls for national and international professional societies and related organizations to work to educate scientists about the risk that life-science research results will be misused and about scientists' responsibility to mitigate the risk.

Recommendation 1 of *Biotechnology Research in an Age of Terrorism* calls for "national and international professional societies and related organizations and institutions [to] create programs to educate scientists about the dual-use dilemma in biotechnology and their responsibilities to mitigate its risks." As noted under our Recommendation 1 above, we believe that although the risk that the growing power of biological and medical research could be applied for destructive purposes is unknown, it is not zero. All life scientists must be sensitized to the potential for the harmful misuse of the knowledge they create. The committee recognizes and applauds the efforts to date of numerous professional societies to educate their members and the public about these issues, and it suggests that such professional societies are the natural home for further efforts in this respect. They should expand efforts to engage their members in discussion of the potential benefits and dangers of the widespread availability of genome sequences and functional genomics data. Professional codes of conduct should explicitly require scientists to act to mitigate the risk of misuse of scientific progress to cause environmental or medical harm and require them to carry out their research with integrity to minimize the risk of misuse of life-science research for destructive purposes.

1

Introduction

THE LIFE-SCIENCE REVOLUTION

The ultimate goal of biology, medicine, and other life sciences is to build a complete understanding of the function of all living things, both as discrete molecular components and as integrated complex interactive systems. Until recently, such an ambitious undertaking has been little more than a distant dream. That dream began to take shape in the 1950s and 1960s when the DNA code was deciphered, and its realization accelerated in the 1970s as new tools that were developed to read and manipulate gene sequences increased the pace of discovery many times over (Hood and Galas, 2003).

The revolution in the life sciences that began to take shape in the 20th century is no longer a promise; it is happening now. The major technologies catalyzing this revolution are sequencing of the entire genetic codes of organisms (including humans), mapping of genome variability between individuals of a species, and microarray technology that allows observation and analysis of genome-wide patterns of gene activity under different conditions. All those technologies allow the rapid analysis of the entire repertoire of proteins and other macromolecules produced by a cell, and they have generated very large biological databases and associated analytic software tools. Those and other advances now allow life scientists to assemble the mass of new data into an accurate and detailed way to reach the goal of understanding how organisms function.

All branches of the life sciences have entered a period of unprecedented research productivity, and the pace will only increase. As a result,

the idea of obtaining a reasonably complete understanding of living systems, although still some distance off, is in view (Kanehisa and Bork, 2003; Venter et al., 2003). And just as it was difficult even a decade ago to envision all the changes that the widespread use of computers and global computer networks would bring, it is impossible now to foresee all the effects that the life-sciences revolution will have over the next decade.

The rapidly growing understanding of natural systems has tremendous potential to create better lives for people the world over. For example, understanding fully how pathogens and hosts interact is a major long-term research goal. To reach it, scientists must gain a detailed understanding of what makes the immune response effective and of how pathogens cripple or evade the immune response to cause disease. As more details of the interplay between pathogenic microorganisms and the immune system become known, scientists will probably be able to create new and powerful strategies to fight infection, create better vaccines, and develop faster, more precise diagnostic tools (Moxon and Rappuoli, 2002; Rappuoli and Covacci, 2003). Perhaps scientists will someday be able to deliver those benefits in a matter of days or weeks, so that when new pathogens emerge, treatments and vaccines will become available quickly enough to contain what might otherwise be catastrophic outbreaks of infection and disease. The benefits of the life-science revolution are broad and include treatments and preventive measures for conditions as varied as sudden infant death syndrome, cancer, autoimmune diseases, infectious diseases, and such neurological disorders as Alzheimer's disease. In addition, agriculture, energy production, chemical manufacturing, and even computing all stand to be transformed by the genome revolution.

Of course, such powerful technology can also be used for destructive purposes. This is the "dual-use" problem familiar to those who work on arms-control and disarmament issues: most technologies that are important in the peacetime economy—including communications, cryptography, computers, materials science, aeronautics, and nuclear energy—are also technologies for weapons. The products of life-science research must be included prominently in any list of technologies that can be used for good or ill. Just as fundamental knowledge about how pathogens interact with the immune system will lead to new ways to prevent and cure infections, it could also help someone bent on designing genetically altered versions of natural pathogens that could be exploited as weapons by governments or terrorists. Some types of research has been called "contentious research" (Epstein, 2001) or is said to fall into a gray zone where the benefits of publication may not outweigh the dangers. In a 2001 publication, Gerald Epstein described this category as "fundamental biological or biomedical investigations that produce organisms or knowledge that could have immediate weapons implications and that therefore raise questions

concerning whether and how that research should be conducted and disseminated" (Epstein, 2001). The conduct of such contentious research is beyond the charge to this committee, but the dissemination of the results falls within our purview. Two examples of such potentially contentious research are given later in the report: work with a fungal pathogen of plants and work with a virus of mice.

RELATED NATIONAL ACADEMIES PROJECTS

On January 9, 2003, the National Academies convened a workshop titled "Scientific Openness and National Security." The day-long workshop had sessions on assessing the threat posed by life-science knowledge and current policies related to openness, and four case studies of how "sensitive" information could be handled were discussed. Two members of the committee that wrote the present report participated in that workshop.

On January 10, 2003, a meeting of journal editors was held in Washington, DC. The editors discussed their role in determining which articles are published, including decisions as to what constitutes sensitive or dangerous information and what steps journal editors might take to decrease the chances that published material would facilitate efforts of bioterrorists. These editors later published a joint statement in three journals (*Science*, *Nature*, and the *Proceedings of the National Academy of Science;* Atlas et al., 2003a,b,c). The statement indicated that the scientific review process must be safeguarded and issues of security risks acknowledged. They called for journals to devise appropriate procedures for reviewing security risks and to encourage scientists to communicate their data in ways that minimize risk and maximize benefits.

On October 8, 2003, the National Academies released a report, *Biotechnology Research in an Age of Terrorism,* written by the Committee on Research Standards and Practices to Prevent the Destructive Application of Biotechnology, chaired by Gerald Fink, of the Whitehead Institute in Cambridge, Massachusetts. The report examined the dual-use problem as related to applications of life-science research. The charge to that committee was to "consider ways to minimize threats from biological warfare and bioterrorism without hindering the progress of biotechnology" (NRC, 2003a), and the committee's report identified seven categories of "experiments of concern." They included experiments that would

- Demonstrate how to render a vaccine ineffective.
- Confer resistance to therapeutically useful antibiotics or antiviral agents.

- Enhance the virulence of a pathogen or render a non-pathogen virulent.
- Increase transmissibility of a pathogen.
- Alter the host range of a pathogen.
- Enable the evasion of diagnostic or detection methods.
- Enable the weaponization of a biological agent or toxin.

The same report proposed modifications of the system of review of biological experiments that would address concerns about misuse of results without unduly limiting work in the life sciences. Among several recommendations, the report urged that

- The Department of Health and Human Services (DHHS) expand and augment its system of scientific review to include the consideration of the potential for misuse of results of proposed research.
- Life scientists educate themselves and policy-makers about the kinds of misuse of scientific results that are possible.
- A permanent expert committee be set up in DHHS to provide advice and leadership for the expanded system of review.
- An international forum be convened to attempt to harmonize policies on dangerous life-science research results around the world.

CHARGE TO THE COMMITTEE

Discussions among members of the National Interagency Genomics Sciences Coordinating Committee (NIGSCC), which comprises representatives of several federal agencies that have an interest in genome research, had been held on the topic of the release to the public domain of genome data as it pertains to likely agents of bioterrorism. Given that complete genomes of more than 100 microbial pathogens—including those for smallpox, anthrax, Ebola hemorrhagic fever, botulism, and plague—are already in Internet-accessible databases freely open to all and that the genomes of hundreds more pathogens will be sequenced with the support of government funds in the next few years (Fraser, 2004), representatives of those agencies discussed whether current policies regarding release of genome sequence data were appropriate. As a result of the discussions, some NIGSCC members decided to seek advice from the scientific community. The National Science Foundation, the National Institutes of Health, the Department of Homeland Security, and the Central Intelligence Agency funded the National Academies to convene a committee, to hold a workshop, and to produce a report about how biological scientists view the potential for misuse of genome sequence data and the policies governing access to databases that contain them.

At the first meeting of the committee, the sponsors indicated that they hoped the report would present the perspective of working biological scientists so that readers in the policy and intelligence communities could use the report when considering potential changes in policy regarding access to genome sequence data. The sponsors specifically requested that the report capture input from presentations and discussions by workshop participants, identify general issues surrounding the publication of genome data on bioterrorism-threat agents, develop a list of pros and cons associated with the release to the public domain of such data, and present recommendations for policy options and decision-making frameworks concerning release to the public domain of genome information.

The National Academies Committee on Genomics Databases for Bioterrorism Threat Agents organized a 1-day workshop on the public release of genome data on bioterrorism-threat agents, which was held in Washington, DC, on October 1, 2003. About 40 invited scientists and policy experts who work in government, private industry, and academic laboratories attended. Workshop participants were asked to address three questions concerning genome data on possible biological-weapons agents:

- What categories of genome data present the greatest concern?
- What are the pros and cons of unlimited vs. restricted access to such data, including threats posed to the scientific community or to national security?
- What are some options for making decisions about release to the public domain?

The genome data considered at the workshop included not only raw DNA sequences but also annotated sequences and interpretations of sequence data (for example, identification of protein motifs or functional genomics data). Proteome data (for example, data on protein expression patterns) was also be considered by the participants. Various venues for publication of such data (such as deposition into electronic banks and publication in mainstream journals) were considered, as well as possible mechanisms to constrain access to data.

The workshop agenda and a list of the participants are appended to this report. Although the questions posed to the committee were limited to consideration of genome sequences of bioterrorism-threat agents, these were by no means the only kind of data that workshop participants discussed. The broader context is complex, and there is no clear demarcation between bioterror-agent genome sequences and other genome data, gene-expression data, protein structures, and other kinds of research results. The key advances in modern life science are not readily apparent in any particular piece of genome data. Instead, the growing set of full-length

sequences of many organisms can be thought of as "raw material" for modern biological research or as the platform from which research can be launched. Data on one organism often prove to be invaluable for building a better understanding of other organisms, and data from many organisms taken together and compared, analyzed, and applied to new questions will allow new and fundamental insights into biological processes.

2

Genome Databases Today

GENOME DATABASES

Genome Sequencing

An organism's genome is the sum of its entire genetic potential, stored as an encoded sequence of the nucleotides adenine, thymine, guanine, and cytosine (A, T, G, and C) that make up its nucleic acids. Bacteria are prokaryotic: they have no organized nucleus, and most of their genes—units of heredity—are in a single large circular chromosome floating free in the cell, although some do have multiple circular chromosomes, and a few have linear chromosomes. Smaller loops of extrachromosomal DNA called plasmids can also be present. Plasmids can be passed between cells, and the instructions they contain can allow bacteria to quickly acquire properties they would not otherwise have, such as resistance to various antibiotics. The cells of more complex organisms, the eukaryotes, store most of their genomic DNA on tightly organized paired chromosomes in a membrane-bound nucleus. A few genes in eukaryotes are outside the chromosomes, in energy-processing cellular organelles called mitochondria or in chloroplasts. Viruses, which contain relatively few genes in their genomes, are parasites that pirate prokaryotic or eukaryotic cells' replication and protein-synthesis machinery to reproduce.

Some genes encode proteins. The cells use the DNA sequences in these genes to make the corresponding sequences of amino acids. The amino acid sequences in turn determine the proteins' structures and functions, which can be structural or functional components of cells or used to cata-

lyze the production of virtually every other building block of life. Some genes are regulatory and are involved in controlling the activity of other genes. The environmental and chemical sensing mechanisms that regulate gene activity are extraordinarily complex and are the target of a great deal of research.

The fundamental principles used today to sequence DNA were developed in the middle 1970s. The speed with which sequencing can be carried out has increased exponentially and has been largely driven by the development of automated sequencing machines and new technologies. The per-nucleotide cost of sequencing has similarly decreased; it fell by about 2 orders of magnitude between 1998 and 2003 and reached about 2 cents per nucleotide by early 2004. The increase in speed and decrease in cost are expected to continue in much the same way that the power and cost of computer processing chips have changed over the years (Carlson, 2003) (see Figure 1). Indeed, the power of sequencing technology is now

FIGURE 1 Productivity improvements in DNA synthesis and sequencing. SOURCE: Reprinted, with permission, from Carlson, 2003.

such that obtaining sequence data is no longer considered to be research, but merely a routine technical procedure carried out in the course of research, largely with fully automated and easy-to-use equipment operated by technicians. Many laboratories own automated sequencers or use central sequencing facilities in their research institutions for routine sequencing tasks. Others contract the work out to private companies or sequencing centers around the world.

The first complete genome sequence of a virus was determined in 1975 when the 3,569-nucleotide genome of MS2, an RNA virus that infects bacteria, was sequenced (Fiers et al., 1976). By the end of 2003, the complete genome sequences of more than 1,100 viral species were available in public databases. The genomes of bacteria and eukaryotic species are far larger, but in recent years determination of these sequences has also become routine. The Institute for Genomic Research (TIGR), a nonprofit institution in Rockville, Maryland, that has been a major participant in many whole-genome sequencing projects, has built a powerful infrastructure for determining DNA sequences accurately and quickly. In 1995, TIGR scientists published the first complete genome sequence of a free-living organism, the pathogenic bacterium *Haemophilus influenzae,* which contains 1.8 million nucleotides (Fleischmann et al., 1995). By November 2003, complete sequences of 140 bacteria had been deposited in genome databases worldwide, and at least 181 more were being determined. The genomes of dozens of other eukaryotic organisms had also been completed by then, including plants, animals, insects, fungi, and the human.

