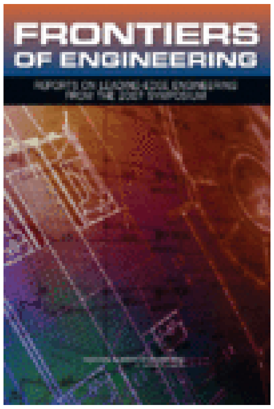


Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2007 Symposium

National Academy of Engineering of the National Academies

ISBN: 0-309-11254-0, 208 pages, 6 x 9, (2008)

FRONTIERS OF ENGINEERING

REPORTS ON LEADING-EDGE ENGINEERING FROM THE 2007 SYMPOSIUM

NATIONAL ACADEMY OF ENGINEERING
OF THE NATIONAL ACADEMIES

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

THE NATIONAL ACADEMIES PRESS • 500 Fifth Street, N.W. • Washington, DC 20001

NOTICE: This publication has been reviewed according to procedures approved by a National Academy of Engineering report review process. Publication of signed work signifies that it is judged a competent and useful contribution worthy of public consideration, but it does not imply endorsement of conclusions or recommendations by the NAE. The interpretations and conclusions in such publications are those of the authors and do not purport to represent the views of the council, officers, or staff of the National Academy of Engineering.

Funding for the activity that led to this publication was provided by the Air Force Office of Scientific Research, Defense Advanced Research Projects Agency, Department of Defense–DDR&E-Research, National Science Foundation, Microsoft Corporation, Cummins, Inc., and Dr. John A. Armstrong.

International Standard Book Number-13: 978-0-309-11253-6
International Standard Book Number-10: 0-309-11253-2

Additional copies of this report are available from The National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20001; (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. Ralph J. Cicerone 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. Charles M. Vest 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. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

ORGANIZING COMMITTEE

JULIA M. PHILLIPS (Chair), Director, Physical, Chemical, and Nano Sciences Center, Sandia National Laboratories
ANA I. ANTÓN, Associate Professor, Department of Computer Science, North Carolina State University
JOHN DUNAGAN, Researcher, Distributed Systems and Security Group, Microsoft Research
RICHARD T. ELANDER, Advanced Pretreatment Team Leader, BioProcess Engineering Group, National Bioenergy Center, National Renewable Energy Laboratory
CHRISTIAN LEBIERE, Research Scientist, Human-Computer Interaction Institute, Carnegie Mellon University
DONALD J. LEO, Associate Dean of Research and Graduate Studies, College of Engineering, Virginia Tech
CAROL R. REGO, Vice President, CDM
VIJAY SINGH, Associate Professor, Department of Agricultural and Biological Engineering, University of Illinois at Urbana-Champaign
PAUL K. WESTERHOFF, Professor, Department of Civil and Environmental Engineering, Arizona State University
ROBERT WRAY, Chief Scientist, Soar Technology

Staff

JANET R. HUNZIKER, Senior Program Officer
VIRGINIA R. BACON, Senior Program Assistant

Preface

This volume highlights the papers presented at the National Academy of Engineering's 2007 U.S. Frontiers of Engineering Symposium. Every year, the symposium brings together 100 outstanding young leaders in engineering to share their cutting-edge research and technical work. The 2007 symposium was held September 24-26, and was hosted by Microsoft Research in Redmond, Washington. Speakers were asked to prepare extended summaries of their presentations, which are reprinted here. The intent of this volume, and of the volumes that preceded it in the series, is to convey the excitement of this unique meeting and to highlight cutting-edge developments in engineering research and technical work.

GOALS OF THE FRONTIERS OF ENGINEERING PROGRAM

The practice of engineering is continually changing. Engineers today must be able not only to thrive in an environment of rapid technological change and globalization, but also to work on interdisciplinary teams. Cutting-edge research is being done at the intersections of engineering disciplines, and successful researchers and practitioners must be aware of developments and challenges in areas other than their own.

At the 2-1/2–day U.S. Frontiers of Engineering Symposium, 100 of this country's best and brightest engineers, ages 30 to 45, have an opportunity to learn from their peers about pioneering work being done in many areas of engineering. The symposium gives engineers from a variety of institutions in academia, industry, and government, and from many different engineering disciplines, an opportunity to make contacts with and learn from individuals whom they would not meet in the usual round of professional meetings. This networking may lead to collaborative work and facilitate the transfer of new techniques and approaches. It is hoped that the exchange of information on current developments in many fields of engineering will lead to insights that may be applicable in specific disciplines.

The number of participants at each meeting is limited to 100 to maximize opportunities for interactions and exchanges among the attendees, who are chosen through a competitive nomination and selection process. The choice of topics and speakers for each meeting is made by an organizing committee composed of engineers in the same 30- to 45-year-old cohort as the participants. Each year different topics are covered, and, with a few exceptions, different individuals participate.

Speakers describe the challenges they face and communicate the excitement of their work to a technically sophisticated, but non-specialized audience. Each speaker provides a brief overview of his/her field of inquiry; defines the frontiers of that field; describes experiments, prototypes, and design studies that have been completed or are in progress, as well as new tools and methodologies, and limitations and controversies; and summarizes the long-term significance of his/her work.

THE 2007 SYMPOSIUM

The five general topics covered at the 2007 meeting were: engineering trustworthy computer systems, control of protein conformations, biotechnology for fuels and chemicals, modeling and simulating human behavior, and safe water technologies. The Engineering Trustworthy Computer Systems session focused on the security challenges of current computing infrastructure. Speakers described new software engineering tools and technologies to improve computer security and the public policy issue of whether owners of devices have the right to tinker with them, which might lead to the discovery of system vulnerabilities and, ultimately, to more trustworthy systems.

Control of protein conformations has been made possible by advances in imaging technologies and techniques using mechanical, optical, and magnetic forces to manipulate proteins directly in order to control protein function. These advances are important because understanding cell signaling and protein-protein interactions is central to understanding the properties of biological systems. The presentations described these techniques and their impact on applications in drug discovery, vaccine development, and new methods for tunable biosensors and bioassays.

There is growing concern about the United States' dependence on imported petroleum for its energy and chemical feedstock supply, particularly in terms of availability and security of future petroleum supplies and impacts of a petrochemical-based economy on the climate. Talks in the session on biotechnology for fuels and chemicals covered the development and commercial deployment of renewable, sustainable, and cost-effective technologies to meet transportation fuel and chemical feedstock needs. Presenters covered applications of corn products in the polymer and pharmaceutical industries, cosmetics, and drug delivery; biochemical processes, operations, and trends in converting biomass to ethanol; and sustainable biorefineries.

The Modeling and Simulating Human Behavior presentations described how advances in functional brain imaging techniques, computational cognitive architectures, and artificial intelligence have led to a better understanding of brain organization and human performance. With the convergence of disciplines in this field, researchers are contributing to the building of an empirical and computational framework for understanding, modeling, and simulating human behavior. Speakers described the current state and future outlook for research in this area and applications for military training and serious games.

The final session was on safe water technologies. In the United States, approximately 40 billion gallons of water are treated to drinking water standards every day. As safe water resources become scarcer globally and in the United States, it has become increasingly important to create inexpensive, high-volume, and dependable water treatment technologies. Three of the talks featured the latest advances in three technologies for producing high-quality water—ultraviolet irradiation, membrane processes, and biological water treatment. In the final talk, the speaker described challenges in managing the water distribution system, the final frontier in providing safe water to consumers.

In addition to these plenary sessions, there were many opportunities for more informal interactions, including breakout sessions on the first afternoon where participants discussed such issues as, What does sustainability mean for the engineering community? How does one balance family life with a demanding engineering career? What is the role of the social sciences in engineering? How does one balance rigor and creativity in information technology, science, engineering, and design? How can we better prepare Ph.D. engineers for the competitive global marketplace? What are the responsibilities of engineers as members of society, and how are those responsibilities carried out? A summary of the discussions from the breakout groups is included in the Appendixes. On the second afternoon, Microsoft Research engineers and scientists set up a number of displays and demos about research they are doing in surface computing, web security, tablet PCs, and visualization of data in graphs and charts. Participants circulated among the displays and asked questions about the various technologies.

Every year, a distinguished engineer addresses the participants at dinner on the first evening of the symposium. The speaker this year, Henrique (Rico) Malvar, managing director of Microsoft Research, gave the talk. His description of Microsoft Research's operational underpinnings—an open academic model, long-term focus, informal atmosphere of exchange between researchers and product-development groups, and recruitment of world-class researchers with diverse backgrounds—provided valuable insight into how one corporate research entity organizes itself to achieve its goals.

NAE is deeply grateful to the following organizations for their support of the 2007 U.S. Frontiers of Engineering Symposium: Microsoft Research, Air Force Office of Scientific Research, Defense Advanced Research Projects Agency, Department of Defense–DDR&E-Research, National Science Foundation, Cummins Inc., and Dr. John A. Armstrong. NAE would also like to thank the members of the Symposium Organizing Committee (p. iv), chaired by Dr. Julia M. Phillips, for planning and organizing the event.

Contents

ENGINEERING TRUSTWORTHY COMPUTER SYSTEMS

CONTROL OF PROTEIN CONFORMATIONS

BIOTECHNOLOGY FOR FUELS AND CHEMICALS

MODELING AND SIMULATING HUMAN BEHAVIOR

SAFE WATER TECHNOLOGIES

APPENDIXES

Engineering Trustworthy Computer Systems

Introduction

Ana I. Antón
North Carolina State University
Raleigh, North Carolina

John Dunagan
Microsoft Research
Redmond, Washington

Improving the trustworthiness of computer systems is a very broad research challenge in computer science. The negative consequences of our current computing infrastructure are sufficiently severe that they are discussed in the popular press, in academic forums, and in the legislative arena. While identity theft is perhaps the most widespread negative consequence today, we also face threats ranging from vote fraud to failure to protect consumer privacy.

Two of the presentations in this session cover two public concerns that remain mostly unaddressed in practice: vote fraud and consumer privacy. The other two presentations in this session cover infrastructural approaches to improving computer trustworthiness: software engineering tools and new security technologies. In total, we hope these convey the breadth of effort being expended in the research community to develop computing technologies that merit the trust society wants to place in them.

Privacy in a Networked World

Rebecca N. Wright
Rutgers University
Piscataway, New Jersey

ABSTRACT

Networked electronic devices have permeated business, government, recreation, and almost all aspects of daily life. Coupled with the decreased cost of data storage and processing, this has led to a proliferation of data about people, organizations, and their activities. This cheap and easy access to information has enabled a wide variety of services, efficient business, and convenience enjoyed by many. However, it has also resulted in privacy concerns. As many recent incidents have shown, people can be fooled into providing sensitive data to identity thieves, stored data can be lost or stolen, and even anonymized data can frequently be reidentified. In this paper we discuss privacy challenges, existing technological solutions, and promising directions for the future.

Changes in technology are causing an erosion of privacy. Historically, people lived in smaller communities, and there was little movement of people from one community to another. People had very little privacy, but social mechanisms helped prevent abuse of information. As transportation and communications technologies developed, people began to live in larger cities and to have increased movement between communities. Many of the social mechanisms of smaller communities were lost, but privacy was gained through anonymity and scale.

Now, advances in computing and communications technology are reducing privacy by making it possible for people and organizations to store and process personal information, but social mechanisms to prevent the misuse of such information have not been replaced. While a major issue in computing and communications technology used to be how to make information public, we now have to work hard to keep it private.

The issue of confidentiality,[1] or protecting information in transit or in storage from an unauthorized reader, is a well-understood problem in computer science. That is, how does a sender Alice send a message *M* to an intended receiver Bob in such a way that Bob learns *M* but an eavesdropper Eve does not, perhaps even if Eve is an active attacker who has some control over the communication network? Although some difficulties remain in practical key management and end-host protection, for the most part this problem is quite well solved by the use of encryption.

In contrast, a modern view of privacy is not simply about Eve not learning the message *M* or anything about its contents, but includes other issues such as Eve learning whether a message was sent between Alice and Bob at all, and questions about what Bob will and will not do with *M* after he learns it. This view of privacy in the electronic setting was first introduced in the early 1980s by David Chaum (1981, 1985, 1992). A growing interest in privacy in the computer science research community can be seen by the large number of annual conferences, workshops, and journals now devoted to the topic.[2]

It is easy to describe how to achieve privacy when it is considered in isolation: one can live in seclusion, avoid any use of computers and telephones, pay cash for all purchases, and travel on foot. Obviously, this is not realistic or even desirable for the vast majority of people. The difficulty of privacy arises because of the apparent conflict between utility and privacy, that is, the desire of various parties to benefit from the convenience and other advantages provided by use of information, while allowing people to retain some control over "their" information, or at the very least to be aware of what is happening to their information.

The problem is further complicated by the fact that privacy means different things to different people. What some people consider a privacy violation is considered completely innocuous by other people. Or what seems a privacy violation in one context may not seem to be one in another context. For this reason privacy is not a purely technological issue and cannot have purely technological solutions. Social, philosophical, legal, and public policy issues are also important. However,

[1] One should note that the terminology distinction of "confidentiality" and "privacy" is not thoroughly standardized, and some people use the term "confidentiality" in the way the term "privacy" is used in this paper.

[2] For example, ACM Conference on Computers, Freedom, and Privacy (yearly since 1991); ACM Workshop on Privacy in the Electronic Society (yearly since 2002); Carnegie Mellon University Symposium on Usable Privacy and Security (yearly since 2006); and Workshop on Privacy-Enhancing Technologies (yearly since 2002).

technology can enable new policy decisions to be possible by instantiating solutions with particular properties.

Given the pervasive nature of networked computing, privacy issues arise in a large number of settings, including the following:

- **Electronic health records.** In the United States there is a large effort to move toward electronic health records in order to improve medical outcomes as well as reduce the exorbitant cost of health care. Unless solutions can be developed that allow medical practitioners' access to the right personal health information at the right time and in the right circumstances, while also ensuring that it cannot be accessed otherwise or used inappropriately even by those who have legitimate access, privacy will remain a barrier to adoption.
- **Government surveillance.** In the interest of protecting national security and preventing crime, governments often wish to engage in surveillance activities and link together huge amounts of data in order to identify potential threats before they occur and learn more about incidents and their perpetrators if they occur. Most people want increased security, but many remain concerned about invasion of privacy.
- **Commerce and finance.** Consumers have eagerly adopted online commerce and banking because of the great convenience of being able to carry out transactions from home or anywhere else. However, due to the use of personal information such as social security numbers and mother's maiden name being used for authentication coupled with the ease with which this kind of personal data can be learned, this has resulted in a huge increase in identity theft. It also allows companies (particularly when data from multiple sources is aggregated by data aggregators) to gain great insight into customers and potential customers, often to the point where customers feel unsettled.
- **Pervasive wireless sensors.** Sensors such as RFID tags (inexpensive tiny chips that broadcast a unique 96-bit serial number when queried and are used in mobile payment applications such as EZPass and ExxonMobil's Speedpass, as well as increasingly being embedded in consumer products—or even in consumers themselves in some cases), mobile telephones and other personal devices that broadcast recognizable identification information, and GPS transmitters such as those used in popular car navigation systems. These devices can potentially be used to track individuals' locations and interactions.