Genome Data and Analysis

The primary data that DNA sequencing generates consist of a long list of the letters A, T, C, and G in what looks like no order. For whole genomes, the list can be very long. The human genome is more than 3 billion nucleotides long, for example, and the genome of *Yersinia pestis*, the bacterium that causes plague—and that devastated Europe in the Middle Ages—has about 4 million nucleotides.

To keep track of sequence data, the Los Alamos National Laboratory in 1982 opened a data repository called GenBank. The purpose was to create a single repository that would allow easy access to all sequence data as they became available. GenBank moved to the National Center for Biotechnology Information (NCBI) on the National Institutes of Health (NIH) campus in Bethesda, Maryland, in 1988 and has grown to an extraordinary degree in recent years. It now contains more than 30 million gene sequences from more than 130,000 species, comprising more than 36 billion nucleotides (GenBank, http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html).

Since the middle 1980s, GenBank has coordinated its activities, policies, and data with two other large genome-sequence repositories overseas: the European Bioinformatics Institute (EBI) and the DNA Data Bank of Japan (DDBJ). Under the terms of the International Nucleotide Sequence Database Collaboration (INSDC), these three repositories exchange sequence data daily, and thereby each maintain essentially the same set of sequence data as workers around the world submit new data daily.

In addition to those three repositories, however, many other sites provide access to genome data. Some sites are comprehensive, such as those run by TIGR; the Whitehead Institute in Cambridge, Massachusetts; and the Wellcome Trust Sanger Institute in Cambridge, England. Others specialize in particular organisms or topics. WormBase.org, for example, is devoted entirely to the genome of *Caenorhabditis elegans*, a simple nematode often used in basic-science experiments, and the Jackson Laboratory in Bar Harbor, Maine, maintains an extensive on-line library of information and tools relevant to the mouse genome. In addition, numerous private facilities periodically download all new genome data submissions.

Raw gene sequences, however, would be of little use without computers and analytic tools to decipher them. Bioinformatics specialists have been working to create and improve such tools since gene sequences were first obtained. In those early years, sequences consisted of relatively short single genes. Researchers compared the amino acid sequences of proteins from different species by printing them out, cutting them into strips, and examining them by eye to find similarities and differences.

Today, however, entire genomes consisting of thousands of genes have been sequenced, and powerful programs running on large, networked computer systems are needed to analyze, compare, and interpret the data and store the results in an accessible form. The first level in genome analysis is called annotation. After assembling the entire sequence, computers scan the data for landmarks, such as start and stop signals for genes that encode proteins. The protein sequences of putative structural genes are predicted, and, if possible, a potential physiologic or regulatory function of each encoded protein is assigned on the basis of similarity to known proteins. Three-dimensional structural models of proteins encoded by genes can also be constructed. And the full-length genome is analyzed for the presence of entire biochemical and regulatory pathways.

Just as computer-based translations of human language can yield peculiar results, the results of computerized genome annotation are often flawed. NCBI, EBI, TIGR, and other organizations employ teams of experts to constantly check and edit the results of computer-generated annotation. Borrowing a word from museum-exhibit management, genome scientists refer to this editing of computer-generated genome sequence annotations as curation. The resulting curated annotations are stored in standard

formats in databases for easy access. The entire set of annotations for one genome can then be compared with those derived from other whole genomes. As new analytic tools become available, existing sequence data are constantly reanalyzed by both machines and people to keep them as up to date and accurate as possible.

Although annotation and curation are the first steps in making sequence data comprehensible, scientists need sophisticated tools to make efficient use of them. A remarkable array of such tools has been placed in the public domain, and new ones are under development constantly as genomics expands. NCBI, EBI, DDBJ, TIGR, the Sanger Institute, and other institutions provide many Internet-accessible analytic tools that scientists can use to query genome and other databases to solve questions relevant to their work (see, for example, the list of tools available at NCBI at http://www.ncbi.nlm.nih.gov/About/tools/index.html). The tools can also be downloaded free with annotated data for use on users' personal computers and shared over local computer systems. Thus, the vast majority of known genome sequences and the tools to analyze them are freely available to anyone in the world who has a computer and Internet access.

In addition to gene sequences, biologists have constructed databases that contain many other kinds of data, including protein amino acid sequences, three-dimensional structures, protein functions, organism taxonomy, and protein-protein interactions. The scientific literature is indexed, and abstracts made available, through freely accessible databases maintained by the National Library of Medicine. Most newly published scientific articles are available in electronic form, although many journals limit full access to paid subscribers. In addition, databases have recently been established to catalog and make available experimental data on changing gene-expression patterns obtained in microarray experiments. Those experiments often generate far more data than can be easily interpreted by a single laboratory; the databases are intended to let other scientists access and interpret the data for their own work. EBI, for example, has a Web site called ArrayExpress for this kind of information (http://www.ebi.ac.uk/arrayexpress/), and the Microarray Gene Expression Data Society was founded in 1999 to facilitate the exchange of such data (http://www.mged.org/index.html).

Database-Access Policy

Science thrives on free and open exchange of ideas, results, data, and materials. It is a long-standing principle of scientific practice that all experimental protocols and data be completely described at the time of publication of a scientific finding. That allows others to evaluate thoroughly whether the analysis was done correctly, to repeat the experiment

if they so desire, and to use the published work to further their own research.

GenBank and other international gene sequence databases were set up to allow the free and open exchange of genome data. Consequently, the policy of the INSDC has always been to offer worldwide users open, unfettered access to all data, including genome annotations (http://www.ebi.ac.uk/embl/Documentation/INSD_policies.html). Moreover, privacy policies for all three of the member institutions state that no personally identifiable information is collected about what data a user might access or how these data might be analyzed. The largest nonprofit organizations that provide genome information and analytic tools, such as the Sanger Institute and TIGR, have similar policies mandating free, unfettered, and anonymous access.

The open-access policies are guided by broader U.S. government policy statements concerning the release of results of scientific research funded by the federal government. National Security Decision Directive 189 (NSDD-189), promulgated by the Reagan administration in 1985, states that access to fundamental research results should be unrestricted to the greatest possible extent. If such access is deemed a threat to national security, the research results should be formally classified as secret. Classified documents are available only to people who have undergone an approval process controlled by the government. Individual classified documents are then made available on a need-to-know basis and are subject to regulations on the locations and situations in which they may be viewed and stored. Recently, however, there has been considerable discussion of "sensitive but unclassified" information. How a sensitive-but-unclassified label might be used to categorize the products of life-science research remains to be seen. The issue is discussed further in the recent National Research Council report *Biotechnology Research in an Age of Terrorism* (NRC, 2003a). For the present, policies at the major U.S. funding agencies for life-science research still adhere to the principles set forth in NSDD-189 and long-standing scientific practice, and they require that grant recipients make research results and data publicly available. The NIH Grants Policy Statement, for example, says that "it is NIH policy to make available to the public the results and accomplishments of the activities that it funds" (NIH grants Web site, http://grants1.nih.gov/grants/policy/nihgps_2001/part_iia_6.htm).

Scientific journals also strongly support openness as a scientific norm and require the deposition of primary gene sequence data into a free and open database as a condition of publication of research results based on them. *Science,* for example, states that authors must "agree to honor any reasonable request for materials and methods necessary to verify the conclusions of experiments reported, and must also agree to make the data

upon which the study rests available to the scientific community. For large data sets such as DNA sequences, *Science* advises authors that this means deposition in GenBank or some other open database prior to publication of the paper . . ." (*Science* web site, http://www.sciencemag.org/feature/contribinfo/home.shtml). Most life-science journals have similar policies.

USE OF GENOMICS IN MODERN LIFE-SCIENCE RESEARCH

Genome data have become indispensable to the conduct of much life-science research—to the point where not many life scientists would consider starting a project without thinking about how existing genome data could be used in their experimental design. The growing importance of genomics cuts across all divisions of the life sciences to include biomedical, agricultural and environmental-biology topics; basic and applied research; and science in academic, government and industrial laboratories. Just as no entrepreneur would start a business without thinking about how to use computer technology, most biological scientists today do not go into a laboratory without incorporating available genome data into their plans.

Although access to whole-organism genome sequences has become vital to life-science research, the data do not immediately provide understanding of any organism's natural properties, nor do they furnish a road map for manipulating the organism to give it new properties. A credible attempt to do that requires substantial experience, knowledge, training, and a great deal of patient thoughtful experimentation. And because biological systems are so intricate and finely tuned, attempts to manipulate genome structure rarely work out as the experimenter expects. Nonetheless, the growing library of genome data is an extraordinarily potent research tool. Malefactors might make use of genome data on organisms engineered by others (and selected for particular traits of interest); this would almost certainly be easier than trying to create new phenotypes themselves, but the full implications of the modification may not be clear from the publicly available data.

Genome data allow investigators to use and develop experimental tools that are far more potent than those available just a few years ago. One of the most powerful is the whole-genome microarray (Duggan et al., 1999). DNA consists of two strands that bind to one another when their sequences are complementary (see Figure 2). In the 1970s, molecular biologists developed techniques that allowed them to isolate a specific DNA sequence, make multiple copies of it, and add it to an experimental solution to "probe" for the presence of the complementary sequence. If present, the complement would bind to the probe, and the resulting double-stranded molecule could be readily detected. Finding the complementary sequence might, for example, indicate that the gene associated

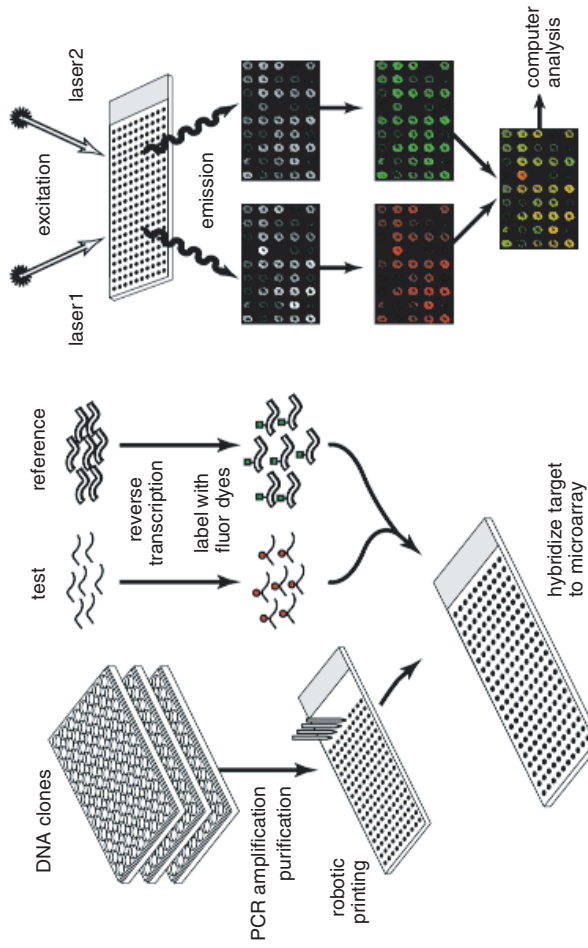

FIGURE 2 cDNA microarray schema. SOURCE: Reprinted, with permission, from Duggan et al., 1999.

with it was being transcribed and the protein it encoded was being synthesized.

Microarray technology developed in the 1990s allows an investigator to carry out thousands of DNA probe experiments simultaneously. A microarray is generally a set of single-stranded DNA sequences, each representing a single genetic feature bound at known locations onto a suitable surface, often an ordinary microscope slide. Each bound sequence acts as a specific target for the presence of its complementary sequence found in an experimental solution, for example, material extracted from cells or from animal tissue. The presence of combined target and complement is commonly detected by using sensitive fluorescent markers (to mark the DNA in the experimental solution) and a laser scanning device. Microarrays can be constructed by using the same principle as an inkjet printer; an area scarcely the size of a coin can contain individual samples of DNA sequences corresponding to every one of the thousands of genes in an organism's genome. Scientists now routinely use microarrays to determine how gene-expression patterns change in cells in response to experimental variables or to study the DNA sequence of a microbial agent. Studies in differential gene expression are performed in organisms as diverse as bacteria and human cells and, indeed, can be used to monitor the changes that occur when cells are infected by parasites, become cancerous, or are just growing normally. They have made it possible to detect unique molecular "signatures" of biological events. This powerful technology depends entirely on the availability of genome data, and today a single scientist performing a single experiment can obtain data that would have taken years to obtain with older techniques—if they would have been collectible at all.

The impact of genomics goes well beyond microarrays and other technical advances that improve the efficiency of data collection. As the entire genomes of many organisms have been sequenced and made widely available, research scientists have begun to analyze genomes as an individual complex biological system and compare them with each other. This kind of comparative analysis, called systems biology, will greatly accelerate understanding of how entire organisms work and how organisms interact with one another. In particular, it will initially facilitate understanding of bacterial genomes (Rappuoli and Covacci, 2003). As more genomes are sequenced, the power of comparative genomics will increase our ability to understand larger biological systems (Kanehisa and Bork, 2003). Two examples—development of meningococcus B vaccine candidates and the recent experience with the SARS coronavirus—illustrate how genome data and contemporary experimental techniques have permitted the rapid development of new products and tools to fight infectious diseases. These examples focus on human health but the techniques used are equally

applicable to the development of countermeasures to plant and animal pathogens.

Meningococcus B Vaccine

Neisseria meningitidis (the meningococcus) is a bacterial species that can cause meningitis (an infectious disease of the fluid and membranes surrounding the brain and spinal cord) and septicemia (infection of the blood). Those infections are fatal if untreated. The bacterium is spread through intimate contact and airborne droplets.