There are a large number of different kinds of solutions to various privacy problems, with different kinds of properties. Some aspects in which solutions differ are: transparent (users cannot even tell they are there, but they protect privacy in some way anyway) or visible (users can easily tell they are there and can, or even must, interact with the system while carrying out their tasks); individual user (in which an individual can unilaterally make a choice to obtain more privacy, say, by using a particular software package or configuring it in a certain way) or infrastructure (in which some shared infrastructure, such as the Internet itself, is modified or redesigned in order to provide more privacy); focused only on notification (allowing or requiring entities to describe their data practices) or also on compliance (ensuring that stated practices are observed); and in other ways. However, unless a solution primarily favors utility and functionality over privacy when choices must be made, it will tend not to be widely adopted. I describe one kind of solution, privacy-preserving data mining, in further detail in the following section.

PRIVACY-PRESERVING DATA MINING

One area of investigation in privacy-related technologies is the use of distributed algorithms to process data held by one or more parties without requiring the parties to reveal their data to each other or any other parties. While this may seem contradictory at first, sophisticated use of cryptography can yield solutions with unintuitive properties. Specifically, the elegant and powerful paradigm of general secure multiparty computation (Goldreich et al., 1987; Yao, 1986) shows how cryptography can be used to allow multiple parties, each holding a private input, to engage in a computation on their collective inputs in such a way that they all learn the result of the computation but nothing else about each other's data. This is achievable with computation and communication overhead that is reasonable when described as a function of the size of the private inputs and the complexity of the nonprivate computation.

One area in which one might seek to use these kinds of algorithms is in the area of data mining. However, because the inputs to data-mining algorithms are typically huge, the overheads of the general secure multiparty computation solutions are intolerable for most data-mining applications. Instead, research in this area seeks more efficient solutions for specific computations. Most cryptographic privacy-preserving data-mining solutions to date address typical data-mining algorithms, such as clustering (Vaidya and Clifton, 2003; Jagannathan and Wright, 2005), decision trees (Lindell and Pinkas, 2002), or Bayesian networks (Meng et al., 2004; Yang and Wright, 2006). Recent work addresses privacy preservation during the preprocessing step (Jagannathan and Wright, 2006) and the postprocessing step (Yang et al., 2007), thereby working toward maintaining privacy throughout the data-mining process.

Cryptographic techniques provide the tools to protect data privacy by allowing exactly the desired information to be shared while concealing everything else about the data. To illustrate how to use cryptographic techniques to design privacy-preserving solutions to enable mining across distributed parties, we describe a privacy-preserving solution for a particular data-mining task: learning Bayesian networks on a dataset divided among two parties who want to carry out data-mining algorithms on their joint data without sharing their data directly. Full details of this solution are described in Yang and Wright (2006).

Bayesian Networks

A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest (Cooper and Herskovits, 1992). This model can be used for data analysis and is widely used in data-mining applications.

Formally, a Bayesian network for a set *V* of *m* variables consists of two parts, called the network structure and the parameters. The network structure is a directed acyclic graph whose nodes are the set of variables. The parameters describe local probability distributions associated with each variable. There are two important issues in using Bayesian networks: (1) learning Bayesian networks and (2) Bayesian inference. Learning Bayesian networks includes learning the structure and the corresponding parameters. Bayesian networks can be constructed using expert knowledge or derived from a set of data, or by combining those two methods together. Here we address the problem of privacy-preserving learning of Bayesian networks from a database vertically partitioned between two parties. In vertically partitioned data one party holds some of the variables and the other party holds the remaining variables.

Privacy-Preserving Bayesian Network Learning Algorithm

A value *x* is said to be secret shared (or simply shared) between two parties if the parties have values (shares) such that neither party knows (anything about) *x*, but given both parties' shares of *x*, it is easy to compute *x*. Our distributed privacy-preserving algorithm for Bayesian network learning uses composition of privacy-preserving components in which all intermediate outputs from one component that are inputs to the next component are computed as secret shares. In this way it can be shown that if each component is privacy preserving, then the resulting composition is also privacy preserving.

Our solution is a modified version of the well-known K2 algorithm of Cooper and Herskovits (1992). That algorithm uses a score function to determine which edges to add to the network. To modify the algorithm to be distributed and privacy preserving, we seek to divide the problem into several smaller subproblems that we know how to solve in a distributed, privacy-preserving way. Specifically, noting that only the relative score values are important, we use a new score function *g* that approximates the relative order of the original score function. This is obtained by taking the logarithm of the original score function and dropping some lower order terms.

As a result, we are able to perform the necessary computations in a privacy-preserving way. We make use of several existing cryptographic components, including secure two-party computation (such as the solution of Yao [1986], which we apply only on a small number of values, not on something the size of the original database), privacy-preserving scalar product share algorithm (such as the solutions described by Goethals et al., 2004), and a privacy-preserving algorithm for computing $x\ln x$ (such as Lindell and Pinkas, 2002). In turn, these can be combined to create a privacy-preserving approximate score algorithm and a privacy-preserving score comparison algorithm, as described in Yang and Wright (2006).

The overall distributed algorithm for learning Bayesian networks is as follows. In keeping with cryptographic tradition, we call the two parties engaged in the algorithm Alice and Bob.

Input: An ordered set of m nodes, an upper bound u on the number of parents for a node, both known to Alice and Bob, and a database containing n records, vertically partitioned between Alice and Bob.

Output: Bayesian network structure (whose nodes are the m input nodes, and whose edges are as defined by the values of π_i at the end of the algorithm).

Following the specified ordering of the m nodes, Alice and Bob execute the following steps at each node v_i in turn. Initially each node has no parent; parents are added as the algorithm progresses. The current set of parents of v_i is denoted by π_i.

1. Alice and Bob execute the privacy-preserving approximate score algorithm to compute the secret shares of $g(i,\pi_i)$ and $\mathrm{g}(i,\pi_i \cup \{z\})$ for any possible additional parent z of v_i.

2. Alice and Bob execute the privacy-preserving score comparison algorithm to compute which of those scores in Step 1 is maximum.

3. If $g(i,\pi_i)$ is maximum (i.e., no new parent increases the score) or if v_i already has the maximum number of parents, then Alice and Bob go to the next node to run from Step 1 until Step 3. Otherwise, if $\mathrm{g}(i,\pi_i \cup \{z\})$ is maximum (i.e., potential parent z increases the score the most), then z is added as the parent of v_i by setting $\pi_i = \pi_i \cup \{z\}$ and Alice and Bob go back to Step 1 on the same node v_i.

Further details about this algorithm can be found in Yang and Wright (2006), where we also show a privacy-preserving algorithm to compute the Bayesian network parameters. Experimental results addressing both the efficiency and the accuracy of the structure-learning algorithm can be found in Kardes et al. (2005).

CHALLENGES FOR THE FUTURE

Many challenges, both technical and political, remain regarding privacy. These include social and political questions regarding who should have the right or responsibility to make various privacy-related decisions about data pertaining to an individual, as well as continued development and deployment of technologies to enable these rights and ensure that such privacy decisions, once made by the appropriate parties, are respected.

REFERENCES

Chaum, D. 1981. Untraceable electronic mail, return addresses, and digital pseudonyms. Communications of the ACM 24:84-88.

Chaum, D. 1985. Security without identification: Transaction systems to make big brother obsolete. Communications of the ACM 28:1030-1044.

Chaum, D. 1992. Achieving electronic privacy. Scientific American 267(2):96-101.

Cooper, G. F., and E. Herskovits. 1992. A Bayesian method for the induction of probabilistic networks from data. Machine Learning 9(4):309-347.

Goethals, B., S. Laur, H. Lipmaa, and T. Mielikäinen. 2004. On private scalar product computation for privacy-preserving data mining. Pp. 104-120 in Proceedings of the Seventh Annual International Conference in Information Security and Cryptology, volume 3506 of LNCS. Berlin, Germany: Springer-Verlag.

Goldreich, O., S. Micali, and A. Wigderson. 1987. How to play ANY mental game. Pp. 218-229 in Proceedings of the 19th Annual ACM Conference on Theory of Computing. New York: ACM Press.

Jagannathan, G., and R. N. Wright. 2005. Privacy-preserving distributed k-means clustering over arbitrarily partitioned data. Pp. 593-599 in Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press.

Jagannathan, G., and R. N. Wright. 2006. Privacy-preserving data imputation. Pp. 535-540 in Proceedings of the ICDM International Workshop on Privacy Aspects of Data Mining. New York: ACM Press.

Kardes, O., R. S. Ryger, R. N. Wright, and J. Feigenbaum. 2005. Implementing privacy-preserving Bayesian-net discovery for vertically partitioned data. Pp. 26-34 in Proceedings of the ICDM Workshop on Privacy and Security Aspects of Data Mining. New York: ACM Press.

Lindell, Y., and B. Pinkas. 2002. Privacy preserving data mining. Journal of Cryptology 15(3): 177-206.

Meng, D., K. Sivakumar, and H. Kargupta. 2004. Privacy-sensitive Bayesian network parameter learning. Pp. 487-490 in Proceedings of the Fourth IEEE International Conference on Data Mining. Piscataway, NJ: IEEE.

Vaidya, J., and C. Clifton. 2003. Privacy-preserving k-means clustering over vertically partitioned data. Pp. 206-215 in Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press.

Yang, Z., and R. Wright. 2006. Privacy-preserving computation of Bayesian networks on vertically partitioned data. IEEE Transactions on Data Knowledge Engineering 18:1253-1264.

Yang, Z., S. Zhong, and R. N. Wright. 2007. Towards privacy-preserving model selection. Pp. 11-17 in Proceedings of the First ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD. New York: ACM Press.

Yao, A. C.-C. 1986. How to generate and exchange secrets. Pp. 162-167 in Proceedings of the 27th IEEE Symposium on Foundations of Computer Science. Piscataway, NJ: IEEE.

Unifying Disparate Tools in Software Security

GREG MORRISETT
Harvard University
Cambridge, Massachusetts

How much can you trust software? When you install a piece of code, such as a video game or a device driver for a new camera, how can you be sure the code won't delete all of your files or install a key-stroke logger that captures your passwords? How can you ensure that the software doesn't contain coding bugs or logic errors that might leave a security hole?

Traditional approaches to software security have assumed that users could easily determine when they were installing code and whether or not software was trustworthy in a particular context. This assumption was reasonable when computers controlled few things of real value, when only a small number of people (typically experts) installed software, and when the software itself was relatively small and simple. But the security landscape has changed drastically with advances in technology (e.g., the explosive growth of the Internet, the increasing size and complexity of software) and new business practices (e.g., outsourcing and the development of open-source architecture). Furthermore, the more we rely on software to control critical systems, from phones and airplanes to banks and armies, the more desperately we need mechanisms to ensure that software is truly trustworthy.

ATTACK TECHNIQUES

Malicious hackers have serious financial incentives to break into and control personal and business computers. Once they break into a machine, they can gather personal information (e.g., passwords, credit card information, and social security numbers) and use it to clean out bank accounts, apply for credit cards, or gain access to other machines on the same network. In addition, if attackers break into and control machines, they can use them to send "spam" (unsolicited e-mail advertisements), to store illicit digital goods (e.g., pirated music or pornography), or to launch denial-of-service attacks against other computing systems.

One technique attackers use to gain a foothold on a machine is to trick the user into thinking a piece of code is trustworthy. For example, an attacker may send an e-mail that contains a malicious program, but with a forged return address from someone the user trusts. Alternatively, an attacker may set up a website with a name that appears to be trustworthy, such as "*www.micr0s0ft.com*" (note that the letter "o" has been replaced by the numeral "0"), and send an e-mail with a link to the site suggesting that the user download and install a security update.

Even sophisticated users can be fooled. A clever attacker might contribute "Trojan horse" code to a popular open-source project, turning an otherwise useful program (e.g., a music player) into undetected mal-ware.[1] In short, it is relatively easy for an attacker to gain the trust of a user and get him or her to install code.

Today we use a combination of digital signatures and virus scanners to try to stop these attacks. But digital signatures only tell us who produced the code, not whether it is trustworthy. And virus scanners, which look for snippets of previously identified mal-ware at the syntactic level, can be easily defeated through automated obfuscation techniques. Furthermore, virus scanners cannot detect new attacks. These relatively weak tools can tell us something about the *provenance* of software and its *syntax*, but they cannot validate the software's *semantics*.

EVEN GOOD GUYS CAN'T BE TRUSTED

A second technique attackers use to gain control of a machine is to find a bug in an already installed program that communicates with the outside world. The classic example is a *buffer overrun* in a service, such as a log-in daemon, or a network-based application, such as a web browser. A buffer overrun occurs when a program allocates *n* bytes of memory to hold an input (e.g., a password), but the attacker provides more than *n* bytes. A properly written program will check and reject any input that is too long. Unfortunately, many programmers fail to insert appropriate checks, partly because commonly used programming languages (C and C++) make it easy to forget those checks, and partly because programmers often think that "no one will have a password of more than a thousand characters."

When programmers fail to put in a check and the input is too long, the extra bytes overwrite whatever data happen to be stored next to the input buffer. In many cases, the buffer is allocated on the control stack near a return address

[1] See Thompson (1984) for a particularly devious example.

for a procedure—a location where the program transfers control once the input routine is completed.

A clever attacker will enter an input long enough to overwrite the return address with a new value, thereby controlling which code is executed when the input routine ends. In an ideal attack, the rest of the input contains executable instructions, and the attacker causes control to transfer to this newly injected code. In this way, the attacker can cause the program to execute arbitrary code.[2]

Attacks based on buffer overruns are surprisingly common; at one point, they accounted for more than half of the security vulnerabilities reported by the Computer Emergency Response Team (Wagner et al., 2000). But the failure to check input lengths is only one of a large number of coding errors attackers use to gain a foothold on a machine or extract private information. Other examples include integer overflows; format-string attacks; script-injection attacks; race conditions; bad, pseudo-random number generators; and improper use of cryptographic primitives.

Thus, even well-meaning software vendors cannot seem to produce trustworthy code. Part of the problem is that vendors must keep producing new features to sell new versions of software, which simply adds complexity to the code base. And part of the problem is that current development practices, including design, coding, review, and testing, cannot rule out simple errors, such as buffer overruns, much less deeper problems, such as the inappropriate use of cryptographic primitives or covert-information channels.

PREVENTING BUGS THROUGH REWRITING

One way to defend against attacks is to build tools that automatically insert checks at program locations where a policy violation might occur. For example, we might insert extra code at each buffer update to ensure that we do not write beyond its bounds.

Compilers for high-level, type-safe languages, such as Java and C#, already insert checks to prevent a large category of language-level errors, including buffer overruns. However, the vast majority of systems and application software, including security-critical operating-systems code, is still written in C or C++, so the compiler does not have enough information to insert appropriate checks.