Several distinct strains of the meningococcus have been identified, and vaccines for many of them have been developed with standard methods (Moxon and Rappuoli, 2002). All vaccines rely on the principle that after the immune system has been primed to mount a protective response to an agent, it can mobilize quickly to defeat infection when the immunized person is exposed to the agent again (Grifantini et al., 2002). Classical vaccines are based on the administration of killed or attenuated versions of pathogenic agents and increasingly on the use of purified molecules from cultured bacteria that can elicit a protective immune response when injected into a susceptible person.

In the case of *N. meningitidis,* it was known that the injection of a polysaccharide capsule from the meningococcal cell wall worked well as a vaccine for preventing meningococcal disease caused by different bacterial types—A, C, Y, and W135 (Pizza et al., 2000; Adu-Bobie et al., 2003). However, the corresponding capsular polysaccharide from type B meningococci was ineffective for protection because it is very similar to a molecule that humans also produce; the immune system usually fails to produce antibodies against such "self" antigens and thus avoids harming the host. Despite years of effort, biomedical scientists failed to find a protective molecule that would induce immunity to type B meningococcal disease (Moxon and Rappuoli, 2002).

The complete genome of *N. meningitidis* type B was sequenced in 2000 (Tettelin et al., 2000). Scientists at Chiron Corporation then used the sequence in "reverse vaccinology." That is, they worked from gene to protein to vaccine candidate rather than purifying various bacterial constituents for testing as protective antigens. Reverse vaccinology uses the analysis of every gene in the type B meningococcal genome, looking for gene products that might encode proteins likely to be "seen" by the immune system during infection; these are molecules predicted to be either on the bacterial surface or secreted by the microorganism into its surroundings. Of the 600 such type B genes identified in their computer analysis, they inserted 350 into *E. coli* that then manufactured the corresponding encoded type B proteins. The recombinant proteins were puri-

fied and injected into mice. Blood serum from the immunized mice was then analyzed to see which might contain antibodies that would bind to the *N. meningitides* type B cell surface and kill the bacteria. They discovered 28 candidate antigens that could induce this killing (bactericidal) activity. Five of them are now in initial clinical testing—less than 3 years after the genome data first became available and after more than 2 decades of failure with standard pregenomic vaccine-development methods. It is possible that products could enter the marketplace within 5 years.

That work is the first example of how a genomic approach can lead to novel vaccine candidates. It will certainly not be the last. Similar reverse-vaccinology efforts are currently under way to apply the strategy to several other microbial pathogens, including those of malaria, plague, and anthrax (Mora et al., 2003).

Those efforts and a second generation of vaccine-discovery strategies aimed at pathogenic microorganisms make use of proteomics and whole-genome microarray analysis of gene transcription to find potential protective antigens. Proteomics refers to the rapidly advancing ability to isolate and analyze large numbers of proteins from a cell efficiently. It provides additional structure and function data about various candidate protein antigens that will help to identify the most promising among them. In addition, microarray analysis of bacterial gene activation when the bacteria first encounter host cells can complement the kind of genome analysis described above. Specifically, these experimental approaches allow identification of genes that are not active in cell culture but produce their encoded protein *only* when the bacteria are actively interacting with an infected host. For example, when this type of analysis was applied to *N. meningitides* type B, several additional protective antigens were discovered that are made only in the presence of human cells. Those antigens could not have been characterized with genome analysis alone or by simply isolating various proteins from cultured *N. meningitidis* cells grown on ordinary laboratory growth media (Grifantini et al., 2002). It requires the use of genome information and carefully executed laboratory experiments.

SARS Coronavirus

The disease that became known as severe acute respiratory syndrome (SARS) first came to world attention on March 12, 2003, when the World Health Organization (WHO) issued a global health alert about an atypical pneumonia in Viet Nam, Hong Kong, and Guangdong Province, China. The global research community responded vigorously with all the tools of modern science. Within 6 weeks, the virus that causes the disease, dubbed SARS coronavirus (SARS-CoV), had been isolated and cultured

and its 29,727-nucleotide genome completely sequenced and posted on the Internet. In the months that followed, dozens more SARS-CoV isolates were sequenced and published.

The availability of the sequence data quickly put to rest fears that SARS was the result of a laboratory-fabricated agent. The sequence data also allowed research scientists throughout the world to begin immediately to analyze viral structure, function, and the molecular basis of how it might cause illness. The sequence quickly revealed that the new virus was related to other coronaviruses and provided key insights into its potential pathogenic mechanisms. The sequence data were also crucial to global efforts to develop candidate vaccines, antiviral drugs, and especially accurate, sensitive diagnostic tests.

Vaccines. Within 3 months of the initial WHO alert, workers in academic, government, and industrial laboratories had created several SARS-vaccine candidates and were moving to test them in animal models. Many more vaccine candidates have since been created. In most cases, vaccine development relied entirely on knowledge of the viral sequence. Anti-SARS DNA vaccines, for example, are based on DNA sequences that encode portions of a viral protein. Those DNA sequences are injected directly into a vaccine recipient, whose cells take up the injected DNA and express the viral protein in a way that stimulates the immune system. Scientists have also created several live attenuated vaccine candidates by inactivating specific genes in the viral genome. Without sequence information, those vaccine strategies could not have been pursued so quickly. Several of the candidate vaccines have shown initial promise, and some are in preclinical testing in nonhuman primates.

Drug Therapies. In a search for antiviral drugs, the goal is to find a compound that can disrupt the viral life cycle without harming the infected host. Screening efforts were begun immediately after SARS emerged, looking for compounds that could prevent viral replication in cell culture. Many compounds screened were obvious choices; they included every known drug that might have antiviral activity. The choice of other potential antiviral-drug candidates, however, relied on insights provided by the SARS sequence. For example, analysis of the sequence made it clear that SARS-CoV enters cells by fusing with their outer membrane. That immediately suggested that drugs that inhibit membrane fusion might be active against SARS. Furthermore, knowledge of the amino acid sequence and the three-dimensional structures of the SARS proteins involved in fusion provided clues for targeted development of more-efficient fusion inhibitors.

The sequence data also provided clues about how to design drugs that could interfere with other viral proteins. For example, the SARS virus

contains a protease enzyme that cleaves and thereby activates many other viral proteins. Possession of the sequence of the protease allowed protein chemists to quickly construct three-dimensional molecular models, which revealed it to be structurally similar to a protease from rhinoviruses, a separate family of viruses that cause the common cold. Research workers at the pharmaceutical company Pfizer took particular note of that similarity because they had recently designed and synthesized a series of peptides intended to inhibit the rhinovirus protease. Some of those compounds were found to partially block SARS-CoV replication in cell culture. Pfizer scientists then began to refine the SARS protease structural model and to design new drug candidates that might bind the SARS protease more tightly and therefore more effectively inhibit its activity and serve as a potent anti-SARS therapeutic agent.

Diagnostics. The initial symptoms of SARS, like those of many other viral infections, are fever, malaise, and other nonspecific "flu-like" symptoms. Thus, fast and accurate diagnostic tests are needed to separate SARS patients from those with less serious illnesses. The availability of the SARS sequence has greatly accelerated diagnostic development. One standard method for detecting virus in a clinical sample involves the use of antibodies that can bind to viral proteins. Creating SARS-specific antibodies requires purification of viral proteins from cultured virus, which are injected into animals to produce antibodies. The availability of the SARS sequence allows the use of more-efficient genetic-engineering techniques to make recombinant versions of the SARS proteins in bacteria to produce antibodies that can be used in diagnostic tests.

Two of the most promising new diagnostic strategies, polymerase chain reaction (PCR) and diagnostic DNA microarrays, rely entirely on sequence data. In PCR, DNA sequences that are complementary to specific viral sequences are synthesized in the test tube and added to a clinical sample with appropriate transcription enzymes. If SARS viral sequences are present in clinical material, such as sputum, they are easily identified with specific enzyme-detection methods. Thanks to the availability of the SARS sequence, many PCR kits and procedures for SARS detection have been developed around the world.

GENOMICS AND BIOTERRORISM

Modern Technology

As noted above, access to sequence data and the associated tools needed to analyze them are indispensable tools for life-science research. But genome databases are also of interest to anyone who might want to

enhance pathogens for destructive purposes. Just as any scientist planning research aimed at finding new cures for infectious diseases would tap the power of genomics, so would any malefactor setting out to create engineered pathogens for use as biological weapons.

In the future, newly engineered agents will be a growing concern, but even natural pathogens can be used to carry out devastating attacks. The naturally occurring forms of the "category A" infectious agents that the Centers for Disease Control and Prevention considers to be the worst potential bioterror threats—those of anthrax, smallpox, botulism, plague, tularemia, and viral hemorrhagic fevers—already have inherent properties that give them terrible destructive potential. The most difficult step in carrying out an attack with such an infectious agent is neither obtaining a starter culture of the organism nor expanding it to produce a quantity needed to conduct an effective bioterror attack. Instead, the highest hurdle is preparing and disseminating the agent so that it can be delivered effectively to a dispersed target population; for contagious agents, this problem is less difficult because infection of a relatively small number of people can spread widely. (That is not the case for smallpox, however, which is known to exist only in two high-security laboratories and would therefore be very difficult for a terrorist to obtain.) For an attack intended to affect many people simultaneously, the delivery vehicle of choice would probably be an aerosol or the food chain. If effectively executed under even less than optimal conditions, such an attack could be catastrophic. Even a small-scale attack can have serious consequences, as shown by the 1984 *Salmonella typhimurium* salad bar attacks in The Dalles, Oregon, and the better known 2001 anthrax attacks (Torok et al., 1997). Those consequences can include illness, death, and social and economic disruption. Although the dissemination of an infectious agent for either a large-scale or a small-scale bioterrorism attack may be difficult, it is important to acknowledge that relatively unsophisticated dissemination methods are effective.

Regulations on access to genome data would not affect the ability of a terrorist to carry out an attack with naturally occurring pathogens. The techniques used to prepare agents for this kind of attack were mastered by workers in biological-warfare programs in the 1950s, and a large, unclassified technical literature relevant to the methods required to aerosolize pathogens already exists. Although the technical hurdles that would confront a bioterrorist intending to deploy a naturally occurring agent to cause large numbers of casualties are substantial, they are much lower than those associated with enhancing the virulence of a known pathogenic species with genetic manipulation. Thus, an attack with a natural pathogen is still the most likely; however, given the developments in biotechnology described in this report, a more sophisticated attack with an engineered pathogen is a serious concern.

Potential Malefactors

Those who might attempt to make an engineered pathogen that has enhanced properties are in several categories. Nations conducting dedicated biological-warfare research and development programs are likely to have access to substantial funding and the requisite expertise. However, attempting to create an enhanced pathogen would not be beyond the capabilities of a lone person with the appropriate background, a relatively modest budget, and a destructive agenda (Carlson, 2003). A nation-state or a group might be able to recruit a scientist who already has access to facilities and could use them for biological-weapons research. Subnational terrorist groups, such as Al Qaeda, or apocalyptic religious groups, such as Aum Shinrikyo, which released the nerve gas Sarin in the Tokyo subway in 1995 and experimented with biological weapons, might also attempt such a project (Lifton, 1999). Press reports assert that Al Qaeda is attempting to achieve biological-weapons capability, although of a conventional variety involving naturally occurring pathogens and toxins (Petro and Relman, 2003). It is possible that fanatical religious groups with apocalyptic fantasies, such as Aum Shinrikyo, would be interested in developing enhanced pathogens intended to cause an indiscriminate global catastrophe (Kaplan, 2000).

As the technology for manipulating DNA becomes more widespread and easier to use, some observers have suggested that large numbers of amateur experimenters might begin to dabble in the molecular engineering of organisms (Carlson, 2003). One technology columnist, for example, went so far as to suggest that perhaps "bathtub biotech" will do for biology what garage hackers did for information technology (Schrage, 2003).

> Maybe bioinformatics and the diffusion of genetic engineering technologies and techniques will inspire a new generation of bio-hackers. Certainly the technologies are there for those inclined to genetically edit their plants or pets. Maybe a mouse or *E. coli* genome becomes the next operating system for hobbyists to profitably twiddle.

Whether many "bio-hackers" will actually emerge remains to be seen. Most amateurs today would be unlikely to achieve much through manipulation of microbial genomes, which is far more difficult than many people outside the scientific community recognize. In addition to relevant biological training, a potential terrorist interested in engineering new pathogens would need access to appropriately equipped experimental facilities. Unexpected difficulties often arise in this type of work. For some organisms, genetic systems are not well understood, so many details must be determined from scratch. And random and targeted gene insertions or deletions can have unintended consequences that change the phenotype of an organism in ways not anticipated.

Apart from ability, the proportion of amateurs who would be interested in deliberate manipulation of pathogens is not known. Given the record of destructive computer viruses created by computer hackers, however, the possibility that the tools needed to carry out 21st-century genetic engineering will be available to hobbyists the world over is unsettling and underscores the fact that individuals or groups that might want to enhance an organism's potential to cause harm theoretically have powerful tools within their reach. Large-scale bioterrorism is unlikely, but the possibility of such a rare devastating event dictates that we not dismiss it and that we be vigilant.

How Genome Data Might Be Misused

One way that governments, groups, or individuals might misuse genome data would be to conduct primary research on pathogen enhancement, starting with hypotheses that they generate themselves from genome analysis and pursuing them experimentally in the laboratory. Creating a genetic construct that an experimenter believes might be more virulent than its naturally occurring form is not very demanding technically. However, experimental evidence has shown that enhancing pathogenicity is quite difficult. Manipulations of biological systems rarely turn out as planned; attempts to change one property, even if successful, usually have consequences that the experimenter does not want. Pathogens have been highly refined by nature over many millennia, and even minor manipulation can lead to unexpected consequences. The scientific community does not understand virulence and pathogenesis well enough to predict the results of genetic engineering reliably. If the genome is changed to overproduce a known virulence factor, for example, the organism could be unchanged phenotypically or not be able to infect its host efficiently in a real-world setting. For example, Isberg and Falkow in the 1980s showed that the gene for a protein they called invasin, when transferred from *Yersinia pestis* into *E. coli*, permitted entry of *E. coli* into cultured mammalian cells (Isberg and Falkow, 1985); this was discussed in the book *Germs: Biological Weapons and America's Secret War* (Miller et al., 2002). However, not everyone who read the Isberg and Falkow paper understood that the inheritance of invasin did not turn *E. coli* into a pathogen. When investigators put the invasin gene into *Salmonella* and *Shigella* to determine its effects on virulence, it had *no* effect (Voorhis et al., 1991). In addition, although invasin is active in the closest known relative of *Yersinia pestis*, it is inoperative in *Yersinia pestis* itself. (Rosqvist et al., 1988)

Instead of pursuing enhancement strategies of their own, however, malefactors might limit themselves to replicating or adapting published results that have revealed how pathogens can be enhanced. As the

example above illustrates, replicating the same type of genetic modification in a new organism may or may not have a similar result. Such experiments generally are in the category of "functional genomics" because they go beyond obtaining gene-sequence data to tie specific genome information to the implied specific functions, capabilities, or vulnerabilities of an organism. Successful manipulation of microorganisms to create more-efficient biological weapons probably requires methodical investigation of the types of changes that individual kinds of microorganisms can handle, a long time spent on trial-and-error experiments, or much good luck. The cases presented here are only examples. They illustrate that it is no easy feat, but certainly is possible, to use genome data to design an enhanced biological weapon.

As stated above, results that have immediate implications for pathogen enhancement or weapons development have been called contentious research (Epstein, 2001) or are said to be in a gray zone where the benefits of publication might not outweigh the dangers. At the workshop one speaker suggested that the *Nature Biotechnology* paper "Engineering hypervirulence in a mycoherbicidal fungus for efficient weed control" (Amsellem, 2002) might fall into this zone because it uncovered unanticipated lethality for tomato and tobacco plants. This presenter also discussed his own views that the gray zone has an evolving nature, as well as differing shades of gray, and called for continuing review and discussion of the gray zone and of whether any monitoring or control efforts would be beneficial. He called for absolutely transparent discussions that involve multiple communities (scientific, intelligence, public, and policy).

Another example of such gray zone research was publicized in late October 2003 when scientists at St. Louis University extended previously published work in which the mousepox virus was made hypervirulent and capable of overcoming an effective vaccine (Washington Post, 10/31/03, pg A1). In 2001, an Australian research group had shown that insertion of a gene for the mouse version of an immune regulator called interleukin-4 (IL-4) into the mousepox virus increases the virus's virulence; cells infected with the modified virus produce excess IL-4, which "jams" the IL-4 signal and thereby disrupts the normal immune response to infection. As part of a broader effort to explore possible countermeasures to address engineered pox viruses, the St. Louis scientists extended the Australians' work, inserting the IL-4 gene into a different part of the mousepox genome so that it came under the control of different regulatory sequences that increased the amount of IL-4 generated in infected cells. The result was an extraordinarily potent virus that killed every one of the mice it infected, including those previously vaccinated. It is interesting to note, however, that, unlike the wild-type virus, the altered virus reportedly was not transmitted from animal to animal.

Those results are significant in two ways. First, they provide an obvious starting point for anyone who might want to create and release an enhanced pox virus. Mousepox is closely related to several viruses that can cause disease in humans, including viruses that cause smallpox, cowpox, and monkeypox. It would not be difficult for a skilled scientist or technician to use the published results to carry out an analogous genetic manipulation of one of those viruses; the effect of the manipulation of the other pox viruses on virulence and on the ability to overcome vaccine in these is not known. Second, they clearly demonstrate the important principle that gene sequences in a host can be as important to people who intend to enhance a pathogen as gene sequences in the pathogen itself. In this case, replicating the work requires the *human* IL-4 gene sequence. It should be noted that the effect of the modification in the mousepox virus was to make the host more susceptible. That requires an understanding of the host in addition to an understanding of the virus. A number of microbial pathogens cause disease by manipulating host immune function. One important implication is that a potential bioterrorist could use human, animal, or plant genome sequences to create a more dangerous pathogen.

3

Issues in the Control of Genome Information: From Discussions at the Committee's Workshop

The committee held a workshop on October 1, 2003, to gather input from a diverse group of people concerned with the control of genome information, science, and security. A list of the participants and the agenda for the workshop are appendixes to this report. At the workshop, presentations described existing databases and how they are used to advance research, the international issues that arise when one country discusses controlling data, and potential ways to classify genome data with respect to possible threats. Discussions were held on the pros and cons of unlimited or restricted access to data, and breakout sessions addressed the security effects of free release of data, the scientific effects of restricting release of data, and potential mechanisms for controlling release.

Two distinct concerns were apparent throughout the workshop discussions. On one hand, given the enormous potential for human benefit from the accelerating progress of the life sciences and the extent to which data from one field of research might shed light on others, workshop participants were deeply concerned that any policy to withhold genome data would slow the advance of science and would thus impair scientists' ability to improve understanding of pathogenesis and to develop countermeasures to future biological threats, whether natural or human-made. Therefore, any policies that had the effect of constraining science would have to be justified by identifiable security benefits. On the other hand, however, participants clearly understood that the power of the growing human understanding of the life sciences is such that individuals, groups, or nations could someday use the information to cause terrible harm.

The first section of this chapter summarizes the different points of view on the issue of control by grouping stakeholders into broad categories. The ideas come from statements that workshop participants made about themselves and about their communities. The section includes the international interconnections within the life-science community and implications for the control of genome information; this discussion is based on the presentations of Lord May of the Royal Society, Rino Rappuoli of Chiron-Italy, and Michael Morgan of the Wellcome Trust and on input from various other participants during the workshop. The second section of the chapter discusses ways that genome data could be categorized and whether any individual category of data might present an enhanced threat; this was the subject of discussion for much of the afternoon portion of the workshop. The third section identifies potential mechanisms for controlling data, a topic that came up repeatedly during the workshop. The fourth and final section of the chapter summarizes the arguments made for and against instituting restrictions on data; it draws on discussions throughout the workshop, especially the two breakout sessions on the security and scientific effects of releasing and restricting data. Two major foci were the feasibility and desirability of instituting registration requirements for access to genome databases.

STAKEHOLDERS IN THE DEBATE OVER RELEASE OF GENOME DATA

The crux of the dual-use dilemma in the life sciences is this: It is difficult or impossible to limit the application of ideas and data generated through research to beneficial purposes. At the broadest level, all humanity has a stake in how scientists and policy-makers confront the dual-use nature of modern life-science research. The problems posed by naturally occurring emerging and re-emerging infectious diseases—such as HIV/AIDS, influenza, multiple-drug resistant tuberculosis, foot and mouth disease, and SARS—present difficult challenges to global health and security and to the global economy. Scientific research has the potential to deliver powerful new tools to meet the challenges that infectious diseases present. The consequences of retarding scientific progress must be considered in any decision to restrict access.

At the same time, the growing power of the life sciences permits humanity to manipulate nature in new ways, including, in theory, the creation of pathogens with destructive properties that would be unlikely to emerge naturally. For example, the Australian scientists who published the 2001 finding that interleukin-4 (IL-4) increases mousepox virulence made that finding as part of a project to engineer a mousepox virus variant that would induce the mouse immune system to attack proteins displayed

on the surfaces of fertilized mouse oocytes and thereby render infected mice infertile. It is possible that the results of such a project, if successful, could be adapted and expanded to create a contagious virus that could make humans infertile. But the growing power of biological research and technology could also work to counter any kind of human-generated threat, just as it would in response to natural infectious-disease threats.

Domestic Interest Groups and Perspectives

The main question before this committee concerns the degree to which access to genome sequences and related information should be restricted or left open to all. Different groups and communities have different perspectives on this aspect of the dual-use problem. At the risk of creating imprecise caricatures of various approaches to this question, the most important of these groups in the United States can be said to be the scientific community, the security community, and the general public.

- ***The Scientific Community.*** This group includes practicing scientists and administrators in government, academic institutions, and the private sector who are involved in basic scientific research. Generally speaking, members of this group view their work as part of a much wider effort to improve health and welfare. Basic scientists are intimately aware of how important open communication is for rapid scientific progress, and many members of this group favor the maintenance of free and open sharing of data, materials, and ideas among scientists everywhere (Salyers, 2002; Check, 2002). Exceptions to complete openness, of course, exist in routine scientific practice. For example, results are often not shared in the open literature until those who obtained them have had the chance to exploit them fully in their own laboratories or to patent them.
- ***The Security Community.*** This group includes people in the military, intelligence and other federal agencies, law enforcement, and industry whose main concerns are the protection and maintenance of national security. Members of this diverse group have much more experience in handling classified information than do most life scientists. They are therefore not only used to situations in which disclosure of information can seriously undermine security but also well acquainted with the costs of compartmentalization of information, such as the difficulty of getting important information to the people who can use it best. Given the nature of their work, it is difficult to make broad generalizations about what members of the security community think about open information exchange within the life sciences as they typically operate under some constraints. Although some members of the security community might look askance at the current high degree of openness in the life sciences and suggest that

greater restrictions on the flow of research information might reduce the risk of harmful misuse of new results, this view is not widely held (Vastag, 2003; Franz, personal communication). Many in the security community favor retaining the current openness of biological research, arguing that openness and free exchange of information enhance security by strengthening biodefense response capabilities. Some favor achieving openness and transparency by fostering international collaboration in research; others favor the creation of formal international agreements and regulatory regimes to achieve the same end (Epstein, 2001).

- ***The General Public.*** Acting through their elected representatives, Americans have provided strong support for life-science research, especially biomedical research. The National Institutes of Health budget, for example, rose from $13.6 billion in 1998 to $27.2 billion in 2003 (AAAS, 2003; http://www.aaas.org/spp/rd/nih04p.pdf). Congress, the executive branch, and the public seem to have reached a consensus that investment in such research will provide a good return in the form of better health and longer life. It is safe to say, however, that most people do not have a thorough understanding of how fast the life sciences are advancing, nor are they fully aware of how open exchange of data accelerates scientific progress. But some members of the public are clearly troubled by the possibility that biological research could be used for destructive purposes. For example, one scientist received hate mail after announcements in the media that he had created a genetically engineered mousepox virus (Weiss, 2003).

The interplay among the various stakeholders will be complex as the debate on what to do about dual-use biological research moves forward (Kwik et al., 2003; NRC, 2003). The scientific and security communities, it is often said, do not understand one another and in fact seem to represent "two cultures" (Kennedy, 2003). Scientists tend to oppose calls for restrictions on data accessibility or results that might inhibit their work, and others object to what they see as an irresponsible aversion among scientists to facing the growing threat emanating from the life sciences. Elected representatives (like the people they represent) and other policy-makers are found on both sides of the divide and can be expected to look for ways to preserve the advantages that flow from life-science research while limiting the danger that the research may present (Atlas, 2002). Support for biological research will probably remain strong. Fear of bioterrorism, however, like the fear of naturally acquired infection, is not without foundation, and in the aftermath of an accidental or deliberate release of an enhanced pathogen public opinion could swing in favor of limiting the exchange of life-science results.

International Issues

Modern life-science research is an international enterprise. Many nations are vigorously pursuing all aspects of biological research. In addition to the United States, the countries of the European Union, and Japan, nations making large investments in the life sciences include Israel, China, Singapore, Russia, South Korea, India, Brazil, and Cuba. International collaborations between laboratories that might have been unusual a decade ago have become routine. Similarly, results, data, personnel, and experimental materials in the life sciences regularly move across borders. (For further discussion, see NRC, 2003.) In this context, the International Nucleotide Sequence Database Collaboration (INSDC), in which policies are coordinated and data are routinely shared among the world's three largest genome sequence repositories, is a natural consequence of the international and cooperative traditions of life-science research.

Because biological research is a global activity, any actions taken in the United States to restrict access to genome databases would inevitably have international ramifications. Any restrictions placed on access to data generated in the United States or put into databases under U.S. jurisdiction would affect the operation of databases in other countries, including the INSDC partners in Europe and Japan. Such policies would therefore have to be coordinated with those partners or the collaboration terminated. Any restrictions on access to U.S. genome databases would not keep such data out of the hands of potential malefactors unless all other genome databases formulated similar policies. Those databases are available to anyone with Internet access, so any restrictions on U.S. sites could easily be circumvented simply by navigating to another site. There is no international consensus that restricting data access is warranted; indeed, some workshop participants expressed the belief that sentiment abroad was firmly in favor of maintaining free and open access to genome data. Any restrictions that would limit access by countries with small scientific resources would be controversial and might be seen as an attempt by wealthy nations to prevent developing nations from using biological advances effectively. The Biological and Toxins Weapons Convention of 1972 (http://www.opbw.org/) addresses some aspects of international scientific research, and restrictions on sharing genome data might be seen as counter to the spirit of Article X of the convention, which enjoins parties to the treaty to cooperate on scientific discoveries.

The workshop participants discussed a current situation in which the U.S. government must decide whether information will be made public. The decision will have international ramifications. Federal agencies have recently obtained DNA sequences for about 20 smallpox-virus samples held by the Centers for Disease Control and Prevention. The samples were

sequenced so that scientists could look for correlations between sequence, virulence, and the clinical presentation of the disease. That may aid in the development of new anti-infective countermeasures, which might be needed if smallpox is used as a weapon. In making a decision on release of the sequences, the federal government must consider the possibility that a decision to withhold information will convey the misleading impression abroad that the United State is engaged in research connected with the hostile use of biological agents.