Unfortunately, the cost of rewriting the millions of lines of code that make up big software systems (e.g., Windows or Oracle) is prohibitive, and doing so is likely to introduce as many bugs as it eliminates. Furthermore, because certain services, such as device drivers and real-time embedded software, need control over memory layout and timing, they are not suited to high-level languages.

Some effective tools for rewriting low-level legacy code are emerging. For example, the StackGuard tool (Cowan et al., 1998) and the Microsoft "/gs"

[2] For a more detailed explanation, see <http://www.phrack.org/archives/49/P49-14/>.

compiler option rewrite C/C++ code to insert a secret "cookie" next to buffers allocated on the control-flow stack and check that the cookie has been preserved when a procedure is about to jump to a return address. In this way a buffer overrun can often be detected and stopped with relatively low overhead. Unfortunately, this technique does not protect against other forms of attack, such as overflowing a buffer allocated on the heap.

Another example of automated rewriting is the CCured compiler (Necula et al., 2002). CCured provides a strong type-safety guarantee by inserting extra metadata and run-time checks into C code. Metadata make it possible to determine, for instance, the size of a buffer at run time, and the checks ensure that *all* buffer boundaries and typing constraints are satisfied.

Although CCured can entail higher overhead costs than StackGuard, the security guarantees are much stronger. Nevertheless, the advantage of StackGuard is that it can be applied to almost any C program without change, whereas CCured requires a number of changes to the source code. In practice, the tool that is easier to use is almost always used first.

Many other tools also try to enforce security policies on legacy code through rewriting. These include software-based fault isolation (McCamant and Morrisett, 2006; Wahbe et al., 1993), control-flow isolation (Abadi et al., 2005), in-lined reference monitors (Schneider, 2000), and others. Each of these tools strikes a different balance among the expressiveness of the policies that can be enforced, the code base to which the techniques are applicable, and run-time overhead.

PREVENTING BUGS THROUGH STATIC ANALYSIS

Security-conscious companies are beginning to use tools that either prevent or detect a wide range of coding errors at compile time. Microsoft, for example, uses a static analysis tool called Prefast (based on an early tool called Prefix) that scans C and C++ code, looking for common errors, such as buffer overruns (Bush et al., 2000). Other companies, such as Fortify and Coverity, produce similar tools that can analyze software written in a variety of languages and support customizable rule-sets for finding new classes of bugs as they arise.

A static analysis tool attempts to symbolically execute all paths in the code, looking for potential bugs. Of course, a program with n conditional statements can have 2^n distinct paths, and a program with loops or recursion can have an unbounded number of paths. Thus the static analysis approach is both naive and infeasible. Furthermore, the analysis does not know the actual values of inputs to the program, so it cannot always determine the value of variables or data structures.

Consequently, static analysis tools construct models of the program that abstract details and reduce the reasoning to finite domains. For example, instead of tracking the actual values that integer variables might take on, an analysis might track only upper and lower bounds. To ensure termination, the analysis might only consider a few iterations of a loop. More sophisticated techniques based on abstract interpretation (Cousot and Cousot, 1977) make it possible to determine an approximation of behavior for all iterations.

Although automated static analysis has improved tremendously, especially in the past few years, a number of problems remain to be solved. One problem, introduced by the necessary approximation described above, is false positives (i.e., potential errors may be indicated where there are none) or false negatives (i.e., errors that are not reported).

If the approximations are constructed so that the analysis is *sound* (i.e., no false negatives), programmers tend to be flooded with false positives, making it difficult to find and fix real bugs in the program. Therefore, few of these analyses are sound, and bugs can potentially sneak through. For example, a buffer overrun recently found in the Windows Vista custom-cursor animation code was not detected by Prefast.

LOOKING TO THE FUTURE

One problem with static analysis tools is that, like optimizing compilers, they tend to be large, complicated programs. As a result, a bug in the analysis can mean that a security-relevant error goes undetected. Furthermore, because most tools operate on source code (or byte code), they must make assumptions about how the compiler will translate the program to machine code. Consequently, an inconsistent assumption or a bug in the compiler can result in a security hole.

An ideal static analysis would fit the following description:

- The analysis can operate at the machine-code level (to avoid having to reason about the compiler).
- The analysis is small, simple, and formally verified.
- The analysis is sound (i.e., no false positives).
- The analysis is complete (i.e., no false negatives).

Although these goals may seem unattainable, the *proof-carrying code* (PCC) architecture suggested by Necula (1997) comes remarkably close to satisfying them. The idea behind PCC is to require that executable programs come equipped with a formal, machine-checkable "proof" that the code does not contain bugs. By "proof" we mean a formal, logical argument that the code, when executed, will not violate a pre-specified (and formalized) policy on behavior.

The key insight of PCC is that constructing a proof is hard (generally undecidable), but verifying a proof is easy. After all, a proof is meant to be mechanically checkable evidence. Furthermore, assuming we have a sound, *relatively* complete logic for reasoning about program behavior, we can, in principle, accept precisely those programs that are *provably* safe.

The really nice thing about PCC is that, as a framework, it benefits both the user who is worried about the safety of installing a program downloaded from the Internet and the software producer who has outsourced the development of modules. With PCC, we are no longer dependent on *who* produces the code but on what the code actually does.

In practice, the problem with PCC is that someone, namely the code producer, still has to come up with proof that the code satisfies a given policy. Thus none of the difficult problems has really been eliminated. We still need other techniques, such as static analysis and rewriting, to find an effective way to construct the proof. However, with PCC we no longer have to trust the tools or compiler. Therefore PCC helps minimize the size of the trusted computing base.

One way to construct a proof that machine code satisfies a policy is by using a *proof-preserving compiler*. For example, if we start with source code and a proof that the source satisfies a policy, then a proof-preserving compiler can simultaneously transform the code and proof to the machine-code level. *Type-preserving compilers* are a particular instance in which the policy is restricted to a form of type safety (Colby et al., 2000; Morrisett et al., 1999; Necula and Lee, 1998).

Next-generation programming languages are incorporating increasingly sophisticated type systems that allow programmers to provide stronger safety and security properties through automated type checking. For example, the Jif dialect of Java lets programmers specify secrecy requirements for data, and the type system ensures that secret inputs (i.e., private fields) cannot flow to public outputs (Myers, 1999). Thus, when combined with a type-preserving compiler, the Jif-type system makes it possible for a third party to verify that a program will not disclose information that is intended to be private.

At an extreme, arbitrary policies can be captured with a *dependent type-system*, as found in emerging languages, such as ATS (Xi, 2004), Concoqtion (Fogarty et al., 2007), Epigram (McBride, 2004), and HTT (Nanevski et al., 2006). These languages make it possible to specify everything from simple typing properties to full correctness. The trade-off, of course, is that type checking is only semiautomatic. Simple properties can be discharged using a combination of static analysis, constraint solving, and automated theorem proving, but ultimately, programmers must construct explicit proofs for deeper properties.

Thus the success of these languages will be determined largely by how much can truly be automated. Nevertheless, for safety and security-critical software systems, these languages, in conjunction with PPCs and the proof-carrying code architecture, provide a compelling research vision.

REFERENCES

Abadi, M., M. Budiu, U. Erlingsson, and J. Ligatti. 2005. Control-flow integrity: Principles, implementations, and applications. Pp. 340-353 in Proceedings of the 12th ACM Conference on Computer and Communications Security (CCS'05). New York: ACM Press.

Bush, W., J. Pincus, and D. Sielaff. 2000. A static analyzer for finding dynamic programming errors. Software—Practice and Experience 30(7):775-802.

Colby, C., P. Lee, G. C. Necula, F. Blau, K. Cline, and M. Plesko. 2000. A certifying compiler for Java. Pp. 95-107 in Proceedings of the 2000 ACM SIGPLAN Conference on Programming Language Design and Implementation. New York: ACM Press.

Cousot, P., and R. Cousot. 1977. Abstract interpretation: A unified lattice model for static analysis of programs by construction or approximation of fixpoints. Pp. 238-252 in Proceedings of the Symposium on Principles of Programming Languages. New York: ACM Press.

Cowan, C., C. Pu, D. Maier, H. Hinton, J. Walpole, P. Bakke, S. Beattie, A. Grier, P. Wagle, and Q. Zhang. 1998. StackGuard: Automatic adaptive detection and prevention of buffer-overflow attacks. Pp. 63-78 in Proceedings of the 7th USENIX Security Symposium. Berkeley, CA: USENIX Association.

Fogarty, S., E. Pasalic, J. Siek, and W. Taha. 2007. Concoqtion: Indexed types now! Pp. 112-121 in Proceedings of the 2007 ACM SIGPLAN Workshop on Partial Evaluation and Semantics-Based Program Manipulation. New York: ACM Press.

McBride, C. 2004. Practical programming with dependent types. Pp. 130-170 in Lecture Notes in Computer Science 3622, V. Vene and T. Uustalu, eds. New York: Springer-Verlag.

McCamant, S., and G. Morrisett. 2006. Evaluating SFI for a CISC architecture. Pp. 209-224 in Proceedings of the 15th USENIX Security Symposium. Berkeley, CA: USENIX Association.

Morrisett, G., D. Walker, K. Crary, and N. Glew. 1999. From system F to typed assembly language. ACM Transactions on Programming Languages and Systems 21(3):527-568.

Myers, A. C. 1999. Practical mostly-static information flow control. Pp. 228-241 in Proceedings of the ACM Symposium on Principles of Programming Languages. New York: ACM Press.

Nanevski, A., G. Morrisett, and L. Birkedal. 2006. Polymorphism and separation in hoare type theory. Pp. 62-73 in Proceedings of the ACM International Conference on Functional Programming. New York: ACM Press.

Necula, G. C. 1997. Proof-carrying code. Pp. 106-119 in Proceedings of the ACM International Symposium on Principles of Programming Languages. New York: ACM Press.

Necula, G. C., and P. Lee. 1998. The design and implementation of a certifying compiler. Pp. 333-344 in Proceedings of the ACM Conference on Programming Language Design and Implementation. New York: ACM Press.

Necula, G. C., S. McPeak, and W. Weimer. 2002. CCured: Type-safe retrofitting of legacy code. Pp. 128-139 in Proceedings of the ACM Symposium on Principles of Programming Languages. New York: ACM Press.

Schneider, F. B. 2000. Enforceable security policies. ACM Transactions on Information Systems Security 3(1):30-50.

Thompson, K. 1984. Reflections on trusting trust. Communications of the ACM 27(8):761-763.

Wagner, D., J. S. Foster, E. A. Brewer, and A. Aiken. 2000. A first step towards automated detection of buffer overrun vulnerabilities. Pp. 3-17 in Network and Distributed System Security Symposium. Reston, VA: Internet Society.

Wahbe, R., S. Lucco, T. E. Anderson, and S. L. Graham. 1993. Software-based fault isolation. ACM SIGOPS Operating Systems Review 27(5):203-216.

Xi, H. 2004. Applied type systems (extended abstract). Pp. 394-408 in Lecture Notes in Computer Science 3085. New York: Springer-Verlag.

Usable Security: Oxymoron or Challenge?

Diana K. Smetters
PARC
Palo Alto, California

Security is rapidly emerging as one of the greatest challenges of our modern computer-centric society. One of the least addressed factors crucial to achieving effective computer security is usability. Too often users and their pesky focus on the tasks they are actually trying to accomplish are considered primarily as an impediment to systems security, rather than the reason for building those systems in the first place.

Over the last several years there has been a rapid expansion of research into making systems that are both usable and secure. Beginning with studies that simply characterize the overwhelming flaws of current technologies designed without any thought to the user, it has expanded into interest from the Human-Computer Interaction (HCI) community in designing improved user interfaces to existing security technologies, in the hopes of increasing their usability. While both of these approaches play important roles, we have argued (Smetters and Grinter, 2002; Balfanz et al., 2004a) that only through designing (or redesigning) security technologies and secure systems from the ground up with usability in mind will there be systems sufficiently usable and secure to meet the demands of modern computing environments.

In this paper I will briefly review recent work both on improving the usability of security technologies and in designing systems to be simultaneously usable and secure, with an eye toward the challenges still faced in marrying these two seemingly opposing goals.

PASSWORDS: THE OLD STANDBY

There is no more common security technology in use today than that of the ubiquitous user name and password. Easy to understand and implement, passwords play a critical functional role in online society as the secrets that bind real people to their digital information and personas. They require no more of their users than to simply remember them. Unfortunately, they are flawed in both nature and execution—as the simplest form of secret, they are easily given away and reused[1] by attackers. As such, they and other similar forms of "secret" information such as social security numbers are subject to increasingly sophisticated "phishing" attacks, which attempt to trick users into revealing them. They are also deployed in a fashion that overtaxes users' abilities to manage them securely and effectively. A number of studies have demonstrated clearly what most password users (i.e., all of us) know by experience: people don't know how to pick good passwords and are asked to remember far too many of them or change them far too often, resulting in poor password choice and passwords being written down or shared (Adams and Sasse, 1999).

Standard password policies are striking in how effectively they minimize usability. Even simple changes in typical password policies can demonstrably increase usability without decreasing security, for example, not requiring password change or increasing the number of password input errors allowed without requiring administrator intervention (e.g., from 3 to 10) (Brostoff and Sasse, 2003).

Unfortunately, the increasing body of research on both traditional text passwords and various forms of new graphical passwords appears to be moving inexorably to the conclusion that passwords in any form cannot be used securely in their "naked" form (i.e., based only on what a user is able to remember on his own) without technological support. Most interestingly, it seems the very universality of passwords is also their downfall. For example, studies suggested that a simple technique for constructing passwords through mnemonic phrases could result in passwords that were as difficult to guess as randomly generated passwords, while at the same time being easy to remember (Yan et al., 2005). Unfortunately, it turns out that in practice most of us pick the same mnemonic phrases from which to generate our passwords, making these hard passwords vulnerable to simple dictionary attacks (Kuo et al., 2006).

Because of their strong appeal, more design activity is going into attempts to improve the security and usability of passwords—with mixed results. Mutual authentication systems, such as SiteKey™, attempt to reduce the risk that users will reveal passwords to other than the websites they intend by providing a supposedly user-friendly image-based method for users to authenticate the website

[1] In contrast to more effective approaches utilizing cryptography, which allows secrets, such as keys or passwords, to be used without being revealed in a form that allows an attacker to intercept and reuse them.

they are communicating with prior to entering their password. Unfortunately, recent research shows that users do not notice the absence of the correct SiteKey™ security indicators, still entering their passwords into possibly malicious websites (Schechter at al., 2007). Two-factor authentication systems require users to present something other than just a password for access, and attempt to ensure that the other factor is one that is difficult for an attacker to steal. However, perhaps in order to minimize changes required to infrastructure, such systems have been deployed most commonly in a form that gives users one-time passwords to be entered in addition to their standard credentials, either from a list or generated automatically by an electronic token, rather than giving them cryptographic keys to use to authenticate themselves. Unfortunately, these one-time password systems have shown themselves vulnerable in practice to attackers who interpose themselves between the user and the resource as a so-called "man in the middle," capturing the password from the user and then handing it to the resource provider to gain access, while returning an innocuous-looking error message to the original user (e.g., "You typed your password incorrectly; please try again.").