CATEGORIES OF GENOME DATA

The workshop participants considered what categories of genome data present the greatest concern. This was the major topic of discussion of a breakout session, and it was addressed by the full group at multiple points throughout the workshop. Moreover, various ways of categorizing genome information were often implicit in the discussions.

However, the study of microbial pathogenic mechanisms, like other fields of biological research, lacks neat compartments into which data can be categorized. The committee did not see evidence that identifying data as belonging to any category would necessarily make them a greater threat. It is important to remember that the focus here is on access to *data* pertaining to organisms, not on access to the organisms themselves; for example, U.S. government regulations on select agents apply to the possession of the *organisms* and not to their genome sequences.

There are many reasons why it is difficult to categorize genome data by risk. First, the study of nonpathogenic microorganisms is often closely related to the study of pathogenic species. The ubiquitous soil bacterium *Bacillus cereus*, for example, is closely related to *Bacillus anthracis*, the bacterium that causes anthrax; insights gained from the genome of one have been directly applicable to the other (Parkhill and Berry, 2003). Second, biological-weapons developers and those studying ways to counter biological weapons both use model strains to simulate real agents so that they can do development work and trials more safely. One classical model of anthrax is the insect pathogen *Bacillus thuringiensis*, which is widely used as a microbial pesticide. It could be argued that knowledge of its genome would be beneficial to a malefactor hoping to genetically enhance *B. anthracis*. Third, data derived from a single microbial species are not the only data relevant to understanding it. Instead, the ability to compare genes, genetic control mechanisms, and protein function among the entire growing and diverse catalog of completely sequenced microbial genomes is what drives many current research efforts (Frazer et al., 2003; Kanehisa and Bork, 2003). Such comparisons among species have already proved to be a productive approach to deciphering how pathogenic and non-

pathogenic species function as complex biological systems. Fourth, genome data that help scientists to clarify how pathogenic microorganisms cause disease are by no means limited to microorganisms. Human gene sequences and sequences from other "host" species are crucial data for those seeking to understand the intricacies of the interactions between the immune system and microbial pathogens, including specific immune mechanisms and vulnerabilities. The gene sequences of humans and other host species and the insights derived from them therefore would be crucial "enabling data" both for those who would work to find new ways to defeat pathogens and for those who might hope to modify pathogens to exploit immune vulnerabilities and create pathogens with unusual or particularly destructive properties.

Categories of information that might be made subject to access restrictions were discussed during the workshop and can be summarized as follows.

Data from Bioterror Agents vs. Other Pathogens

This classification labels microorganisms on the basis of whether they have been designated as potential biological-terrorism threats. One approach to controlling access would be to withhold genomes of organisms that are on such a list of bioterror threat agents while continuing to release all others into the public domain; it was the original paradigm suggested as an example by the sponsors when the committee was assembled.

There was no support for this approach among workshop participants. It is too late, in that the sequences of most of the known bioterror threat agents, including all six Category A agents (anthrax, smallpox, botulinum toxin, plague, tularemia, and some viral hemorrhagic fevers), have already been released into the public domain. Moreover, free access to genome information about these agents is of tremendous value to research scientists who are attempting to create new countermeasures to combat them in case they are used in a bioterrorist attack. And pathogens not normally considered to pose bioterror risks might still be used by a bioterrorist, modified or not, in an attack on civilian populations.

Data from Naturally Occurring vs. Genetically Engineered Pathogens

Some participants suggested that even if all sequences for naturally occurring pathogens should be accessible, perhaps the sequence modifications for some genetically engineered organisms should not be. Support for distinguishing engineered from natural organisms was mixed. For example, it was argued that access to the changes in genome sequence that led to antibiotic resistance (either naturally occurring or selected in

the laboratory) should be restricted; in other words, restricting specific pieces of information might hinder potential terrorists in constructing potentially more dangerous microorganisms. It should be noted that many sequences for antibiotic resistance are already in the public domain, and in some cases the molecular basis of the resistance is well understood. Others argued that withholding such information would deprive the broader scientific community of insights that might be gained from understanding how specific genetic changes affect the properties of organisms and would impede understanding of the kinds of enhanced pathogens that might one day be created and released; these participants did not see a net advantage in saddling the current dynamic and productive system of scientific discovery with regulations that would slow the communication of results and ideas among legitimate investigators and thereby slow scientific progress.

Primary Genome Sequences vs. Annotations

Primary sequence data—the raw sequence of As, Ts, Gs, and Cs—are not particularly useful without the tools to analyze them. Annotations are the first level of analysis, so the question arises as to whether limiting access to the annotations might be more effective than withholding raw sequence files. Most participants thought that annotations were not in themselves dangerous. It was pointed out that up to one-third of the putative proteins encoded by putative genes in microbial genomes are unlike any that have been previously characterized, so no functions have been assigned to them. It is also clear to those who analyze genomes that the assigned functions are not necessarily all correct; for example, even though many genes are annotated as "virulence factors", such putative gene assignments are often not supported later by experimental data (Fraser, 2004). Therefore, a gene annotation alone may not be sufficient to assist someone who is seeking to increase a microorganism's virulence for weapons purposes.

Microarray and Other Functional Genomic Data

Databases that will archive functional genomic results from microarray experiments, such as the European Bioinformatics Institute ArrayExpress site mentioned in Chapter 2, are still in their formative stages. In the absence of centralized sites, some scientists routinely make microarray data available through their laboratory Web sites. Workshop participants indicated that these databases are not likely to be useful to potential terrorists now but may become so in the future—provided that a potential malefactor is sufficiently knowledgeable to detect the few useful

pieces of data scattered among hundreds of thousands of data points derived from a single experiment. Microarray data are notoriously hard to interpret; large amounts of data make analysis difficult, and it is challenging to tease apart results that are due to the intended variable and results that are due to factors for which there was not an adequate control. The scientific community today does not fully understand what the transcriptional data from microarray experiments mean with respect to cellular function, and it would be hard to put the data to practical use in enhancing a pathogen.

Tools for Analyzing Genome Data

It might be possible to distinguish access to genome data, such as primary sequences and annotations, from access to sophisticated analytic tools that allow the assembly of biological data into a coherent picture. Tools that link many kinds of biological data to computer programs that can be used to mine and analyze them are themselves among the most potent tools for conducting biological research ever constructed, (see, for example, the work being done by the Synthetic Biology group at Massachusetts Institute of Technology—www.syntheticbiology.org). As the power of computer systems that integrate various kinds of data grows, one might argue that it will become easier for someone to use these tools anonymously through the Internet to further attempts to enhance pathogens. By the same token, that risk is balanced by the even higher likelihood that the data and tools to analyze them will be used to create new therapies and prevention measures to control natural outbreaks and bioterror attacks.

The committee was charged with determining which types of pathogen-related genome data present the most concern. As evidenced by the categories above, it is possible to identify categories of data, but it is not clear that some types of data can be correlated with a specific level of risk of misuse for bioterrorist purposes. Data on all organisms present some level of concern but, although some organisms are inherently more dangerous, it does not necessarily follow that their genome sequences are more dangerous. The organisms themselves are beyond the scope of this study, and many organisms relevant here are governed by the select agent rules.

POTENTIAL DATA-CONTROL MECHANISMS

Access to digital data is notoriously difficult to limit to approved users. The recent experience of the recording and motion-picture industries with illicit transmission of copyrighted material is well documented. Files containing genome information would likewise be resistant to effec-

tive control by anything short of the most stringent restrictions. And like other kinds of digital data made available on the Internet, sequences, once released to the public domain, cannot be retrieved. All sequence information that has already been released resides on computer servers and in downloaded files on personal computers around the world. It would be impossible to legislate the return of those data from those who might be considered to be unauthorized users.

Even if the data were not difficult to control, whole-genome sequencing projects are becoming technically much easier and less expensive to carry out. For example, as noted, the genome of *Yersinia pestis* contains about 4 million nucleotides. At the 2003 price of about $0.02 per base, this genome could be sequenced for a marginal cost of about $80,000, assuming that the work were done at a well-equipped facility by experienced staff. If current trends continue, the cost will continue to decline (Carlson, 2003). This means that even if governments choose to attempt to limit access to sequence data, it would be feasible for those who are barred from such access to do the work themselves.

At the workshop, more time was devoted to discussion of the kinds of data that might be restricted and the possible costs and benefits of restriction than to the precise mechanisms by which restriction could be achieved. However, three possible strategies could be pursued.

Classify Some Data

The U.S. government has traditionally used a classification system to restrict access to information that poses a national security risk. Under this system one must obtain a government-issued security clearance to access classified information. A review of the U.S. system of classification is beyond the scope of this report. However, the committee acknowledges that there may well be sequence data that pose a risk to national security because of how the information will be used, not because of the inherent scientific information. For example, there are reasons not to publicize information that might expose vulnerabilities in environmental sensors based on the polymerase chain reaction (PCR) or plans for practical applications of medical countermeasures. We leave it to others to evaluate whether the current system is being used appropriately.

Withhold Some Data from Widespread Public Release

Detailed drawings of chemical-manufacturing plants and bioterrorism-emergency response plans for large cities are examples of information that is sometimes withheld. Similarly, research into the genetic and molecular basis of bacterial pathogenicity is of legitimate interest but might be con-

sidered sensitive by many people. For example, the sequences of several isolates of *Bacillus anthracis* generated by the National Institute of Allergy and Infectious Diseases and the Federal Bureau of Investigation during the investigation into the anthrax attacks of 2001 have not been publicly released. How to control access to information while allowing vigorous scientific inquiry is a difficult or impossible issue to resolve easily with legislation or judicial fiat. Some workshop participants suggested that at least a subset of genome data might be restricted and access to them accordingly limited to bona fide scientists as determined by some new oversight process. However, there was no consensus on the point, and most of the participants opposed such a step. Restricted access would require some sort of screening and registration of scientists authorized to use sensitive data by an as yet undefined process. The qualification process would have to be set up carefully to strike the proper balance between allowing scientists reasonably convenient access and screening out users who might be suspect.

Allow Unlimited Access but Require Registration

The workshop participants spent considerable time in discussing the merits of requiring users of genome databases and analytic tools to register with database administrators. To some, that would not amount to restricting access in that anyone could obtain access by answering a few questions. To others, it might be a substantial deterrent to making use of genome databases. A requirement for registration would constitute a major change from the current practice that allows users of many on-line databases to be unrestricted and entirely anonymous. If the United States enacted laws requiring registration, users of databases could potentially be tracked. That might help to deter malefactors, but it would be of concern in the competitive field of biological research. Scientists are often concerned that they will be "scooped" and another laboratory will be the first to publish. Moreover, pharmaceutical companies take great care to protect their early-stage investigations from competitors; companies' willingness to invest in drug discovery could decrease if others could determine what data they are using. In fact, pharmaceutical companies, large research centers, and others download many of the available data onto their own networks so that they can be used privately. Many of the data have been in the public domain for years and may well be stored in dozens or even hundreds of locations around the world. Given the international availability of the data, many people could access sequence information without relying on a database that requires registration.

At the workshop, representatives of the National Center for Biotechnology Information (NCBI) and The Institute for Genomic Research (TIGR)

stated unequivocally that any barrier between scientists and genome data would have a deleterious effect. For example, NCBI Director David Lipman cited the experience of a Web site called GeneTests that offers information about genetic tests and a peer-reviewed journal called *Gene Reviews* (www.genetests.org). Lipman said that when this organization removed registration requirements, use went up severalfold in a short period. Other workshop participants argued, however, that scientists could be persuaded to accept the relatively minor inconvenience of being required to register if they could be convinced that it would reduce the chance of bioterrorism and if their privacy interests could be protected by controlling who could access the registration information and searches associated with each user.

If properly instituted, requiring users of genome databases to register could provide data on who was accessing various types of information. Such tracking data might be used to investigate people's actions after they have been associated with a crime, and they might be used to identify malefactors in time to prevent them from acting. Alternatively, some type of automated program could be constructed to alert authorities to particular types of searches independently of the identities of the searchers. Such mechanisms would provide a public check on the actions of scientists and potential malevolent actors. For any registration system to be effective, however, broad international cooperation would be required.

SUMMARY OF ISSUES RELATED TO RESTRICTING ACCESS TO GENOME DATA

Restrictions on access would limit the ability of individuals or organized groups to use Internet-based genome databases and analytic tools to construct enhanced infectious-disease agents. Reagents and hardware for genetic manipulation of pathogens are increasingly easy to use and relatively straightforward to acquire. It is possible that specific kinds of data would provide a disproportionate advantage to malevolent users over benevolent users. Denying access to databases might deter or slow the progress of malefactors. Some research findings based on genome data might fall into the gray zone discussed earlier, for example, those which exploit vulnerabilities in measures to protect public health. Under a system in which some data are restricted, these specific results could be withheld for use only by designated persons.

On the other hand, open access to genome information preserves a fundamental principle of scientific inquiry, namely, that scientists must reveal, in exhaustive detail, what they found and how they found it. This principle allows working scientists to verify the accuracy of published scientific information, to design experiments to confirm scientific

hypotheses and to use work published by others to make new advances in their own research. Openness allows science to move faster, and this could lead to new biodefense strategies and products. It is impossible to predict who will benefit from having access to different kinds of data, and it could be argued that the data most likely to be restricted are those most important to biodefense research. Science relies on people's being open to unexpected connections, and these connections can offer opportunities for important scientific advances. The more available the data, the more likely that novel findings will be discovered. Another argument against U.S. restriction of access to genome databases concerns how the action would be viewed globally. International cooperation is facilitated by transparency. Restricting access to data could arouse suspicions of policymakers and security experts in other countries about the types of research being conducted in secret. Because of the similarities between some offensive and defensive research, some legitimate classified threat-analysis work conducted in the United States has already caused concern among our closest allies. Many feel that that it is safer to have results and data available to all so that others can verify or refute the results or question the propriety of continuing lines of research.

Requiring registration to access genome databases might be less controversial than directly restricting access to data in that the information would be available to all who were willing to identify themselves. Databases and some computer tools can be accessed anonymously without specialized equipment, and this accessibility has benefits to those who wish to use the data to create bioweapons. Requiring users to register may deter some potential malefactors from accessing the data and encourage them to move on to other activities. However, registration raises challenging ethical questions concerning the monitoring of database use. Consensus would need to be reached on when database use is analyzed, what constitutes suspicious activity, who is authorized to analyze use, and what actions will be taken in response to suspicious activity. A simple system of registration would not be useful for identifying those who might carry out bioterrorist acts. In addition to the ethical issues, it would be expensive to implement and maintain a system capable of providing informative data on its users. It would also be challenging to determine an efficient way to monitor users for suspicious activity. Effective use of registration would require the cooperation of those managing all known databases and perhaps the international sharing of registration mechanisms.

4

Conclusions and Recommendations

The conclusions and recommendations presented here are the results of deliberations of the committee. The recommendations are not intended to indicate a consensus of the workshop participants. They are based on information from the workshop, the expertise of the committee members, and published references. In committee discussions after the workshop, the members of the committee analyzed the information and debated among themselves on appropriate recommendations. In preparing the following recommendations, the committee considered the reality that advances in and technologies of life-science research could potentially be misused by individuals, groups, or nations to create agents capable of causing great harm. However, given that society has reason to fear natural outbreaks and intentional attacks, the committee concludes that biosecurity would be better served by policies that facilitate, not restrict, scientists' ability to understand infectious disease and to develop countermeasures to both naturally occurring pathogens and biodefense threats.

Recommendation 1: Policies with regard to release of genome data on microbial pathogens should not change. Rapid, unrestricted public access to primary genome sequence data, annotations of genome data, genome databases, and Internet-based tools for genome analysis should be encouraged.

Mechanisms currently exist to cope with sequence data obtained during criminal investigations or for specific intelligence or national-security reasons. The committee did not address any criteria presently used by

investigative or security organizations in control of genome data. Those situations are beyond the scope of the committee's charge. In situations where these specific exceptions do not apply the committee calls for unrestricted public access to genome data.

At the outset of the 21st century the possibility that life-science research might be perverted for destructive ends and that a pathogen could be deliberately enhanced and released to significant harm must be taken seriously. As understanding of host-pathogen interactions grows, national governments, subnational groups, or even individuals might well attempt to apply the growing power of biological science for destructive purposes, and it is possible that they could succeed. By the same token, as our understanding grows, the global health community has a greatly enhanced ability to produce new anti-infective drugs, vaccines, and diagnostic reagents.

The primary question before this committee is which policies regarding release of genome data about pathogens will provide the greatest overall biological security. That question takes into account both the possibility of deliberate pathogen enhancement and release and the fact that new and dangerous pathogens will continue to emerge naturally. After careful deliberation, the committee concluded that preserving open access to genome data and free exchange of knowledge and results that flow from the data will, by a substantial margin, increase biosecurity. Therefore the committee recommends no expansion in the amount of genome data which is classified and no change in the extent of material withheld from widespread public release, they also recommend that no registration system be imposed. The committee's reasoning as it came to that conclusion focused on three sets of arguments:

Current Policies Are Effective. Unfettered, free access to the results of life-science research is the historic norm and has served science and society remarkably well. Open access allows life scientists everywhere to evaluate, interpret, adapt, and extend results from many fields of inquiry for use in their own work and thereby accelerates research and speeds the delivery of life-saving benefits that biological and medical research are so rapidly creating. Science builds on itself, and the sharing of methods and data allows scientists to learn from the work of others and to make unexpected connections. There is no obvious way to predict which scientists will benefit from access to which data, so restricting access poses a risk of slowing the progress of research. The current vigor in the life sciences depends on the free flow of data and ideas, and it is necessary if science is to deliver needed new biodefense capabilities.

Current policies allow for the most rapid and effective scientific response possible during an infectious-disease crisis, such as the SARS outbreak of 2003. At such times, when scientific and public-health

resources must be rapidly mobilized to combat a poorly understood emerging disease, free and rapid exchange of data, results, and ideas is essential to allow scientists to communicate effectively and to build on one another's findings. Restrictions on the flow of information in such circumstances would slow the acquisition of understanding of the emerging infection and the development of countermeasures against it.

A security-classification mechanism already exists to deal with specific and unusual cases in which genome data should be withheld. The committee has not examined the current system used to determine which information should be classified, but it believes that a government system of classification should be adequate for restricting disclosure and that an additional system of control is not necessary to address security needs.

Some genome-sequence data acquired with federal funds have not been released immediately. For example, anthrax sequences that were obtained during the criminal investigation of the anthrax attacks of 2001 have not yet been released and might not be for some time, perhaps not until a legal case is resolved. Some sequence data have been classified, such as the sequences of certain PCR primers designed to be used in environmental sensors; this was done to reduce the likelihood that pathogens would be altered to make the primers useless. Mechanisms now in place to cope with sequence data obtained during criminal investigations or for specific intelligence or national-security reasons should not be used to limit scientific research but may be necessary to cope with unusual situations in the future.

Effective Restriction of Genome Data Is Not Practical. As a practical matter, restricting access to genome data would be difficult, expensive, and probably counterproductive. First, it is notoriously difficult to control access to digital data. Files that contain entire genomes are not particularly large—generally, several megabytes—and are easily stored, transferred, and exchanged. Second, in the absence of a uniform international agreement to impose similar control measures worldwide, potential users who are denied access because of U.S. policy could direct their Internet browsers to genomics sites in other countries that have the same kind of data. A global consensus on how to implement policies that would be strong enough to keep relevant data out of the hands of potential malefactors would be difficult to achieve. Third, any policy stringent enough to reduce the chance that a malefactor would access data would probably also impede legitimate scientists in using the data and would therefore slow discovery. Penalties would be required to ensure that restrictions were obeyed, and an oversight organization would have to be created to monitor compliance. The international framework needed to make the policies workable would be cumbersome and expensive. It is possible that the

harm done during a process of negotiating such an agreement—through building walls of mistrust between peoples—would be greater than the benefit gained through the sense of security that such a regime might provide. Finally, such a restrictive regime, the committee believes, could seriously damage the vitality of the life sciences.

At the outset of the workshop, the concept of requiring all users of genome databases to register to gain access seemed to many participants to be a reasonable policy compromise. Under such a policy, anyone could gain access but only after stating a name, address, and institutional affiliation. After additional discussion, however, the committee concluded that a registration requirement of this kind would not be an effective way of protecting society from bioterrorism. Registration would not prevent a determined malefactor from accessing genome databases. Registration might deter a less determined malefactor or provide a mechanism for tracing his or her activities, but it would also raise many troubling questions about who could use registration information and under what circumstances. In addition, the lack of an international consensus that registration should be required would render such measures futile. It seems unlikely that a uniform agreement could be generated between all public and private database managers and others who generate genome data, which would be necessary to track those with access to genome sequences. In addition, scientists are wary of efforts to track their use of genome data, especially in the competitive field of biological research. Some are concerned that another laboratory will figure out what they are working on and be the first to publish. Pharmaceutical companies are cautious about protecting their early-stage investigations from competitors; companies' willingness to invest in drug discovery could decrease if others could determine what data they are using. In addition, the fact that pharmaceutical companies, large research centers, and others download many of the available data onto their own networks so that they can be used privately would hinder the usefulness of attempts to track discrete queries to databases. Many of the data have been in the public domain for years and may well be stored in dozens or even hundreds of locations around the world. Given the international availability of the data, many people could access sequence information without relying on a database that requires registration. For all the above reasons the committee feels that the benefits of registration do not outweigh the costs to society from the resulting slowdown in research on infectious diseases.

Pathogen Genome Sequences Are Not Uniquely Dangerous. Primary sequence data on pathogens become dangerous only if the user has a sophisticated ability to exploit them and a malevolent goal. Mere possession of the sequence of a pathogen does not confer the ability to

enhance the virulence of the organism to which it pertains, nor would it help to solve the demanding technical problems associated with conducting a terrorist attack. Although a potential malefactor might be able to adapt published research results that reveal genetic manipulations that would enhance the virulence of a pathogen, discovering which genetic change would enhance virulence is difficult and would require a substantial and sophisticated effort.

Pathogen genome sequences are not uniquely dangerous, because sequence data from non-pathogenic organisms could also be used to enhance a pathogen's virulence or create a new pathogen. For example, sequence data from a close relative of a pathogen, such as *Bacillus cereus*, could be useful to a potential terrorist. Alternatively, sequence data from a pathogen's host could be used to engineer a microorganism. For example, mousepox virus has been shown to become more virulent when engineered to contain the gene for the mouse version of interleukin-4, an immune regulator. Moreover, many nonpathogenic symbiotic or commensal microorganisms could conceivably be made pathogenic by the addition of genes that encode human immune system signals; this would disrupt the normal immune response and allow otherwise harmless bacteria to cause harm. Access to all pathogen and host sequences could not be restricted without severely damaging the fabric of the entire global scientific enterprise; such information is essential to current life-science research efforts. In the end, the availability or nonavailability of a pathogen's genome sequence will not deter a dedicated actor from using a naturally occurring infectious strain in a terrorist attack.

The above discussion focuses mainly on naturally occurring organisms. For the most part, the issues are the same for genetically engineered organisms. Information on the altered sequences and the resulting phenotypes can provide insight into basic biology, and most alterations are not particularly useful to a potential bioterrorist. However, sequence data on some genetically engineered organisms could be useful for a potential bioterrorist attempting to create a more dangerous pathogen. Regulation of the conduct of the experiments that might generate such engineered organisms is beyond the charge to this committee, but it is an important issue. Decisions on the appropriateness of conducting particular experiments should ideally be made before the experiments are begun. Local institutional review boards (IRBs) play a large role in that process, and the newly announced National Science Advisory Board for Biosecurity (NSABB), discussed in detail in Recommendation 3, will play a growing role in that regard. The guidelines for IRBs and codes of conduct for individual scientists that the NSABB envisions should help to ensure that appropriate consideration is given to the potential implications of research

approaches before they are begun. In addition, journal editors have a responsibility to consider the national-security implications of the papers they publish. Given all those caveats, if an experiment is published, the accompanying genome data should not be restricted by regulations. The data are essential for others to understand the significance of the research and may be crucial to future experiments that could help to protect us from disease. There is some concern that restricting access to this information might lead to a situation in which the mainstream scientific community is unaware of dangers that may threaten us. And some have proposed that observing changes in the frequency of publications (and conference presentations) of potential malefactors can provide useful clues as to whether they are conducting secret experiments.

For all those reasons, the committee concluded that maintaining the current standard of free access to all genome data is the best policy choice. The problem with which the committee has been charged is not to strike the correct balance between security and openness. That is a false dichotomy; openness has enhanced security in the past and is the best way to ensure security in the future. Instead, the most important task is to be as well prepared as possible to cope with the serious infectious-disease threats that society is sure to face in the coming century, both natural and human-made. The committee believes firmly that the policies currently in place for genome data—immediate release and free access—are correct because openness is essential to maintain the progress needed to stay ahead of those who would attempt to cause harm.

Recommendation 2: Genomics and genome sequence data should be exploited fully to improve our ability to defend against infectious agents of all types, including those which contribute to epidemic diseases and infant mortality and the naturally occurring or genetically enhanced organisms that could be used in a bioterrorist attack.

Maximizing the benefit from research on infectious diseases is important for both public-health and national-security reasons. Even before the increase in attention to national security that followed the attacks of 2001, the U.S. government had considered infectious disease as a security issue. For example, in testimony to Congress on June 29, 2000 (Gordon, 2000), a national intelligence officer discussed the possibility of bioterrorism; the threat to public health from importation of diseases; the impact of troop health on U.S. military readiness; the ability of tuberculosis, malaria, and AIDS to slow economic development and undermine social structures in some regions; and the potential harm from infectious-disease-related embargoes and restrictions on travel.

Since the terrorist attacks of 2001, federal spending intended to improve defenses against bioterrorism and natural infectious-disease out-

breaks has increased markedly. For FY 2003, the total expenditure for bioterrorism preparedness was $5.9 billion (http://www.niaid.nih.gov/biodefense/about/nbe.htm). Of that total, about $1.75 billion was spent on biodefense-related research, most of it channeled through the National Institute of Allergy and Infectious Diseases (Fraser, 2004). Indeed, many of the pathogen whole-genome sequencing efforts that have been recently completed or begun have been funded with money earmarked for biodefense.

Research exploiting the revolution in genomics has an important role to play in increasing our ability to defend against infectious agents of importance to biodefense and global infectious disease. Indeed, research on many of the currently important societal infectious threats, such as antibiotic-resistant bacterial pneumonia and antibiotic-resistant staphylococcal disease, will benefit enormously from the genomic revolution. Extensive sequence comparisons between pathogenic and nonpathogenic organisms, studies of changes in the pattern of gene expression in pathogens and their hosts as they interact, and sequencing of multiple strains of specific pathogens will all contribute to the development of new diagnostics, vaccines, and therapeutics for disease-causing organisms, including those which might be used in a bioterror attack. It is important to maintain policies that allow all medical and agricultural scientists, including those who focus on biodefense, to use genome data to the fullest extent possible in their research. The genomics revolution includes not only human pathogens and their hosts but also the infectious agents that plague agricultural crops and livestock. These are also of critical importance for our economy and for our national security. The biodefense effort should include both human pathogens and pathogens that might be deployed against agricultural interests.[1]

Recommendation 3: Future advances in genome science should be regularly reviewed to keep all relevant government departments and agencies apprised of new developments that may affect national security. Regular meetings of scientific and security experts should be held to discuss the implications of new developments and to develop coherent responses. The newly formed National Science Advisory Board for Biosecurity or another appropriate entity with the ability to connect with diverse federal agencies would be a suitable home for that function.