Password managers range from simple software assistants that help users keep straight their bewildering number of passwords, to complex pieces of software that provide a number of defenses. The best of these, PassPet (Yee and Sitaker, 2006) and Web Wallet (Wu et al., 2006), act to significantly strengthen passwords as a security measure by asking users to remember a single master password, from which they generate unique passwords for every site a user logs into. By enabling users to have distinct passwords for each site they access, these tools can not only minimize the damage caused by a single stolen password (which is usually significant, as passwords are frequently reused in practice), but can also prevent a user from entering a particular password into the wrong site (e.g., a phishing site). While promising, the most widely deployed of these tools (e.g., the simple password managers built into web browsers) lack the significant protective features required to make a significant impact on password theft. Ideally, new tools, combining the best features from all existing systems, will begin to achieve widespread deployment in the near future.

STARTING FROM (MOSTLY) SCRATCH: USABLE WIRELESS SECURITY

In the parlance of the introduction, passwords are an existing security technology that have been subject to research, both by attempting to characterize their flaws and to improve their usability through the design of better interfaces and support tools. In this section we turn to the question of combining new and existing security tools to make qualitative changes in users' experience of secure computing systems—making systems so attractive for their usability that good security is simply a nice bonus.

Take the simple problem of securing a wireless network (WLAN). The security options available to a home user to protect such a network involve providing every device on that network with the same secret key. Removing an unwanted device from the network requires changing that key on every other device. Until sometime in 2006 the protection afforded to users that performed all of these steps correctly was in name only; the security mechanisms (WEP, or Wired Equivalent Privacy) built into IEEE 802.11, the most common WLAN standard, were completely ineffective (Walker, 2000; Stubblefield et al., 2002). In 2006 an improved version of the security standards for WLANs brought protection to end users, but still at a cost: vulnerabilities in the shared key-based security intended for home users allowed offline guessing of secret keys and required users to enter very long keys (e.g., 26 hex digits). While recent work by manufacturers to simplify the key distribution process for home users (e.g., WiFi Protected Setup) has improved usability somewhat, it has not made available to those users the more sophisticated forms of WLAN security geared toward enterprises.

Enterprise WLAN security, in contrast, offers a number of alternatives with much higher security guarantees than that provided to home users. At the limit, enterprise WLAN users can be individually authenticated using digital certificates and provided with separate keys for encrypting data; this provides strong authentication and network access control, the ability to revoke individual user's access easily, and protects network users from one another. However, availing oneself of such high security requires deploying a Public Key Infrastructure (PKI) and issuing digital certificates, something considered so difficult that even most professionally managed enterprises do not attempt it.

Our fundamental approach to building systems that are both usable and secure is to focus on users and their application tasks (which is all they really care about anyway) and see whether we can arrive at system designs that allow them to accomplish those tasks effectively and securely, without adding additional requirements or burdens. In Balfanz et al. (2004b) we turned that approach to the problem of deploying secure WLANs.

One of the most significant usability problems in deploying PKIs, even in enterprise environments, is the idea that such PKIs must be designed in the grand-scale, global infrastructure form in which such tools were first envisioned. If, instead, one changes the focus to building small-scale PKIs, requiring no interoperability or trust with any other infrastructure and targeted only to the task at hand (say the small group of devices allowed onto one WLAN), they turn out to be simple tools that are easy to deploy and manage (Balfanz et al., 2005). By building certification authority software (the sort of software that issues digital certificates) into a wireless access point (AP), as well as an enterprise-style authentication server configured to allow only devices certified by that authority onto the WLAN, we created a standalone, home-use AP that automatically configured itself to provide a highly secure WLAN as soon as it was turned on for the first time, with no user intervention. In a later enterprise form of the same system we wired together our system for enrolling new WLAN clients (see below) into a standard enterprise certification authority and authentication server to achieve the same effect while allowing enterprise-level management of digital certificates and network access policy.

The other, more critical challenge in designing a user-friendly approach to securing WLANs is how users accomplish the one task they must do, namely, specifying the WLAN they wish their devices to join, and indicating to their access points which devices should be allowed to do so. In the simplest usage of WLANs, networks are totally unsecured and users simply "fall onto" any available network. However, it is not possible to do that and achieve any level of security, particularly the high levels of security we are attempting to provide here. Instead, we would like to ask users to do the minimum work possible, namely, to indicate to the network that they would like to join it and from that simple indication achieve fine-grained security and control.

We achieve this through the use of a technique we refer to as "demonstrative identification." It is not possible for two devices that share no *a priori* trust information to authenticate each other over an unsecured medium such as a WLAN without risking a man-in-the-middle attack. So instead of having our candidate wireless device and AP find each other over the airwaves, we ask the user to point out the AP hosting the network they desire to join over what we refer to as a "location-limited channel." Such a channel is one where the nature of the medium itself makes it difficult for an attacker to transmit information in the channel without being detected, for example, infrared (as with a remote control), physical contact (as with cables or USB tokens), and sound. Over this channel we have the user's device and the access point exchange public information—the cryptographic digests of their public keys—which can later be used to allow the devices to authenticate each other over any channel they happen to use to communicate (e.g., the WLAN itself).

From the users' points of view, they are merely indicating their desired WLAN by pointing out the AP (or an enrollment device acting as a proxy for the AP) using infrared, as if they were using a remote control. From the AP's point of view, it implements a simple access control policy that says if a user is able to walk up to it and communicate with it over infrared (or audio or physical connection), that user and device is allowed on the WLAN. Having exchanged public key authenticators in this manner, along with a certain amount of configuration information, the user's device can set up a secure, authenticated connection to the AP over the existing wireless network. The AP can then use that connection to download to the device a digital certificate sufficient to allow it to authenticate as a user of the WLAN using standard protocols, and a small amount of software is sufficient to automatically configure the device to use that certificate in this way in the future. So from the point of view of the user, a small demonstrative act, which in experimental tests is perceived as simpler than the amount of manual configuration required to get a device onto a network providing lower levels of security, is all that is required to set his device up on a highly secure WLAN. After that initial setup phase, use of the secure WLAN is no more complicated than using any other WLAN. From the point of view of the WLAN owner, a highly intuitive security perimeter has been established: to get on the WLAN someone must have physical access (e.g., sufficient to communicate over infrared) to the access point. Without such access they cannot get onto the network. Therefore, securing your WLAN becomes equivalent to locking your front door, or if that isn't enough, locking the AP in a closet. And at the same time, the resulting network provides best available enterprise-class WLAN security, with per-user encryption keys and the ability to revoke access by any device at any time without requiring reconfiguration of any other device.

This simple system serves as just one example of how taking a slightly different, user-focused approach to the design of secure systems can result in systems that are easier to use than their insecure counterparts, while providing very high degrees of security. We have used this approach over the last several years to construct a number of such proof-of-concept systems, as well as components and tools that make building such systems easier in general.

REFERENCES

Adams, A., and M. A. Sasse. 1999. Users are not the enemy: Why users compromise computer security mechanisms and how to take remedial measures. Communications of the ACM 42(12):40-46.

Balfanz, D., G. Durfee, and D. K. Smetters. 2005. Making the impossible easy: Usable PKI. Pp. 319-334 in Security and Usability: Designing Secure Systems that People Can Use, L. F. Cranor and S. Garfinkel, eds. Sebastopol, CA: O'Reilly Media, Inc.

Balfanz, D., G. Durfee, R. E. Grinter, and D. K. Smetters. 2004a. In search of usable security—five lessons from the field. IEEE Security and Privacy Magazine 5(2):19-24.

Balfanz, D., G. Durfee, R. E. Grinter, D. K. Smetters, and P. Stewart. 2004b. Network-in-a-box: How to set up a secure wireless network in under a minute. Pp. 207-222 in Proceedings of the 13th USENIX Security Symposium. Berkeley, CA: USENIX.

Brostoff, S., and M. A. Sasse. 2003. Ten strikes and you're out: Increasing the number of login attempts can improve password usability. Paper presented at Workshop on Human-Computer Interaction and Security Systems. Fort Lauderdale, FL, April 5-10.

Kuo, C., S. Romanosky, and L. F. Cranor. 2006. Human selection of mnemonic phrase-based passwords. Pp. 67-78 in SOUPS '06: Proceedings of the Second Symposium on Usable Privacy and Security. New York: ACM Press.

Schechter, S. E., R. Dhamija, A. Ozment, and I. Fischer. 2007. The emperor's new security indicators. Pp. 51-65 in SP '07: Proceedings of the 2007 IEEE Symposium on Security and Privacy. Piscataway, NJ: IEEE.

Smetters, D. K., and R. E. Grinter. 2002. Moving from the design of usable security technologies to the design of useful secure applications. Pp. 82-89 in New Security Paradigms Workshop. New York: ACM Press.

Stubblefield, A., J. Ioannidis, and A. D. Rubin. 2002. Using the Fluhrer, Mantin, and Shamir attack to break WEP. Paper presented at the Network and Distributed Systems Security Symposium. San Diego, CA. February 6-8.

Walker, J. 2000. Unsafe at any key size: An analysis of the WEP encapsulation. IEEE Document 802.11-00/362. Washington, DC: IEEE.

Wu, M., R. C. Miller, and G. Little. 2006. Web wallet: Preventing phishing attacks by revealing user intentions. Pp. 102-113 in SOUPS '06: Proceedings of the Second Symposium on Usable Privacy and Security. New York: ACM Press.

Yan, J., A. Blackwell, R. Anderson, and A. Grant. 2005. The memorability and security of passwords. Pp. 129-142 in Security and Usability: Designing Secure Systems that People Can Use, L. F. Cranor and S. Garfinkel, eds. Sebastopol, CA: O'Reilly and Associates.

Yee, K.-P., and K. Sitaker. 2006. Passpet: Convenient password management and phishing protection. Pp. 32-43 in SOUPS '06: Proceedings of the Second Symposium on Usable Privacy and Security. New York: ACM Press.

Control of Protein Conformations

Introduction

Donald J. Leo
Virginia Tech
Blacksburg, Virginia

The relationship between protein structure and protein function is central to understanding the properties of biological systems. Recent advances in the imaging of biological systems at the appropriate spatial and temporal scales have provided an improved understanding of biological functions such as cell signaling and the dynamic processes associated with protein-protein interactions. Advances in imaging techniques—when combined with the ability to directly manipulate proteins using mechanical, optical, and magnetic forces—provide a basis for real-time control of protein conformations for the purpose of directly controlling protein function. This session will explore recent advances in the ability to directly control the affinity and selectivity of proteins using an external stimulus. Methods such as mechanical manipulation, optical and magnetic tweezers, and chemical control of protein structure will be compared in terms of their ability to control protein conformation. Seminal advances in imaging will be discussed, and the ramifications of new techniques on the understanding of biological systems will be explored. A central theme of the session will be how new techniques for imaging, manipulating, and directly controlling biological systems will lead to new methods for engineering biological systems for applications in drug discovery, vaccine development, and new methods for tunable biosensors and bioassays.

The Evolutionary Design of Proteins

RAMA RANGANATHAN
Green Center for Systems Biology
University of Texas Southwestern Medical Center

Evolution builds proteins that display structural and functional properties that are beautifully suited for their biological role. They can fold spontaneously under physiological conditions into a compact and well-packed three-dimensional structure, and often display complex functional properties, such as signal transmission, efficient catalysis of chemical reactions, specificity in molecular recognition, and allosteric conformational change. Because these properties require great precision in the positioning and coupling of amino acids, the current view is to regard proteins as finely tuned machines that are exactly arranged for mediating their selected function. However, other aspects of proteins seem less consistent with this view and demand further delineation of the underlying design principles. For example, proteins are well known to be robust to random perturbation; that is, they display tolerance to mutagenesis at many amino acid positions. In addition, they are readily evolvable; that is, they display the capacity for rapid adaptation to changing selection pressures by allowing specific sequence variation at a few sites to profoundly alter function. This curious mixture of robustness at many sites and fragility at a few sites is interesting since it suggests that despite homogeneously good atomic packing, strong heterogeneity exists in the energetic interactions between amino acids that underlie structural stability and function. A major advance in our understanding of proteins would come with the development of methods to systematically map and then mechanistically understand the global architecture of amino acid interactions. In principle, such understanding would (1) provide a new basis for the interpretation of protein sequence and structure,

(2) provide powerful practical rules for the rational engineering of protein structure and function, (3) help predict and better understand the molecular basis for disease-causing mutations, and (4) provide the basis for beginning to understand how proteins are even possible through the random, algorithmic process of mutation and selection that we call evolution.

However, the problem is truly complex; amino acids make unequal and cooperative contributions to protein structure and function, and these contributions are generally not obvious in even high-resolution atomic structures. For example, studies of the interaction between the human growth hormone and its receptor show that the binding interface contains "hot spots" of favorable energetic interactions embedded within an overall environment of neutral interactions (Clarkson and Wells, 1995). Similarly, catalytic specificity in proteases (Hedstrom, 1996), signal transmission within G protein coupled receptors (GPCRs) (Gether, 2000) and the cooperative binding of oxygen molecules in hemoglobin (Perutz et al., 1998), catalysis in the metabolic enzyme dihydrofolate reductase (Benkovic and Hammes-Schiffer, 2003), and antigen recognition by antibody molecules (Midelfort and Wittrup, 2006; Patten et al., 1996) all depend on the concerted action of a specific set of amino acids that are distributed both near and far from the active site. Why are these energetic phenomena not obvious in atomic structures? The main problem is easily stated: we do not "see" energy in protein structures. We might observe an interaction in a crystal structure, but we do not know the net free energy value of that interaction given only the mean atomic positions. Since the native state of a protein represents a fine balance of opposing forces that operate with steep distance dependencies to produce marginally stable structures, complex and nonintuitive arrangements of amino acid interactions are possible.

These observations permit a clear statement of the goals, and that will constitute this lecture. The essence of understanding the evolutionary design of protein structure and function is globally assessing the energetic value of all amino acid interactions. Since the value of interactions is not a simple function of distance, and complex spatial arrangements of interactions between amino acids are possible, we must be open to novel strategies that go beyond structure-based inferences or high-throughput mutagenesis. In this regard we have reported a novel statistical approach (now termed "statistical coupling analysis," or SCA) for globally estimating amino acid interactions (Lockless and Ranganathan, 1999). Treating evolution as a large-scale experiment in mutation allows this method to make the simple proposition that the energetic coupling of residues in a protein (whether for structural or functional reasons) should force the mutual evolution of those sites. That is, the conserved cooperative interactions between amino acids might be exposed through analysis of the higher order statistics of sequence variation between positions in a large and diverse multiple sequence alignment of a protein family. Application of this method in several different protein families—PDZ (Post-synaptic density 95, Discs Large, Zona Occludens 1) domains (Lockless and Ranganathan, 1999), GPCRs (Suel et al., 2003), serine proteases

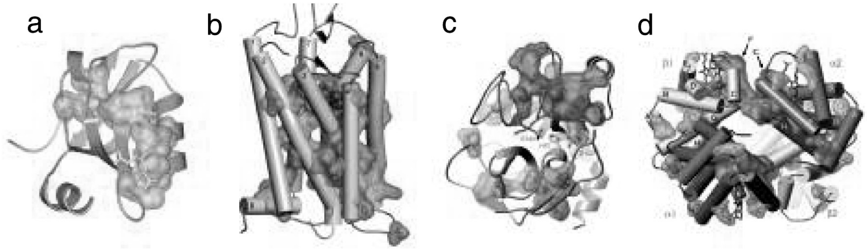


FIGURE 1 Networks of evolutionarily coupled residues in the PDZ (a), GPCR (b), serine protease (c), and hemoglobin (d) protein families. A representative member of each family is shown with a van der Waals surface drawn around the residues that emerge as strongly evolutionarily coupled in the SCA.

(Suel at al., 2003), hemoglobins (Suel et al., 2003), guanine nucleotide binding (G) proteins (Hatley et al., 2003), and the nuclear hormone receptors (Shulman et al., 2004)—suggests two general conclusions (Figure 1): (1) the global pattern of amino acid interactions is sparse, so that a small set of positions mutually co-evolves among a majority that are largely decoupled, and (2) the strongly co-evolving residues are spatially organized into physically connected networks linking distant functional sites in the structure through packing interactions. Importantly, mutagenesis experiments directed by the SCA mapping show strong correlations between the predictions and experimental measurements, implicating the co-evolving networks as hot spots for functional mechanism. Taken together these studies suggest a new model for the architecture of amino acid interactions in natural proteins: most residues interact minimally, acting as if nearly independent or locally coupled, while a few (~10-15 percent) comprise strongly interacting, sparse, and interconnected networks of co-evolving residues that define core aspects of protein function.

The finding that the global pattern of energetic interactions between residues in many protein families is so sparse suggested an interesting and testable hypothesis: proteins may be far simpler in their energetic architecture than we think. Indeed, the SCA suggests the possibility that all the information required for specifying the fold and characteristic function of a protein family may be sufficiently encoded in the matrix of amino acid interactions revealed by the global co-evolution analysis. If so, the proof lies in building artificial members of a protein family using only information extracted by the SCA. We developed a computational method for creation of artificial amino acid sequences according to the statistical rules revealed by the SCA and built large libraries of designed sequences in the laboratory (Russ et al., 2005; Socolich et al., 2005). We find that for the WW (Trp-Trp) domain, a small β-structured protein fold containing

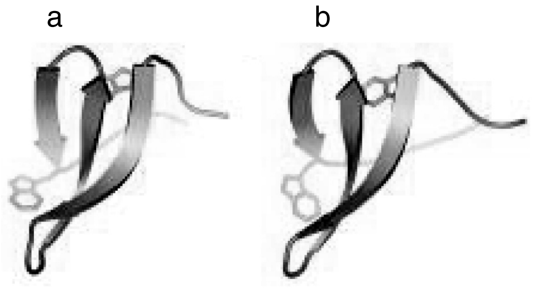


FIGURE 2 A comparison of a 1.45Å crystal structure of the human Pin-1 WW domain (a) and a NMR structure of an artificial WW domain, cc45 (b). The root mean squared deviation of backbone atoms between cc45 and all natural WW domains is within that of natural domains to each other. CC45 shows 35 percent mean identity and 61 percent top-hit identity with natural WW domains.

two conserved tryptophan residues, the computational method produces artificial sequences that in fact efficiently fold into the characteristic WW domain structure (Figure 2). In contrast, randomly designed sequences, or even sequences preserving the conservation pattern of the WW domain but excluding the co-evolution of residues, failed to show native folding. Remarkably, phage display and peptide binding studies indicate that the artificial sequences not only fold but also exhibit binding specificity that quantitatively mirrors that of natural WW domains (Russ et al., 2005). More recently these findings have been extended to the successful design of functional PDZ domains, a roughly 90 amino acid protein interaction module with a mixed α/β fold. These results demonstrate that the information extracted from the SCA is necessary and sufficient to specify the WW and PDZ folds and their characteristic function. The very few numbers we imposed in the computational process in building these artificial sequences suggests that evolution's rules for building proteins could be vastly less complex than theoretically possible. These results suggest a new reparameterization of proteins guided by the higher order statistics of conservation rather than by traditional principles of primary, secondary, or tertiary structure. Recent work on developing and testing this parameterization will be presented.

To date, our work has been focused on mapping and experimentally verifying the architecture of amino acid interactions implied by the evolutionary analysis—essentially, a mechanism-free description of what is built through the evolutionary process. The results have been unexpected and suggest potential practical avenues for directed mutagenesis and design of proteins. However, this work primarily has set the stage for two critical further problems that limit our understanding of the evolutionary design of proteins: (1) how does the SCA-predicted architecture of amino acid interactions physically operate in proteins and (2) why is the SCA-predicted architecture a good solution for specifying protein structure and function through evolution? In essence the goal is to ultimately connect the mechanism-free description of statistical interactions between amino acids with a mechanistic model for the physics of these interactions and with an underlying evolutionary theory. It will be interesting to see what experiments can be designed to test the theory that emerges from this work for the dynamics of the evolutionary process.

REFERENCES

Benkovic, S. J., and S. Hammes-Schiffer. 2003. A perspective on enzyme catalysis. Science 301(5637): 1196-1202.

Clackson, T., and J. A. Wells. 1995. A hot spot of binding energy in a hormone-receptor interface. Science 267(5196):383-386.

Gether, U. 2000. Uncovering molecular mechanisms involved in activation of G protein-coupled receptors. Endocrine Reviews 21(1):90-113.

Hatley, M. E., S. W. Lockless, S. K. Gibson, A. G. Gilman, and R. Ranganathan. 2003. Allosteric determinants in guanine nucleotide-binding proteins. Proceedings of the National Academy of Sciences 100(suppl. 2):14445-14450.

Hedstrom, L. 1996. Trypsin: A case study in the structural determinants of enzyme specificity. Biological Chemistry 377(7-8):465-470.

Lockless, S. W., and R. Ranganathan. 1999. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286(5438):295-299.

Midelfort, K. S., and K. D. Wittrup. 2006. Context-dependent mutations predominate in an engineered high-affinity single chain antibody fragment. Protein Science 15(2):324-334.

Patten, P. A., N. S. Gray, P. L. Yang, C. B. Marks, G. J. Wedemayer, J. J. Boniface, and P. G. Schutz. 1996. The immunological evolution of catalysis. Science 271:1086-1091.

Perutz, M. F., A. J. Wilkinson, M. Paoli, and G. G. Dodson. 1998. The stereochemical mechanism of the cooperative effects in hemoglobin revisited. Annual Review of Biophysics and Bimolecular Structure 27:1-34.

Russ, W. P., D. M. Lowery, P. Mishra, M. B. Yaffe, and R. Ranganathan. 2005. Natural-like function in artificial WW domains. Nature 437(7058):579-583.

Shulman, A. I., C. Larson, D. J. Mangelsdorf, and R. Ranganathan. 2004. Structural determinants of allosteric ligand activation in RXR heterodimers. Cell 116(3):417-429.

Socolich, M., S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, and R. Ranganatahan. 2005. Evolutionary information for specifying a protein fold. Nature 437(7058):512-518.

Suel, G. M., S. W. Lockless, M. A. Wall, and R. Ranganathan. 2003. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nature Structural Biology 10(1):59-69.

Lighting Up the Mechanome

Matthew J. Lang
Massachusetts Institute of Technology
Cambridge, Massachusetts

Genomics and general "omics" approaches, such as proteomics, interactomics, and glycomics, among others, are powerful system-level tools in biology. All of these fields are supported by advances in instrumentation and computation, such as DNA sequencers and mass spectrometers, which were critical in "sequencing" the human genome. Revolutionary advances in biology that have harnessed the machinery of polymerases to copy DNA or used RNA interference to capture and turn off genes have also been critical to the development of these tools. Researchers in biophysics and bio-mechanics have made significant advances in experimental methods, and the development of new microscopes has made higher resolution molecular- and cellular-scale measurements possible.

Despite these advances, however, no systems-level "omics" perspective—the "mechanome"—has been developed to describe the general role of force, mechanics, and machinery in biology. Like the genome, the mechanome promises to improve our understanding of biological machinery and ultimately open the way to new strategies and targets for fighting disease.

Many disease processes are mechanical in nature. Malaria, which infects 500 million people and causes 2 million deaths each year, works through a number of mechanical processes, including parasite invasion of red blood cells (RBCs), increased RBC adhesions, and overall stiffening of RBCs. As the disease progresses, it impairs the ability of infected cells to squeeze through capillaries, ruptures RBC membranes, and releases newly minted malaria parasites.[1]

Cancer, another leading cause of death, also progresses in many mechanical ways, such as the machinery of viral infection, which is responsible for hepatitis B, human papilloma virus, and related cancers. In addition, metastasis involves the detachment of tumor cells from the initial growth site, migration, circulation, and subsequent invasion through reattachment to new locations.

Intracellular changes in cancerous cells bring about cell growth and proliferation, cell motility, and migration. These processes are currently targeted by Taxol, a potent chemotherapy drug that impedes the progression of cancer by stabilizing microtubules in a polymerized state, thus shutting down the machinery of cell division.

The human genome is contained in approximately 3 billion base pairs of DNA. If the pieces of DNA in our cells were connected end-to-end, the chain would be roughly as long as one's outstretched arms. Compared to other cellular filaments, DNA is quite flexible. To fit into cells, it is compacted in highly ordered structures through multiple hierarchical levels.[2] In addition to the conventional Watson and Crick base pairs, DNA forces and interactions include histone protein-DNA nucleosome complexes, chromatin fibers, chromatids, and, ultimately, 46 chromosomes. In general, our biological structures have many levels of structural hierarchy and are highly organized.

MOLECULAR MACHINERY

From our DNA sequence, biological motors, including DNA and RNA polymerases and the ribosome, make RNA, and ultimately proteins, through coupled mechanical and chemical steps. This impressive machinery copies and translates with great accuracy. The ribosome, a huge motor spanning 20 nanometers (nm), which is actually composed of about 65 percent ribosomal RNA, assembles proteins from messenger RNA. Many antibiotics target the critical machinery of ribosomal translation in bacteria.

Other molecular motors, such as myosins and kinesins, run on different cytoskeletal tracks—actin filaments and microtubules, respectively.[3] Myosins, which are responsible for the contraction of skeletal and cardiac muscle, contribute to cell

[1] A red blood cell circulates in the body about half a million times in a 120-day period.

[2] Human cells are typically 10-30 μm in diameter. The smallest, sperm, is only a few microns in diameter, and the largest, the egg, is ~100 μm in diameter. The size of randomly packed DNA can be estimated from the radius of gyration ($b \cdot N^{1/2}$), where N is the number of segments and b is the length of each segment. DNA has a length of 0.3 nm/bp, a persistence length of 50 nm. An average chromosome may contain 1.3×10^8 bp of DNA, yielding a radius of gyration of 44 μm, larger than the typical cell size. Chromosomes were first observed by Karl Wilhelm von Nägeli in 1842.

[3] Mechanically, actin filaments have a Young's modulus of 2.3 GPa, and microtubules have a persistence length on the order of mm.

migration and cell-division processes, such as contractile ring closure. Similarly, kinesins, which are important in transport and cell division, are small, processive motors that run toward the plus (fast polymerizing) end of microtubules. In some cases, such as the kinesins that carry vesicles containing neurotransmitters along the sciatic nerve axon, this progress takes place over the length scale of about a meter and is a critical mechanism for cargo delivery.[4]

The design of biological motors can be classified by cataloging the motor's general structural features, fuel type, stepping distance, stall force, and other mechanical parameters. Detailed measurements of the motility cycles and underlying mechanisms for motility also provide information about how these mechanisms work.[5] Ultimately, "sequencing" the mechanome will lead to the discovery of the design features of biological motors in general, enabling us to catalogue them and outline the rules that govern their behavior.

Comparisons of the family trees of kinesins and myosins reveal a range of structural forms, including single-, double-, and tetramer-headed members of various lengths connecting the motor and cargo-binding domains. They also share some structural elements at the molecular level, such as the binding pocket for adenosine triphosphate (ATP) (Vale, 1996), and both are fueled by hydrolysis of ATP. Another example of a common motor design is in AAA+ motors, which are similarly structured but run on a variety of track types (e.g., microtubules, DNA, and peptides) and are involved in a number of critical cellular processes (Ogura and Wilkinson, 2001). Nature, it appears, copies and pastes design elements for motors with similar roles and physical constraints. Despite different tracks and overall purposes for motility, all of the designs of common biological machines have been optimized.

CELLULAR MACHINERY

Cellular machinery (e.g., the processes of cell division, wound healing, and migration) can also be classified by general design principles. Forces originating from many sources (e.g., biological motors, polymerization of filaments, molecular conformational transitions, and fluid pressure) can act on structures, including cytoskeletal filaments, membrane barriers, receptor ligand linkages, and

[4] Kinesin has a velocity of 800 nm/s. Thus traveling the sciatic nerve might take more than 2 weeks. Diffusion alone would not be complete in a lifetime. This estimate can be obtained from the Einstein-Sutherland-Smoluchowski relation (a 1 μm diameter cargo has a diffusion constant of $D = kT/6\pi\eta a = 2\times10^{-9}$ cm^2/s). Diffusion time is proportional to the distance squared $x^2/D = 7\times10^4$ years.

[5] Kinesin, for example, has been extensively characterized. It has a stepping distance of 8 nm for each ATP molecule consumed. Steps occur upon ATP binding. Kinesin takes about 100 steps/s and runs at 800 nm/s. After about 1 μm, it loses hold of its track. Kinesin coordinates the action of the two heads, shuts down its hydrolysis in the absence of cargo, stalls at a load of about 5-6 pN, and walks much like we do.

protein-complex adhesion sites. These interactions typically occur in cycles, such as migration, extension processes, adhesion, contraction, and detachment.

Highly tuned control over these forces and structures is orchestrated through mechanical and biochemical signals that initiate and terminate motility cycles. Mechanotransduction, a critical aspect of control, refers to the process whereby cells actively respond to mechanical signals through changes in morphology, signaling, and biochemical response.

Migration of the neuronal growth cone exhibits a range of interesting cell machinery. The main core structure of the growth-cone axon consists of bundled microtubules. At the tip of this bundle, actin filaments are structured in both bundled finger-like spikes (fillo-podia) and web-like flattened sheets (lamellipodia). Actin filaments can be arranged in a variety of higher level structures coordinated by a series of actin-binding proteins (ABPs) (Figure 1).