The pace of scientific progress creates a need for continuous and thorough evaluation of science and technology as they affect national security

[1]For more information about biodefense and agriculture, see the 2003 National Research Council report *Countering Agricultural Bioterrorism.*

and the health and welfare of all the inhabitants of the planet. Decisions about policies related to pathogen genomes cannot be properly made unless they are considered in the context of other scientific advances. New developments in law enforcement, forensics, and public health based on continued research may provide better approaches to improving biosecurity than attempts to restrict access to genome data. However, a mechanism is needed to ensure adequate communication between the scientific and security communities. People in the scientific and security communities bring to the table their own cultures and experiences. Extensive discussions are necessary for each to be able to understand the other's perspectives. This sharing of perspectives is crucial if future policy decisions are to reflect the best possible input. Improved communication will help to guide scientific research in fields that will facilitate biodefense, and it will help security experts base their actions on the latest science.

A well-informed body with both scientific and security expertise should provide an accessible link between the scientific and security communities and review advances in genome science in case future developments warrant the creation of additional monitoring of or restrictions on access to genome data. The new body would serve as a communication mechanism between the scientific and security communities and help to decrease the likelihood that new developments will come as a surprise. Extensive participation of the security community in this activity would be a concrete action that could be taken as part of its data-gathering work.

Review should be scientifically broad because the effect of genomics on biosecurity goes far beyond the biology of biothreat organisms and includes both biomedical topics, such as drug and vaccine development, and topics pertaining to forensics, intelligence, agriculture, and the environment. Limiting the evaluation to direct studies on genomes of pathogens would not adequately address threats to biosecurity. Knowledge of the genomes of infectious agents that might be used as weapons of bioterror is obviously important, but the genomes of potential hosts (humans, other animals, and plants) also offer opportunities for manipulation. Over the next 10 years, scientists may learn at least as much about the genetic and molecular basis of genetic resistance and susceptibility to infection as about specific microbial virulence factors and their function. The perspective of those involved in basic research related to humans, plants, animals, and microorganisms is essential for staying on top of new developments that may affect biosecurity. Continuing review of new technology could include the use of functional genomics as it pertains to understanding microbial virulence; host susceptibility and resistance to infectious diseases of plants, domestic animals, and humans; and relevant aspects of the development of new drugs, vaccines, and anti-infective therapies.

To be well informed, the review body must be part of a network for information exchange among academe, industry, international actors, and U.S. government agencies, including those in the threat-response community. Coordination of efforts in all arenas, including the international community and those involved in threat response, would provide a means of assessing the significance of advances in genome research in terms of both increased threats to security and improvements in understanding of the environment and human health and disease. As an additional benefit, providing a network for information exchange would help to further research in disease diagnosis and epidemiologic surveillance on a national and global basis and facilitate communication of information required for the unambiguous identification and attribution of pathogens in forensics.

There are several options for implementing action on the functions described above. For example, a new entity could be created or an existing entity modified. The committee is not aware of any existing entities that would have access to both the scientific expertise and the broad network described above. However, the newly announced National Science Advisory Board for Biosecurity (NSABB) might be a suitable home for those tasks, depending on the focus it takes as it is established. On March 4, 2004, the Department of Health and Human Services (DHHS) announced at a press conference and on its Web site (www.biosecurityboard.gov) the creation of the NSABB to "advise all Federal departments and agencies that conduct or support life sciences research that could fall into the dual use category. The NSABB will be managed by the National Institutes of Health." In announcing the creation of the board, John Marburger, director of the White House Office of Science and Technology Policy, said that "it is imperative that we develop this new framework to address serious concerns that range from personal responsibility to national security."

The NSABB was created in response to Recommendation 4 in the recent National Research Council report (2003a) *Biotechnology Research in an Age of Terrorism.*

> We recommend the formation of a National Science Advisory Board for Biodefense (NSABB)[2] to provide advice, guidance, and leadership for the system of review and oversight that we are proposing. The NSABB would serve a number of important functions for both the scientific community and the government. At the most general (strategic) level, it would serve as a point of continuing dialogue between the scientific community and the national security community and as a forum for addressing issues of interest or concern. At the operational (tactical) level, it would provide case-specific advice on the oversight of research and the

[2]The name of the board announced on March 4 differs slightly from that proposed in the 2003 National Research Council report.

communication and dissemination of life sciences research information that is relevant for national security and biodefense purposes.

The DHHS announcement states that the NSABB will

Advise on strategies for local and federal biosecurity oversight for all federally funded or supported life sciences research.

Advise on the development of guidelines for biosecurity oversight of life sciences research and provide ongoing evaluation and modification of these guidelines as needed.

Advise on strategies to work with journal editors and other stakeholders to ensure the development of guidelines for the publication, public presentation, and public communication of potentially sensitive life sciences research.

Advise on the development of guidelines for mandatory programs for education and training in biosecurity issues for all life scientists and laboratory workers at federally-funded institutions.

Provide guidance on the development of a code of conduct for life scientists and laboratory workers that can be adopted by federal agencies as well as professional organizations and institutions engaged in the performance of life sciences research domestically and internationally.

The NSABB will have up to 25 voting members, to be appointed by the DHHS Secretary in consultation with the heads of relevant federal departments and agencies. Members will be experts in a broad range of fields, including molecular biology, microbiology, infectious diseases, laboratory biosafety and biosecurity, public health/epidemiology, health physics, pharmaceutical production, veterinary medicine, plant health, food production, bioethics, national security, biodefense, intelligence, law and law enforcement, and scientific publishing. The board will also include nonvoting ex officio members from at least 15 federal departments and agencies.

The following agencies were included in the announcement and are expected to be involved in the NSABB: the Executive Office of the President, DHHS, the Department of Energy (DOE), the Department of Homeland Security (DHS), the Department of Veterans Affairs, the Department of Defense, the Department of the Interior, the Environmental Protection Agency, the U.S. Department of Agriculture, the National Science Foundation (NSF), the Department of Justice (DOJ), the Department of State, the Department of Commerce, and the National Aeronautics and Space Administration. The intelligence community is also expected to participate.

There are pros and cons to locating the responsibility for review of genome data in the NSABB. For example, it may not be possible for the NSABB to manage the necessary continuing review of genome information while establishing guidelines for the oversight of biological research. A dedicated subcommittee of the NSABB might be formed to review developments and keep the NSABB as a whole informed. Another issue in assigning the above functions to the NSABB is that the board will be managed by the National Institutes of Health (NIH), and this may hamper its ability to view biosecurity-related issues from all the necessary perspectives. The review of genome research envisioned will require the gathering and analysis of diverse opinions. The partnership and full participation of each of the agencies is crucial to ensure that all the available information and insight are used.

A useful example of cooperation between agencies already exists in the National Interagency Genomics Sciences Coordinating Committee (NIGSCC), which meets on an ad hoc basis and has proved effective in maintaining close contact between the various government agencies with interests in genome research, including NSF, NIH, the Federal Bureau of Investigation, DOJ, the Centers for Disease Control and Prevention, the Central Intelligence Agency, the Defense Advanced Research Projects Agency, DOE, DHS, the U.S. Army, and USDA. The NIGSCC provides a useful model of how the understanding of interdisciplinary issues can be advanced and profited from by successful collaboration among individuals and agencies with diverse perspectives. The NIGSCC, however, has no formal authority to carry out actions that it deems necessary, nor does it include representatives of academe, industry, or international bodies; so it is not ideal for the purpose the committee suggests. The entity that becomes responsible for reviewing scientific advances in genome science for their potential effect on national security must be scientifically respected, have the ability to integrate information from diverse sources, and have the ability to influence discussions in numerous federal departments and agencies.

Recommendation 4: The committee endorses Recommendation 7 of *Biotechnology Research in an Age of Terrorism*, which calls for an international forum to unify the discussion on the effect of genomics on biosecurity.

Life-science research is global, and no single nation can successfully implement policy concerning access to and release of life-science data and results without reference to the rest of the international community. For that reason, it is of the utmost importance that the international community establish a common understanding of security concerns and shared

resources to make the most efficient and safest use of genome data and experimental results, some of which might suggest how pathogens could be successfully enhanced. The committee therefore strongly endorses Recommendation 7 of *Biotechnology Research in an Age of Terrorism*, which calls for "the international policy-making and scientific communities [to] create an international forum on biosecurity to develop and promote harmonized national, regional, and international measures that will provide a counterpart to the system [recommended] for the United States." If conducted openly and in the proper spirit, the process of discussing these issues might actually build understanding, and some trust, among the nations involved and eventually help to establish an international norm against misuse of genetic information.

Since the release of *Biotechnology Research in an Age of Terrorism*, plans have begun for the International Forum on Biosecurity. The event will be coordinated by the Policy and Global Affairs Division of the National Academies and funded by the Sloan Foundation and the Nuclear Threat Initiative. An international steering committee will be formed to develop plans for the forum. The committee membership will include experts with current or past ties to existing international organizations working in this field. The details of the forum will be worked out by the members of the committee as they engage in several outreach activities. Its three key objectives are as follows:

- To advance awareness in the life-science community and the international scientific community about the critical challenges posed by the dual-use dilemma.
- To solidify the commitment of leading scientific organizations to make biosecurity issues part of their regular programming. The forum will also serve as a showcase for the results of other meetings on biodefense and for the programs of major organizations.
- To serve as a major convening and coordinating mechanism for the scientific and policy-making communities. For example, a number of organizations already have or will be developing codes of conduct, some with an eye to the meeting of Biological Weapons Convention States Parties in the summer and fall of 2005 and some for their own purposes. The forum will provide the opportunity to bring these efforts together and to think strategically about how to maximize their impact.

The committee applauds the new initiative and encourages all parties to participate in the activities of the forum to advance the goal of promoting coordination and synergy by linking other efforts headed by established organizations with developed constituencies. In the same way that the

United States should maintain an endeavor in this field (see Recommendation 3), it will be important for international cooperation and coordination to be maintained.

Recommendation 5: The committee endorses Recommendation 1 of *Biotechnology Research in an Age of Terrorism*, which calls for national and international professional societies and related organizations to work to educate scientists about the risk that life-science research results will be misused and about scientists' responsibility to mitigate the risk.

Recommendation 1 of *Biotechnology Research in an Age of Terrorism* calls for national and international professional societies and related organizations and institutions to create programs to educate scientists about the dual-use dilemma in biotechnology and their responsibilities to mitigate its risks. As noted under our Recommendation 1 above, we believe that although the risk that the growing power of biological and medical research could be applied to destructive purposes is unknown, it is not zero. All life scientists must be sensitized to the potential for the harmful misuse of the knowledge they create.

The committee recognizes and applauds the efforts of numerous professional societies to educate their members and the public about these issues, and it suggests that such professional societies are the natural home for further efforts in this respect. They should expand efforts to engage their members in discussion of the potential benefits and dangers of the widespread availability of genome sequences and functional genomics data.

At this writing, the U.S. government has announced that the mission of the NSABB will include the development of professional codes of conduct for scientists and laboratory workers that can be adopted by professional organizations and institutions engaged in life-science research and the development of materials and resources to educate the research community about effective biosecurity (www.biosecurityboard.gov). The work of the NSABB will provide an important opportunity for the professional societies to work with the government so that the educational opportunities provided and the guidelines produced will be most effective.

The committee recommends that professional codes of conduct explicitly require scientists to act to mitigate the risk of misuse of scientific progress to cause environmental or medical harm and require them to carry out their research with integrity to minimize the risk of misuse of life-science research for destructive purposes.

References

AAAS (American Association for the Advancement of Science). 2003. NIH Budget Growth Slows to 2 Percent in FY 2004. February 25, 2003. [Online]. Available: http://www.aaas.org/spp/rd/nih04p.pdf

Adu-Bobie, J., B. Capecchi, D. Serruto, R. Rappuoli, and M. Pizza. 2003. Two years into reverse vaccinology. Vaccine 21(7): 605-610.

Amsellem, Z., B.A. Cohen, and J. Gressel. 2002. Engineering hypervirulence in a mycoherbicidal fungus for efficient weed control. Nature Biotechnology 20(10): 1035-1040.Atlas, R.M. 2002. National security and the biological research community. Science 298: 753-754.

Atlas, R.M., P. Campbell, N.R. Cozarelli, G. Curfman, L. Enquist, G. Fink, A. Flanagin, J. Fletcher, E. George, G. Hammes, D. Heyman, T. Inglesby, S. Kaplan, D. Kennedy, J. Krug, R. Levinson, E. Marcus, H. Metzger, S.S. Morse, A. O'Brien, A. Onderdonk, G. Poste, B. Renault, R. Rich, A. Rosengard, S. Salzburg, M. Scanlan, T. Shenk, H. Tabor, H. Varmus, E. Wimmer, and K. Yamamoto. 2003a. Uncensored exchange of scientific results. Proceedings of the National Academy of Sciences 100(4): 1464.