The overall structures of ABPs vary widely, but many have a common actin-binding domain. In cell division, an actin-based ring-like bundle contracts to separate the cell into progeny cells. The design elements of this process, called cytokinesis, appear to be cut and pasted. For example, in cell division of higher plants, a similar process generates a microtubule-based structure at the cell-division locus (Jurgens, 2005). Even bacteria contain tubulin and actin homologues (FtsZ and MreB, respectively) that form higher level structures and machinery that resembles their counterparts, such as microtubules (MreB) and

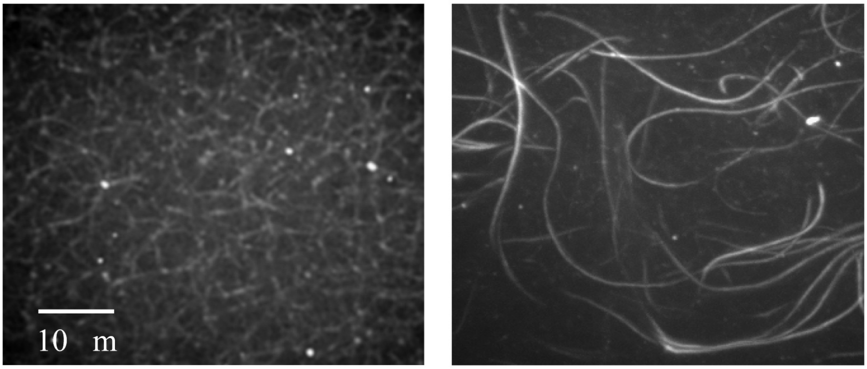


FIGURE 1 The images above show networks of actin filaments organized by different actin-binding proteins (ABPs). The ABP on the left, consisting of filamin, forms a meshwork arrangement of filaments. The ABP on the right, α-actinin, organizes actin filaments into bundles. Despite differences in network morphology, filamin and α-actinin have similar actin-binding domains. Source: Photos courtesy of Hyungsuk Lee and Jorge Ferrer, in collaboration with Prof. Roger Kamm's laboratory, Massachusetts Institute of Technology.

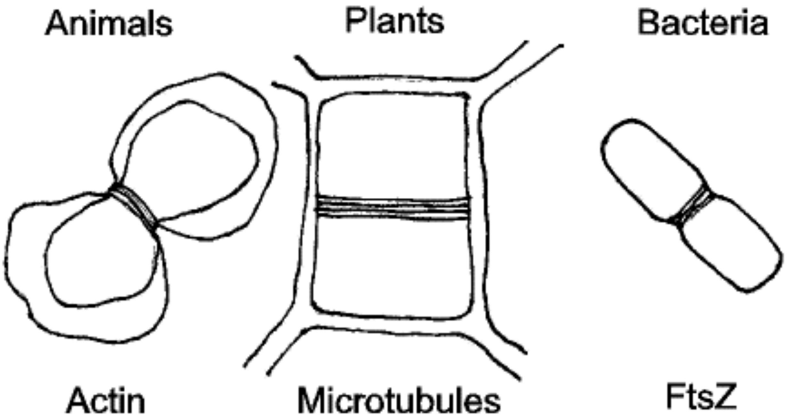


FIGURE 2 This cartoon shows similar stages of cell division in animal cells, plant cells, and bacteria. In all three, a common structural end point is achieved by ring formation through different types of filaments: actin filaments, microtubules, and FtsZ, respectively.

the formation of the cytokinetic ring in bacteria (FtsZ) (Erickson, 1998; van den Ent et al., 2001) (Figure 2).

TISSUE AND HIGHER LEVEL MACHINERY

Mechanical processes are critical, not only at the molecular level, but also at the tissue level and higher levels. In fact, the mechanics of the extracellular environment is an important factor in determining its fate. Stem cells plated on substrates of different stiffnesses differentiate into different cell types, taking on properties of tissues similar to their mechanical environments. For example, mesenchymal stem cells grown on a matrix with elasticity resembling that of tissues such as 1 kPa brain, 10 kPa muscle, and 100 kPa bone differentiate into cells resembling neurons, muscles, and osteoblasts, respectively (Engler et al., 2006).[6]

The external environment of a cell contains many organized mechanical structures. A typical extracellular matrix is composed of cable-like collagen molecules, rubber band-like elastin molecules, and sponge-like proteo-glycans. So, for example, tendons contain collagen organized into a bundle form with many levels of hierarchy, including mineralization. Tendons are designed to withstand

[6] kPA, or kilo-Pascal, is a unit of pressure defined as force for a given area, Newtons per meter squared.

low strain and fatigue and can sustain many cycles of loading and unloading over a long period of time. Tendon failure is a common sports injury.

In contrast to a tendon, the dense collagen network of the chorioamnion membrane surrounding a fetus is designed to eventually fail. Early failure in this structure is responsible for one-third of all premature births (Oyen et al., 2005).

Mechanical forces may also affect the organization of structures. Bone requires mechanical stimulation to maintain density, and collagen stress fibers in bone align in the direction of maximal load. Higher level mechanical systems include musculoskeletal, respiratory, cardiovascular, lymphatic, and integumental systems.

MEASUREMENTS AT THE MOLECULAR LEVEL

With the advanced instrumentation and experimental methods described below, we are poised to measure the mechanome at the molecular level. For example, force microscopy, in which controlled loads are applied to structures while they are tracked with nanometer-level resolution, can be used to measure mechanical transitions in unfolding protein structures, the rupture of individual molecular interactions, and the motility of biological motors.

A huge range of forces are required to make these measurements. At the molecular level, a balance is obtained between stabilizing an interaction against thermal energy (kT) and maintaining the ability to break an interaction through transition-state barriers of a few kT. Given that kT is in the range of 4 pN-nm (or 4.1×10^{-21} J),[7] and typical transition distances are on the length scale of nanometers, forces ranging from 1 pN to a few tens of pN are ideal for probing molecular machinery.

ATP hydrolysis, the energy source for many biological motors, provides about 20 to 25 kT of energy. Unfolding proteins requires a force of tens to hundreds of pN. Furthermore, because cell-level machinery may involve many molecular interactions, force magnitudes on the scale of nN and more are sometimes required.

A broad range of complementary force-probe methods can meet these force requirements. These include the atomic force microscope (AFM), micropipette manipulation, magnetic traps, microfabricated cantilevers, and optical traps (Suresh, 2007). Combinations of these force probes, such as combined optical trapping and magnetic-particle manipulation, can extend the force range. The optical trap, which has been a mainstay technique in single-molecule biophysics research, is an ideal tool for probing the molecular interactions of biological motors and receptor-ligand interactions.

Fluorescence microscopy is another powerful method of observing single

[7] pN, or picoNewton, is a unit of force defined as 10^{-12} Newtons.

molecules. With advances in this technology, it can be used to track fluorophore-labeled biological motors with nanometer-level resolution (Yildiz et al., 2004). Energy-transfer methods using pairs of fluorophores (FRET) provide a fluorescence-based ruler with a $1/r^6$ dependence,[8] on the length scale of nanometers, which is useful for monitoring conformational changes.

Recently, optical traps and FRET have been combined on a single platform (Tarsa et al., 2007). The combination marries the power of mechanical interrogation and control of structures with the ability to observe them with fluorescence. This combined method has been demonstrated using a model system consisting of a DNA-based hairpin that can be opened and closed with about 15 pN of force from the optical trap (Figure 3). A FRET pair placed at the base of the hairpin reported whether the hairpin was open or closed through high and low FRET states, respectively. Many additional advances, such as light-based methods of laser scissors, photobleach recovery, triggering, uncaging, and time-resolved measurements, can be integrated to produce powerful new tools.

Significant advances have also been made in assay technologies, molecular biology, and linkage chemistry. Molecular biology strategies, such as the encodeable green fluorescence protein (GFP), which can genetically label proteins, have advanced the field, as have new probes, such as enhanced organic dyes and quantum dots, which have enabled observations of single molecules over extended periods of time. Molecular biology strategies can also be used to create physical handles for the manipulation of molecules, thus exploiting, for example, the aptamer interactions and tethers based on the M13 phage (Khalil et al., 2007).

CONCLUSION

With the tools and advances in assays I have described, we are beginning to probe nature's machinery. Ultimately, we hope to gain a quantitative understanding of molecular and cellular machinery and to measure the forces and strength of the mechanical steps that underlie the mechanome. Once we understand the physical rules that govern biological systems and can measure nature's machinery, we should eventually be able to "sequence" the mechanome.

[8] The separation, or distance, between fluorophores, *r*, is usually on the order of 5 nanometers, or 5×10^{-9} m.

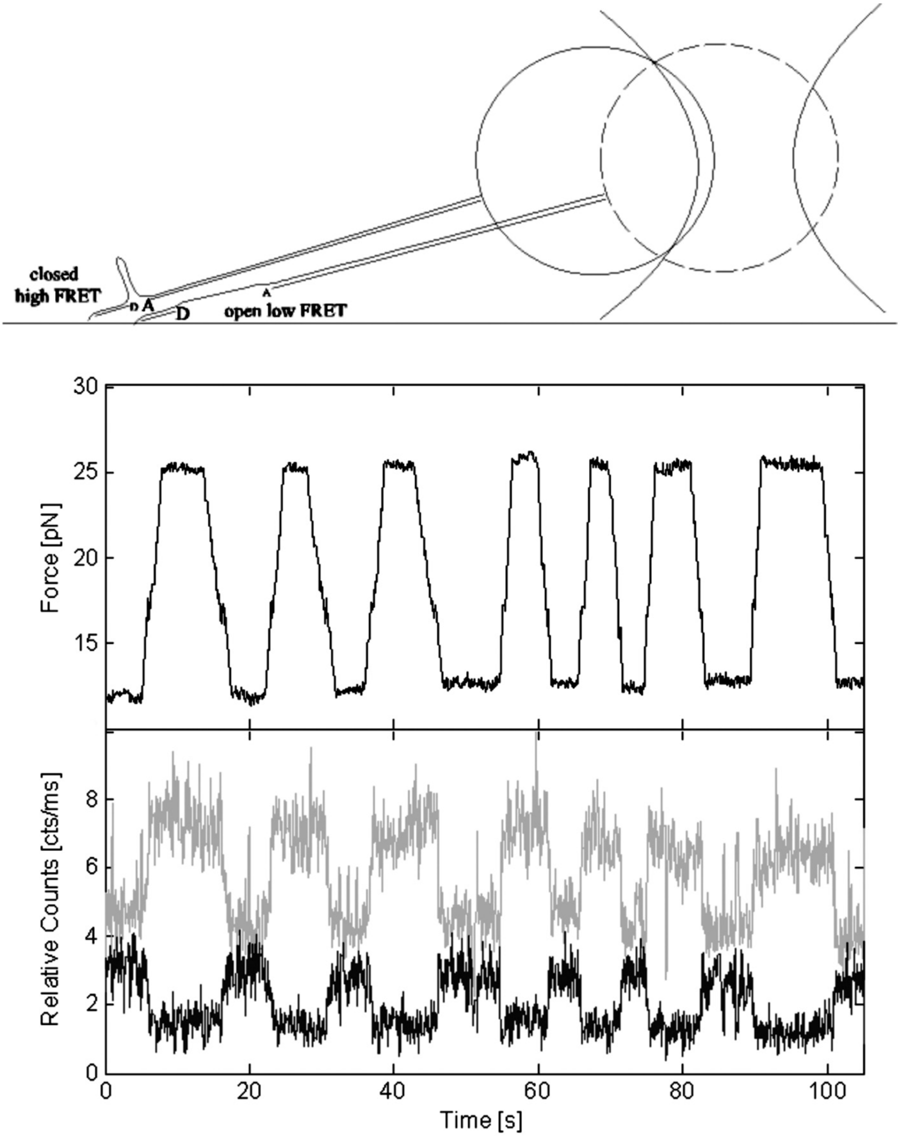


FIGURE 3 The upper frame is a cartoon showing the experimental design with open and closed states of a DNA hairpin complex controlled with force from an optically trapped bead. A donor and acceptor FRET pair is placed at the base of the hairpin to monitor whether or not it is open or closed. The middle frame shows how force on the hairpin is repeatedly transitioned from low to high with hairpin opening occurring at approximately ~18pN. The lower frame shows photon counts with donor (upper) and acceptor (lower) FRET signals changing simultaneously with the opening and closing of the hairpin.

REFERENCES

Engler, A. J., S. Sen, H. L. Sweeney, and D. E. Discher. 2006. Matrix elasticity directs stem cell lineage specification. Cell 126(4):677-689.

Erickson, H. P. 1998. Atomic structures of tubulin and FtsZ. Trends in Cell Biology 8(4):133-137.

Jurgens, G. 2005. Plant cytokinesis: Fission by fusion. Trends in Cell Biology 15(5):277-283.

Khalil, A. S., J. M. Ferrer, R. R. Brau, S. T. Kottmann, C. J. Noren, M. J. Lang, and A. M. Belcher. 2007. Single M13 bacteriophage tethering and stretching. Proceedings of the National Academy of Sciences 104(12):4892-4897.

Ogura, T., and A. J. Wilkinson. 2001. AAA(+) superfamily ATPases: Common structure-diverse function. Genes to Cells 6(7):575-597.

Oyen, M. L., R. F. Cook, T. Stylianopoulos, V. H. Barocas, S. E. Calvin, and D. V. Landers. 2005. Uniaxial and biaxial mechanical behavior of human amnion. Journal of Materials Research 20(11):2902-2909.

Suresh, S. 2007. Biomechanics and biophysics of cancer cells. Acta Biomaterialia 3(4):413-438.

Tarsa, P. B., R. R. Brau, M. Barch, J. M. Ferrer, Y. Freyzon, P. Matsudaira, and M. J. Lang. 2007. Detecting force-induced molecular transitions with fluorescence resonant energy transfer. Angewandte Chemie (International edition in English) 46(12):1999-2001.

Vale, R. D. 1996. Switches, latches, and amplifiers: Common themes of G proteins and molecular motors. Journal of Cell Biology 135(2):291-302.

van den Ent, F., L. A. Amos, and J. Lowe. 2001. Prokaryotic origin of the actin cytoskeleton. Nature 413(6851):39-44.

Yildiz, A., M. Tomishige, R. D. Vale, and P. R. Selvin. 2004. Kinesin walks hand-over-hand. Science 303(5658):676-678.

Biotechnology for Fuels and Chemicals

Introduction

Richard T. Elander
National Renewable Energy Laboratory
Golden, Colorado

Vijay Singh
University of Illinois at Urbana-Champaign

Recent events are once again highlighting that the United States is overly dependent upon imported petroleum for its energy and chemical feedstock supply. This is particularly true in the liquid transportation fuels and chemical production sectors of our economy. Additionally, the ever-increasing use of petroleum as a feedstock for fuels and chemicals in the worldwide economy is causing growing concerns about the availability and security of future petroleum supplies and the implications of a petrochemical-based economy on global climate change. These factors have resulted in a resurgence of interest in the development and commercial deployment of renewable, sustainable, environmentally friendly, and cost-effective technologies to supply a large fraction of our liquid transportation fuels and chemical feedstock needs. Many of these technologies are now achieving or are rapidly approaching technical and economic viability in the commercial marketplace and are utilizing the most advanced biotechnology, chemical reaction, and engineering tools. In this session, we examine recent advances that are bringing "biorefineries" to commercial reality. The unique aspects of grain-based, lignocellulose-based, and integrated grain and lignocellulose biorefineries will be discussed.

Corn-Based Materials

Sanjay V. Malhotra and Vineet Kumar
National Cancer Institute
Frederick, Maryland

Anthony East and Michael Jaffe
New Jersey Institute of Technology
Newark, New Jersey

The United States produces 44 percent of the world's corn. The largest crop grown in the United States, corn occupies some 70 million acres. This versatile, natural, biodegradable, and renewable resource has many commercial applications. As a raw material, it is replacing petroleum in many industrial applications, from plastic containers to clean-burning ethanol. One current goal is to produce 5 billion gallons of ethanol from corn annually. Corn components can also be found in thousands of other products, ranging from foods, drugs, cosmetics, and cleansers to adhesives and shoe polish. For example, corn oil is used in emollient creams and toothpaste, and corn syrup is used as a texturing and carrying agent in cosmetics.

The development of biodegradable polymers is a high priority from the standpoint of environmental preservation; biodegradable polymers from corn chemistry have a number of advantages over other synthetic materials. Polymers derived from starch (a component of corn) or other carbohydrates made from entirely renewable resources have been used in the manufacture of quality plastics, packaging materials, and fabrics (Koch and Röper, 1988), as well a variety of biomedical materials (e.g., bioMEMs, biochips, bioscaffolds, etc.). This article focuses on new technologies arising from corn-based chemistry and their applications in the polymer industry and biomedicine, cosmetics and drug-delivery systems, and pharmaceuticals.

STARCH-BASED, BIODEGRADABLE POLYMERS

Starch, a major component of corn, is a linear polysaccaride made up of repeating glucose groups with glycosidic linkages in the 1-4 carbon positions with chain lengths of 500 to 2,000 glucose units. There are two major polymer molecules in starch—amylose and amylopectin. Amylose, which has alpha linkage, is both flexible and digestible.

As the starch content of a polymer increases, it becomes increasingly biodegradable and leaves fewer intractable residues. Biodegradation of starch-based polymers is the result of enzymatic attack on glycosidic linkages between the sugar groups, which leads to decreases in chain length and the "splitting out" of lower molecular-weight sugar units (monosaccharides, disaccharides, and oligosaccharides) that can be readily used in biochemical pathways.

Biodegradable starch-based polymers are being investigated for potential use in several biomedical applications. For example, new processing techniques and reinforcement with various fillers have led to the development of materials with mechanical properties comparable to those of bone (Reis et al., 1997a). These polymers may be suitable for bone-replacement implants (Reis and Cunha, 1995), bone cements (Reis et al., 1997b), drug-delivery systems (Malafaya et al., 2001), and tissue-engineering scaffolds (Gomes et al., 2001).

Thermoplastic-Starch Products

When biodegradable thermoplastic-starch plastics, which have a starch (amylose) content of more than 70 percent, are combined with specific plasticizing solvents, they produce thermoplastic materials with good performance properties and inherent biodegradability. The hydrophilic nature of high-starch content plastics, which readily disintegrate on contact with water, can be overcome by blending and chemically modifying the material by acetylation, esterification, or etherification. Starch-based polymers are often plasticized, destructured, and/or blended with other high-performance polymers (e.g., aliphatic polyesters and polyvinyl alcohols) to provide useful mechanical properties for different packaging applications.

Starch-Aliphatic Polyester Blends

Blends of biodegradable, synthetic, aliphatic polyesters and starch are often used to produce high-quality sheeting and films for packaging. Approximately 50 percent of synthetic polyester (which costs approximately $4.00/kg) could be replaced with natural polymers, such as starch (approximately $1.50/kg). In addition, polyesters can be modified by incorporating different functional groups (e.g., hydroxy, amine, carbonyl, etc.) that are capable of reacting with natural starch polymers.

Several products made of starch-based plastics are currently available commercially. These include films (e.g., shopping bags, bread bags, bait bags, overwraps, and backing materials for "flushable" sanitary products) and mulch film. One commercial product, the "BioBag," is produced from the Novamont resin (found in Italy), which has been around since 1994. The BioBag is made from corn starch combined with fully biodegradable plastics or polylactic acid.

Other products are produced from starch-based foams. A well-known example is water-soluble beads made from potato starch that can replace polystyrene-foam packaging beads. Starch-based beads, which are readily soluble and biodegradable, have created an early market for biodegradable plastics.

Starch and PBS/PBSA Polyester Blends

Polybutylene succinate (PBS), polybutylene succinate adipate (PBSA), and other polyesters can be blended with starch to improve the mechanical properties of materials. Starch and PBS or PBSA blends produce biodegradable plastic sheeting that can be thermo-formed into biscuit trays and other film products. The tensile strength of these blends is somewhat lower than that of polyester alone, but there is little significant further decrease in strength as the starch content increases, at least to a point. With a very high-starch content (>60 percent), however, sheeting can become brittle. To reduce brittleness and increase flexibility, plasticizers are often added.

Starch-Based Superabsorbent Polymers

Superabsorbent polymers (SAPs) are a unique group of materials that can absorb more than a hundred times their weight in liquid. In addition, SAPs do not easily release these fluids, even under pressure. SAPs were first developed in the 1970s by the U.S. Department of Agriculture in the form of starch/acrylonitrile/acrylamide-based polymers called "superslurpers." These products were originally used in agricultural/horticultural markets as hydrogels to help retain moisture in soil during the growing and shipping of crops and other plants (Yamaguchi et al., 1987).

There are two primary types of SAPs—starch-graft polymers and polymers based on cross-linked polyacrylates. Starch-graft polymers are prepared by graft-polymerizing acrylonitrile onto a starch substrate (Smith, 1972; Yamaguchi et al., 1987).

Saponification of a starch-graft polymer with an alkali yields a final product with nitrile, amide, and carboxyl functionalities (Figure 1). The hydrophilic groups on the composite can be adjusted by controlling the amount of sodium hydroxide and reaction time during the saponification process. The collaborative absorbent effect of -$CONH_2$, -COONa, and -COOH groups present in these poly-

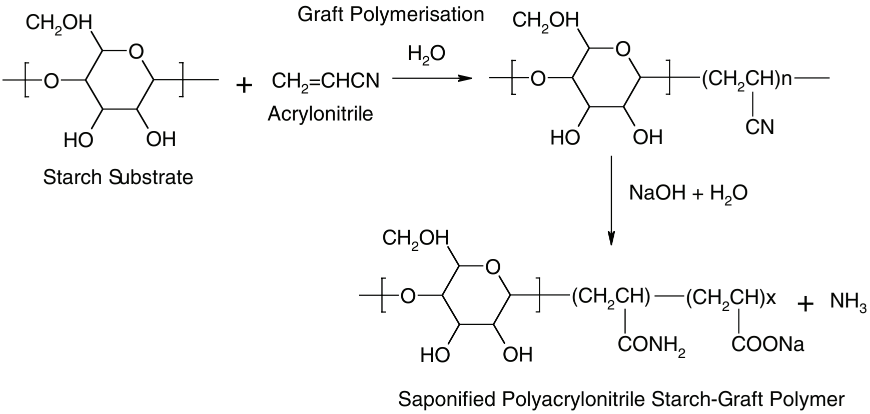


FIGURE 1 Synthesis of a starch-graft polymer.

mers is superior to the absorbance capacity of other polymers containing a single $-CONH_2$, -COONa, or -COOH group (Wu et al., 2003).

Cross-linking during polymerization yields a networked polymer that is not only insoluble in water but is able to absorb and retain water under low load. A typical cross-linking agent is trimethylolpropane triacrylate in concentrations of 0.05 mol percent relative to the monomer. Cross-linking is also possible with ethylene glycol diglycidyl ether, which reacts with carboxyl groups on the polymer molecules to cross-link them.

Global demand for SAPs totaled an estimated 1.05 million tons in 2003, and the demand is expected to increase an average of 3.6 percent per year from 2003 to 2008. Globally, baby diapers account for an estimated 81 percent of SAP demand (Figure 2) (*http://nexant.ecnext.com/coms2/gi_0255-3047/Super-Absorbent-Polymers-SAP.html*). Adult incontinence garments make up the next largest segment of the SAP market (8 percent), followed by other applications (6 percent) and feminine hygiene products (5 percent). The "other applications" category includes de-watering agents for sewage sludge, drying agents for china-clay slurries and pulverized-coal slurries (Farrar et al., 1995), and a wide range of other uses.

Polylactic Acid or Polylactide

Polylactide (PLA) is a biodegradable polymer based on renewable resources with a range of applications similar to polyethylene terephthalate (PET). Uhde Inventa-Fischer, a German-based company, has been developing a PLA process for the last 10 years (*http://www.uhde-inventa-fischer.com*), and Cargill in the

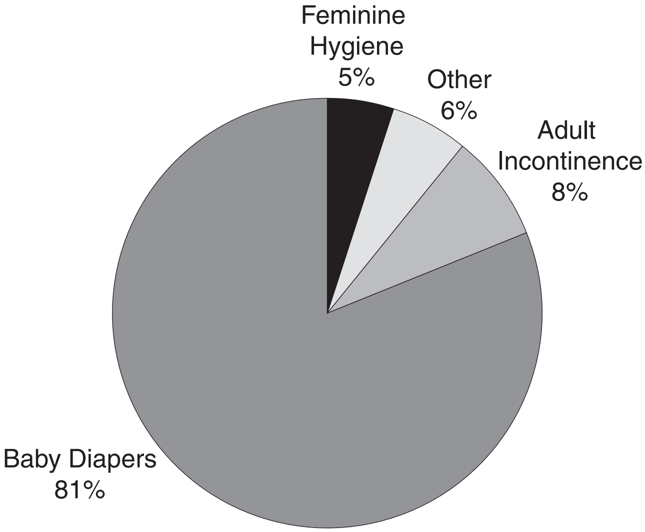


FIGURE 2 The market for superabsorbent polymers in 2003 (total 1.05 million tons).

United States and Teijin in Japan are investing in large-scale production facilities for the manufacture of PLA.

The feedstock for PLA processes is glucose, which can be produced from hydrolysis of starch. Glucose first undergoes fermentation to produce sodium lactate and some impurities, such as proteins, cellular mass, and so on. After a number of purification steps, sodium lactate is transformed into lactic acid, which is then concentrated to remove residual water. This is followed by thermal polymerization (self-condensation), which yields a low-molecular weight PLA prepolymer, which is then thermally depolymerised to yield a distillable cyclic dimer (dilactide, or simply lactide).

Ring-opening polymerization of the purified dilactide with a suitable catalyst, often a tin salt (e.g., stannous octoate), produces high-molecular weight PLA (Figure 3). The final polymer is then granulated and further processed for the desired application. Because of PLA's good mechanical properties, it is used as packaging material (e.g., film, sheet, or bottles), textiles (e.g., filaments or fibers), engineering plastics, and medical polymers (e.g., surgical thread or implants).

1,4:3,6-Dianhydrohexitols-Based Polymers

Incorporating simple carbohydrates directly into polymer structures is difficult because both primary and secondary hydroxyl functionalities in carbo-

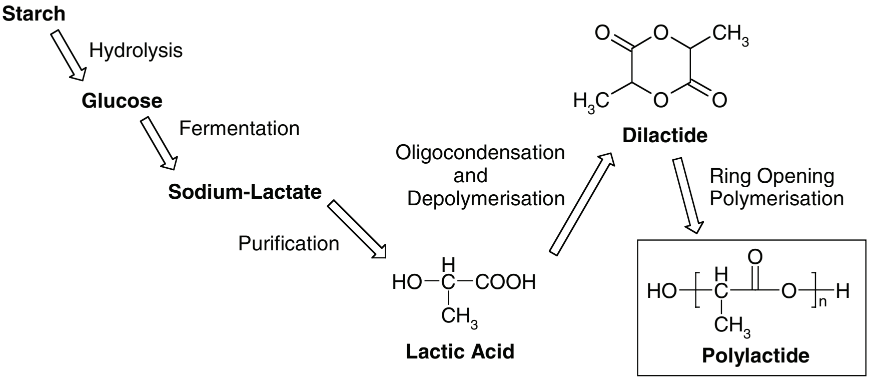


FIGURE 3 The Uhde Inventa-Fischer process of making polylactide from starch. Source: <www.udhe-inventa-fischer.com>.

hydrates interfere with the synthesis of well defined products. The problem can be circumvented by using 1,4:3,6-dianhydrohexitols.

Dianhydrohexitols are well documented by-products of the starch industry obtained by dehydration of D-hexitols, which are made by a simple reduction of hexose sugars (Stross and Hemmer, 1991). About 650,000 tons of dianhydrohexitols are produced annually worldwide (Sheldon, 1993). These chiral biomass-derived products exist as three main isomers (isosorbide, isomannide, and isoidide), depending on the configuration of the two hydroxyl functions (derived from D-glucose, D-mannose, and L-fructose, respectively) (Figure 4). Isosorbide, which is produced from glucose via sorbitol, is the most widely available dianhydrohexitol.

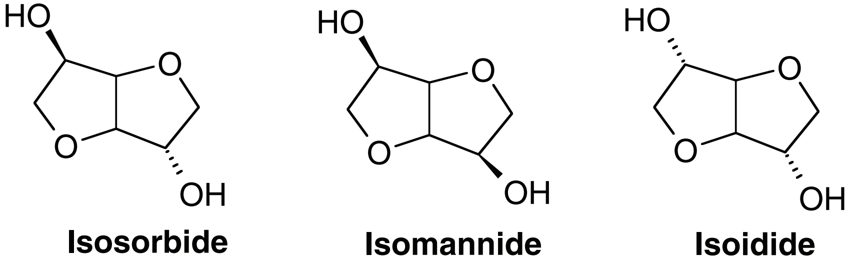


FIGURE 4 The three main isomers of dianhydrohexitols.

Isosorbide

The Iowa Corn Promotion Board supports research on corn-derived products (notably isosorbide) in conjunction with the U.S. Department of Agriculture and U.S. Department of Energy and is partnering with the New Jersey Institute of Technology to identify and assess potential polymer applications for isosorbide. This biodegradable material is classified by the Food and Drug Administration as a "generally recognized as safe," or GRAS, material.

In principle, isosorbide can be incorporated directly into commercial polyesters, such as PET, which is widely used in a variety of food and beverage containers and has a current market of 4.4 billion pounds. Isosorbide, which has a rigid molecular structure, stiffens PET chains and raises the glass-rubber transition temperature (T_g) of PET resin to 85-90°C, depending on the level of incorporation. At that T_g, bottles made from these modified resins can be filled with pasteurized food products while they are still hot or with products that are so viscous they require heating to reduce filling time. Normal, unmodified PET bottles simply become distorted when hot-filled in this way. Preliminary cost estimates show that isosorbide produced by processing of starch is competitive with the cost of petroleum-based building blocks currently used to make plastics (Adelman et al., 2003; Kricheldorf and Gomourachvili, 1997).