Atlas, R.M., P. Campbell, N.R. Cozarelli, G. Curfman, L. Enquist, G. Fink, A. Flanagin, J. Fletcher, E. George, G. Hammes, D. Heyman, T. Inglesby, S. Kaplan, D. Kennedy, J. Krug, R. Levinson, E. Marcus, H. Metzger, S.S. Morse, A. O'Brien, A. Onderdonk, G. Poste, B. Renault, R. Rich, A. Rosengard, S. Salzburg, M. Scanlan, T. Shenk, H. Tabor, H. Varmus, E. Wimmer, and K. Yamamoto. 2003b. Statement on the consideration of biodefence and biosecurity. Nature 421: 771.

Atlas, R.M., P. Campbell, N.R. Cozarelli, G. Curfman, L. Enquist, G. Fink, A. Flanagin, J. Fletcher, E. George, G. Hammes, D. Heyman, T. Inglesby, S. Kaplan, D. Kennedy, J. Krug, R. Levinson, E. Marcus, H. Metzger, S.S. Morse, A. O'Brien, A. Onderdonk, G. Poste, B. Renault, R. Rich, A. Rosengard, S. Salzburg, M. Scanlan, T. Shenk, H. Tabor, H. Varmus, E. Wimmer, and K. Yamamoto. 2003c. Statement on scientific publication and security. Science 299: 1149.

Carlson, R. 2003. The pace and proliferation of biological technologies. Biosecurity and Bioterrorism 1(3): 203-214.

Check, E. 2002. Biologists apprehensive over U.S. moves to censor information flow. Nature 415(6874): 821.

Duggan, D.J., M. Bittner, Y. Cheng, P. Meltzer, and J.M. Trent. 1999. Expression profiling using cDNA microarrays. Nature Genetics 21(supplement): 10-14.

Epstein, G.L. 2001. Controlling biological warfare threats: resolving potential tensions among the research community, industry, and the national security community. Critical Reviews in Microbiology 27(4): 321-354.

Fiers, W., R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, F. Molemans, A. Raeymaekers, A. Van den Berghe, G. Volckaert, and M. Ysebaert. 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature 260(5551): 500-507.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.-F. Tomb, B.A. Dougherty, J.M. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. Fields, J.D. Gocyne, J. Scott, R. Shirley, L.-I. Liu, A. Glodek, J.M. Kelley, J.F. Weidman, C.A. Phillips, T. Spriggs, E. Hedblom, M.D. Cotton, T.R. Utterback, M.C. Hanna, D.T. Nguyen, D.M. Saudek, R.C. Brandon, L.D. Fine, J.L. Fritchman, J.L. Fuhrmann, N.S.M. Geoghagen, C.L. Gnehm, L.A. McDonald, K.V. Small, C.M. Fraser, H.O. Smith, and J.C. Venter. 1995. Whole-genome sequencing and assembly of *Haemophilus influenza* Rd. Science 269(5223): 496-512.

Fraser, C.M. 2004. A genomics-based approach to biodefence preparedness. Nature Reviews Genetics 5(1): 23-33.

Frazer, K.A., L. Elnitski, D.M. Church, I. Dubchak, and R.C. Hardison. 2003. Cross-species sequence comparisons: a review of methods and available resources. Genome Research 13: 1-12.

Gordon, D.F. 2000. Importation of Infectious Diseases into U.S. Testimony before House Committee on International Relations. June 29, 2000.

Grifantini, R., E. Bartolini, A. Muzzi, M. Draghi, E. Frigimelica, J. Berger, G. Ratti, R. Petracca, G. Galli, M. Agnusdei, M.M. Giuliani, L. Santini, B. Brunelli, H. Tettelin, R. Rappuoli, F. Randazzo, and G. Grandi. 2002. Previously unrecognized vaccine candidates against group B meningococcus identified by DNA microarrays. Nature Biotechnology 20(9): 914-921.

Hood, L., and D. Galas. 2003. The digital code of DNA. Nature 421: 444-448.

Isberg, R.R., and S. Falkow. 1985. A single genetic locus encoded by *Yersinia pseudotuberculosis* permits invasion of cultured animal cells by *Escherichia coli* K-12. Nature 317(6034): 262-264.

Kanehisa, M., and P. Bork. 2003. Bioinformatics in the post-sequence era. Nature Genetics 33(supplement): 305.

Kaplan, D.E. 2000. Aum Shinrikyo. Pp. 207-220 in Toxic Terror: Assessing Terrorist Use of Chemical and Biological Weapons. J.B. Tucker, ed. Cambridge, MA: MIT Press.

Kennedy, D. 2003. Two cultures. Science 299(5610): 1148.

Kwik, G., J. Fitzgerald, T.V. Inglesby, and T. O'Toole. 2003. Biosecurity: responsible stewardship of bioscience in an age of catastrophic terrorism. Biosecurity and Bioterrorism 1(1): 27-35.

Lifton, R.J. 1999. Destroying the World to Save It: Aum Shinrikyo, Apocalyptic Violence, and the New Global Terrorism. New York: Henry Holt.

Miller, J., W. Broad, and S. Engelberg. 2002. Germs: Biological Weapons and America's Secret War. New York: Simon and Schuster, Inc.

Mora, M., D. Veggi, L. Santini, M. Pizza, and R. Rappuoli. 2003. Reverse vaccinology. Drug Discovery Today 8(10): 459-464.

Moxon, R., and R. Rappuoli. 2002. Bacterial pathogen genomics and vaccines. British Medical Bulletin 62: 45-58.

NRC (National Research Council). 2003a. Biotechnology Research in an Age of Terrorism. Washington, DC: The National Academies Press.

NRC (National Research Council). 2003b. Countering Agricultural Bioterrorism. Washington, DC: The National Academies Press.

Parkhill, J., and C. Berry. 2003. Relative pathogenic values. Nature 423: 23-25.

Petro, J.B., and D.A. Relman. 2003. Understanding threats to scientific openness. Science 302(5652): 1898.

Pizza, M., V. Scarlato, V. Masignani, M.M. Giuliani, B. Arico, M. Comanducci, G.T. Jennings, L. Baldi, E. Bartolini, B. Capecchi, C.L. Galeotti, E. Luzzi, R. Manetti, E. Marchetti, M. Mora, S. Nuti, G. Ratti, L. Santini, S. Savino, M. Scarselli, E. Storni, P. Zuo, M. Broeker, E. Hundt, B. Knapp, E. Blair, T. Mason, H. Tettelin, D.W. Hood, A.C. Jeffries, N.J. Saunders, D.M. Granoff, J.C. Venter, E.R. Moxon, G. Grandi, and R. Rappuoli. 2000. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science 287: 1816-1820.

Rappuoli, R., and A. Covacci. 2003. Reverse vaccinology and genomics. Science 302: 602.

Rosqvist, R., M. Skurnik, and H. Wolf-Watz. 1988. Increased virulence of *Yersinia pseudotuberculosis* by two independent mutations. Nature 334(6182): 522-524.

Salyers, A. 2002. Science, censorship, and public health. Science 296(5568): 617.

Schrage, M. 2003. Hobbyists play a critical role in the design and diffusion of technology. MIT Technology Review. June, 2003. [Online]. Available: http://www.technologyreview.com/articles/schrage0603.asp

Tettelin, H., N.J. Saunders, J. Heidelberg, A.C. Jeffries, K.E. Nelson, J.A. Eisen, K.A. Ketchum, D.W. Hood, J.F. Peden, R.J. Dodson, W.C. Nelson, M.L. Gwinn, R. DeBoy, J.D. Peterson, E.K. Hickey, D.H. Haft, S.L. Salzberg, O. White, R.D. Fleischmann, B.A. Dougherty, T. Mason, A. Ciecko, D.S. Parksey, E. Blair, H. Cittone, E.B. Clark, M.D. Cotton, T.R. Utterback, H. Khouri, H. Qin, J. Vamathevan, J. Gill, V. Scarlato, V. Masignani, M. Pizza, G. Grandi, L. Sun, H.O. Smith, C.M. Fraser, E.R. Moxon, R. Rappuoli, and J.C. Venter. 2000. Complete genome sequence of *Neisseria meningitides* serogroup B strain MC58. Science 287: 1809-1815.

Torok, T.J. R.V. Tauxe, R.P. Wise, J.R. Livengood, R. Sokolow, S. Mauvais, K.A. Birkness, M.R. Skeels, J.M. Horan, and L.R. Foster. 1997. A large community outbreak of salmonellosis caused by intentional contamination of restaurant salad bars. Journal of the American Medical Association 278(5): 389-395.

Venter, J.C., S. Levy, T. Stockwell, K. Remington, and A. Halpern. 2003. Massive parallelism, randomness, and genomic advances. Nature Genetics 33(supplement): 219-227.

Vastag, B. 2003. Openness in biomedical research collides with heightened security concerns. Journal of the American Medical Association 289: 686-690.

Voorhis, D.L., S. Dillon, S.B. Formal, and R.R. Isberg. 1991. An O antigen can interfere with the function of the *Yersinia pseudotuberculosis* invasin protein. Molecular Microbiology 5(2): 317-325.

Weiss, R. 2003. Engineered Virus Related to Smallpox Evades Vaccine. The Washington Post. November 1, 2003.

Appendix A

Statement of Task

STATEMENT OF TASK

The National Research Council will convene an ad hoc steering committee to oversee a one-day workshop that will identify issues surrounding the release to the public domain of genome data for bioterrorism threat agents. Biological agents considered at the workshop will include those on national "select agent" lists, as well as those that are not but could be considered plausible threats in the future. Questions addressed at the workshop will include but are not limited to the following:

- What are the categories of genome data that present the greatest concern?
- What are the pros and cons of unlimited versus restricted access to such data, including threats posed to the scientific community or to national security?
- What are some options for making decisions about release to the public domain?

The committee will author a report based upon the workshop. The report will 1) capture input from presentations and discussions by workshop participants, 2) identify general issues surrounding the release to the public domain of genome data for bioterrorism threat agents, 3) develop a list of pros and cons associated with the release to the public domain of such data, and 4) present recommendations and/or ideas about policy

options or decision-making frameworks for release to the public domain of pathogen- or pest-related genome information. The workshop will be designed to complement and reinforce related National Academies activities, namely the January 9, 2003 workshop on "Dual-use" Information in the Life Sciences and the ongoing study on Improving Research Standards and Practices to Prevent the Misuse of Biotechnology Research.

Appendix B

Agenda

AGENDA FOR OCTOBER 1, 2003, WORKSHOP OF THE NATIONAL RESEARCH COUNCIL COMMITTEE ON GENOMICS DATABASES FOR BIOTERRORISM THREAT AGENTS

National Academy of Sciences Lecture Room, 2100 Constitution Avenue, NW, Washington, D.C.

8:30 am Welcome from National Academies, and committee chair Stanley Falkow
Overview of charge to committee and goals for the day

9:00 am What database resources are available today and how are they used?
What policies affect their content? Are the answers different in the United States vs. abroad?
David Lipman, NCBI
Rino Rappuoli, Chiron Italy (Industry Perspective)
Rob Heckert, USDA (Agriculture Perspective)

11:00 am International Perspective on data release (with request to touch on legal issues)
Sir Bob May, Royal Society
Michael Morgan, Wellcome Trust

12:15 pm Wrap up and plans for the afternoon
Stanley Falkow

12:30 pm LUNCH-at assigned tables each with its own topic and with a committee member as leader
Topic 1: Security impact of free release
Topic 2: Scientific impact of restricted release
Topic 3: Potential mechanisms for controlling release

2:00 pm Reporting back from the lunch discussions (15 minutes per topic)

2:45 pm Can we classify genome data by threat level? Would this be based on characteristics of the organism or characteristics of the data (such as annotation)?
David Relman, Stanford
Discussant: Art Friedlander

3:15 pm BREAK

3:30 pm Revisit issues from the morning: What are the pros and cons of unlimited vs. restricted access to data, including threats posed to the scientific community or to national security?

4:00 pm Wrap-up talks summarizing the day's ideas
Tara O'Toole, Johns Hopkins (policy perspective) AND
David Relman (biology perspective)

4:30 pm ADJOURN

Appendix C

Participants

PARTICIPANTS IN THE OCTOBER 1, 2003, WORKSHOP OF THE NATIONAL RESEARCH COUNCIL COMMITTEE ON GENOMICS DATABASES FOR BIOTERRORISM THREAT AGENTS

Corrie Brown, University of Georgia
Tom Cebula, Food and Drug Administration
Mary Clutter, National Science Foundation
Joe DeRisi, University of California, San Francisco
Janet Dorigan, Central Intelligence Agency
Gerald Epstein, Defense Threat Reduction Agency
Stan Falkow, Stanford University
David Franz, Southern Research Institute
Claire Fraser, The Institute for Genomic Research
Art Friedlander, U.S. Army Medical Research Institute of Infectious Diseases
Elizabeth George, Department of Homeland Security
Maria Giovanni. National Institutes of Health
Michael Gottlieb, National Institutes of Health
Robert Heckert, U.S. Department of Agriculture
Maryanna Henkart, National Science Foundation
John Houghton, Department of Energy
Barbara Jasny, *Science*
Norm Kahn, Central Intelligence Agency
Paul Keim, Northern Arizona University

James Kvach, Defense Intelligence Agency
Jim LeDuc, Centers for Disease Control and Prevention
Rachel Levinson, Office of Science and Technology Policy
David Lipman, National Institutes of Health
Vahid Majidi, Department of Justice
Bob May, Royal Society (UK)
Michael Morgan, Wellcome Trust
Tara O'Toole, Johns Hopkins University
George Poste, Arizona State University
Rino Rappuoli, Chiron
David Relman, Stanford University
Caird Rexroad, U.S. Department of Agriculture
Janet Shoemaker, American Society for Microbiology
Terence Taylor, International Institute for Strategic Studies
Ron Walters, Central Intelligence Agency
Marion Warwick, National Science Foundation
Mark Wilson, Federal Bureau of Investigation