Besides modifying PET, isosorbide has many potential uses in the polymer industry and can potentially reduce the amount of petroleum necessary to make plastics. The synthesis of numerous polymers, such as polyethers (Chatti et al., 2006; Kricheldorf and Al Masri, 1995; Majdoub et al., 1994), polyesters (Kricheldorf et al., 2003; Thiem and Lueders, 1984), polyurethanes (Bachmann et al., 1998; Cognet-Georjon et al., 1995; Thiem and Lueders, 1986), and polycarbonates (Kricheldorf and Gomourachvili, 1997), is already being investigated.

A few of the potential applications of dianhydrohexitol-based polymeric products are listed below:

- Isosorbide diglycidyl ether resins could replace bisphenol-A epoxies, which are used to line food and beverage cans. Bisphenol-A is a suspected xenestrogen, and replacing it with a naturally derived, biodegradable diol would have clear consumer advantages.
- Isosorbide and dianhydroiditol polycarbonate resins could be used as molding plastics with high T_g, transparency, and UV-resistance to replace bis-A polycarbonates in CD blanks.
- Chiral separation resins (amorphous, insoluble polyurethanes based on isosorbide) have great potential value in the pharmaceutical and fine-chemicals market for resolving enantiomeric molecules. These isosorbide-based materials exploit the basic chirality of anhydrosugar molecules.
- Isosorbide/isoidide-ester copolymers of controlled stereochemistry could lead to high-performance fibers and engineering resins and the creation of new

classes of controlled-performance, cost-effective polyesters through the control of polymer chirality along the chain.

- Potential applications of polyester resins from isosorbide with terephthalic and isophthalic acids and copolymers with monomers (e.g., L-lactide, caprolactone, etc.) range from environmentally friendly molding resins to drug-release and degradable-suture materials.

COSMETICS AND DRUG-DELIVERY APPLICATIONS

Dimethyl isosorbide (DMI) is a water-white liquid with excellent solvent properties that can be easily synthesized by methylation of isosorbide using dimethyl sulfate and aqueous alkali. DMI can be formulated with standard equipment and requires no special materials, such as flammables, in its production.

Used as a sustainable solvent in skin-care products and drug formulations, DMI facilitates the penetration of active ingredients through the epidermis, enabling targeted delivery of self-tanners, makeup removers, anti-acne treatments, and other transdermal products (*http://www.bulkactives.com/dimethylisosobide.htm*). DMI can transport water-soluble active agents into the skin without recrystallizing them and causing skin irritation. And, most important, it does not promote penetration of the ingredients into the bloodstream.

Materials made with DMI require a lower concentration of aggressive active agents, such as salicylic acid, vitamin C, lactic acid, hydrocortisone, and hyaluronic acid, thus reducing the formulation cost of finished products. Other advantages include longer shelf life (particularly for formulations susceptible to hydrolysis or trans-esterification); miscibility with most organic solvents and non-ionic surfactants; and compatibility with many product forms, including clear gels.

All of these properties make DMI a potential choice for transdermal or topical delivery of a variety of drugs. Preliminary studies indicate that DMI is practically nontoxic. Thus, when used internally, it is predicted that DMI will be metabolized to form isosorbide. DMI has already been used successfully for stable liquid formulation of aspirin (Luzzi and Joseph, 1980), which has been a long-standing problem because of the instability of aspirin in other solvents.

PHARMACEUTICAL USES OF DIANHYDROHEXITOLS

The mono- and dinitrate esters of isosorbide and isomannide, both active NO-releasing drugs, are widely used in the pharmaceutical industry as vasodilators. Figure 5 shows the configuration of isosorbide dinitrate (sold under the brand names Isordil® and Sorbitrate®), which is used as a vasodilator in various formulations that work by different routes for the management of angina pectoris and congestive heart failure (Bidoggia, 1987; Marten et al., 1984). Many pharmaceutical preparations of these compounds are trade products.

Isosorbide and isomannide have also been successfully used as templates for

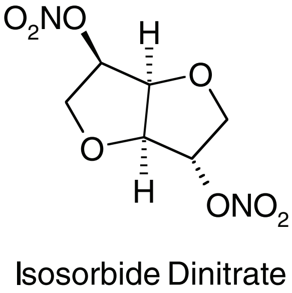


FIGURE 5 Chemical structure of the isosorbide dinitrate molecule.

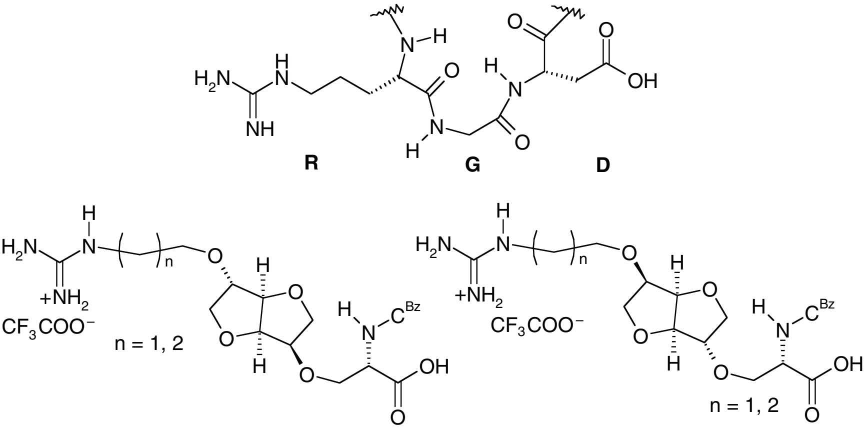


FIGURE 6 RGD-mimetics derived from isosorbide and isomannide.

RGD-mimetics[1] (Figure 6) in the development of integrin antagonists [2](Osterkamp et al., 1999). Because RGD mimetics can inhibit cell attachment and consequently induce apoptosis, they might someday be used as drugs against angiogenesis, inflammation, and cancer mestastasis.

[1] RGD is the single letter code for arginine-glycine-aspartate motif, which can be found in proteins of the extracellular matrix.

[2] Integrins link the intracellular cytoskeleton of cells with the extracellular matrix by recognizing the RGD motif.

CONCLUSION

Corn is a versatile, natural, biodegradable and renewable resource that has many commercial applications. Corn-chemical and biotechnological processes have led to the development of a wide range of products, from polymers to biofuels and vitamins. In the past decade, numerous consumer products based on the results of research have demonstrated the potential of biorenewable sources being used instead of petroleum-based products. With continued work and technological advances, we may soon achieve this potential.

REFERENCES

Adelman, D. J., L. F. Charbonneau, and S. Ung, inventors. E. I. Du Pont de Nemours & Company, assignee. December 2, 2003. Process for Making Poly(ethylene-co-isosorbide) Terephthalate Polymer. U.S. Patent 6,656,577.

Bachmann, F., J. Reimer, M. Ruppenstein, and J. Thiem. 1998. Synthesis of a novel starch-derived AB-type polyurethane. Macromolecular Rapid Communications 19(1):21-26.

Bidoggia, H. 1987. Isosorbide-5-mononitrate and isosorbide dinitrate in the treatment of coronary heart disease: A multi-centre study. Current Medical Research and Opinion 10(9):601-611.

Chatti, S., M. Bortolussi, D. Bogdal, J. C. Blais, and A. Loupy. 2006. Synthesis and properties of new poly(ether-ester)s containing aliphatic diol based on isosorbide effects of the microwave-assisted polycondensation. European Polymer Journal 42(2):410-424.

Cognet-Georjon, E., F. Méchin, and J. P. Pascault. 1995. New polyurethanes based on diphenylmethane diisocyanate and 1,4:3,6-dianhydrosorbitol, 1: Model kinetic studies and characterization of the hard segment. Macromolecular Chemistry and Physics 196(11):3733-3751.

Farrar, D., P. Flesher, M. Skinner, and D. Marshall, inventors. Allied Colloids Limited, assignee. January 24, 1995. Water-Absorbing Polymers. U.S. Patent 5,384,343.

Gomes, M. E., A. S. Ribeiro, P. B. Malafaya, R. L. Reis, and A. M. Cunha. 2001. A new approach based on injection moulding to produce biodegradable starch-based polymeric scaffolds: Morphology, mechanical and degradation behaviour. Biomaterials 22(9):883-889.

Koch, H., and H. Röper. 1988. New industrial products from starch. Starch–Stärke 40(4):121-131.

Kricheldorf, H. R., and M. Al Masri. 1995. New polymer synthesis. LXXXII. Syntheses of poly(ether-sulfone)s from silylated aliphatic diols including chiral monomers. Journal of Polymer Science Part A: Polymer Chemistry 33(15):2667-2671.

Kricheldorf, H. R., and Z. Gomourachvili. 1997. Poly-anhydrides 10: Aliphatic polyesters and poly(ester anhydrides) by polycondensation of silylated aliphatic diols. Macromolecular Chemistry and Physics 198(10):3149-3160.

Kricheldorf, H. R., S. Chatti, G. Schwartz, and R. P. Kruger. 2003. Macrocycles 27: Cyclic aliphatic polyesters of isosorbide. Journal of Polymer Science Part A: Polymer Chemistry 41(21):3414-3424.

Luzzi, L. A., and K. H. Ma Joseph, inventors. Research Corporation, assignee. October 14, 1980. Dimethyl Isosorbide in Liquid Formulation of Aspirin. U.S. Patent 4,228,162.

Majdoub, M., A. Loupy and G. Fleche. 1994. New polyethers and polyesters from isosorbide: Synthesis and characterization. European Polymer Journal 30(12):1431-1437.

Malafaya, P. B., C. Elvira, A. Gallardo, J. San Roman and R. L. Reis. 2001. Porous starch-based drug delivery systems processed by a microwave route. Journal of Biomaterials Science, Polymer Edition 12(11):1227-1241.

Marten, W., M. Weiss, and W. Haase. 1984. Treatment of coronary heart disease with isosorbide mononitrate ('Elantan' 20): A multi-centre study in hospital and general practice. Current Medical Research and Opinion 9(2):96-106.

Osterkamp, F., H. Wehlan, U. Koert, M. Wiesner, P. Raddatz and S. L. Goodman. 1999. Synthesis and biological evaluation of dianhydrohexitol integrin antagonists. Tetrahedron 55(35): 10713-10734.

Reis, R. L., and A. M. Cunha. 1995. Characterization of two biodegradable polymers of potential application within the biomedical field. Journal of Material Science: Materials in Medicine 6(12):786-792.

Reis, R. L., A. M. Cunha, and M. J. Bevis. 1997a. Structure, development and control of injection molded hydroxyapatite reinforced starch/EVOH composites. Advances in Polymer Technology 16(4):263-277.

Reis, R. L., S. C. Mendes, A. M. Cunha, and M. Bevis. 1997b. Processing and in vitro degradation of starch/EVOH thermo-plastic blends. Polymer International 43(4):347-352.

Sheldon, R. A. 1993. Chirotechnology: Industrial Synthesis of Optically Active Compounds. New York: Marcel Dekker.

Smith, T., inventor. Grain Processing Corporation, assignee. May 9, 1972. Water Absorbing Alkali Metal Carboxylate Salts of Starch-Polyacrylonitrile Graft Copolymers. U.S. Patent 3,661,815.

Stross, P., and R. Hemmer. 1991. 1,4:3,6-dianhydrohexitols. Advances in Carbohydrate Chemistry and Biochemistry 49:93-175.

Thiem, J., and H. Lueders. 1984. Synthesis and directed polycondensation of starch-derived anhydroalditol building units. Starch–Stärke 36:170-176.

Thiem, J., and H. Lueders. 1986. Synthesis and properties of polyurethanes derived from diaminodianhydroalditols. Makromolekulare Chemie 187:2775-2785.

Wu, J., Y. Wei, J. Lin, and S. Lin. 2003. Study on starch-graft-acrylamide/mineral powder superabsorbent composite. Polymer 44(21):6513-6520.

Yamaguchi, M., H. Watomoto, and M. Sakamoto. 1987. Super-absorbent polymers from starch-polyacrylonitrile graft copolymers by acid hydrolysis before saponification. Carbohydrate Polymers 7(1):71-82.

Process Review of Lignocellulose Biochemical Conversion to Fuel Ethanol

Bruce S. Dien
National Center for Agricultural Utilization Research
U.S. Department of Agriculture
Peoria, Illinois

The United States is in transition. Each year we use 140 billion gallons of gasoline to fuel our cars. Over 50 percent is imported, and transportation accounts for 50 percent of U.S. greenhouse gas emissions. Increasingly worrisome headlines related to energy security and economic interests and strident warnings from climate scientists all suggest the time has come to reduce oil usage and its related CO_2 emissions.

Currently three routes are available to reduce imported oil and greenhouse gas emissions associated with transportation: conserve, switch to plug-in hybrids, and rely on renewable biofuels. Quite likely all three will be needed to reduce gasoline consumption in a meaningful manner. Coal liquefaction dates back to World War II and can reduce or even eliminate oil imports but will increase, possibly even double, CO_2 emissions.

Renewable fuels are defined as liquid fuels produced from biomass. The concept is simple; plants recycle the CO_2 released from combustion and the plants are in turn harvested and converted to fuels. It is actually more complex because fossil fuels are used to plant, grow, harvest, and process the biomass into fuels. Despite this, lifecycle analysis indicates biofuels, especially from lignocellulose biomass, can greatly reduce CO_2 emissions and very efficiently reduce net gasoline usage (Farrell et al., 2006).

Major sources of biofuels are ethanol and to a much lesser extent biodiesel. Last year 5 billion gallons of ethanol (95 percent from corn) was produced, and production is expected to grow to 10 billion gallons in the next few years

(Westcott, 2007). After that any further growth in biofuel production will need to rely on lignocellulosic feedstocks because of competing demands for corn from food, industrial, and export uses (Westcott, 2007).

Lignocellulose includes agricultural residues, forest industry wastes, and (potentially) perennial energy crops. Agricultural residues include corn stover (e.g., stalks and cobs) and wheat straw. Forest industry wastes include lumber scraps, pulping waste, as well as urban-generated cellulosic wastes. Perennial energy crops include warm season grasses (e.g., switchgrass) and fast growing trees (e.g., loblolly pine and poplar hybrids). Currently no crops are grown for energy, but warm season grasses are grown for forage, and trees, of course, are grown for pulping and to make lumber. Estimated availability of each is in the hundreds-of-million-ton range (Figure 1) and all summed together could theoretically meet 20 percent of our total liquid transportation fuels by 2017 (Perlack et al., 2005).

Recently the U.S. Department of Energy announced that it would help fund six commercialization efforts for converting biomass to biofuels, most of which are related to production of ethanol. While ethanol is the leading candidate for a renewably generated liquid fuel, there are other alternatives, some dating back nearly as far as ethanol. These include butanol produced by acetone-butanol-ethanol (ABE) fermentation of biomass, synthetic gasoline gasification produced by gasifying biomass to syngas followed by Fisher-Tropsch reformation, ethanol

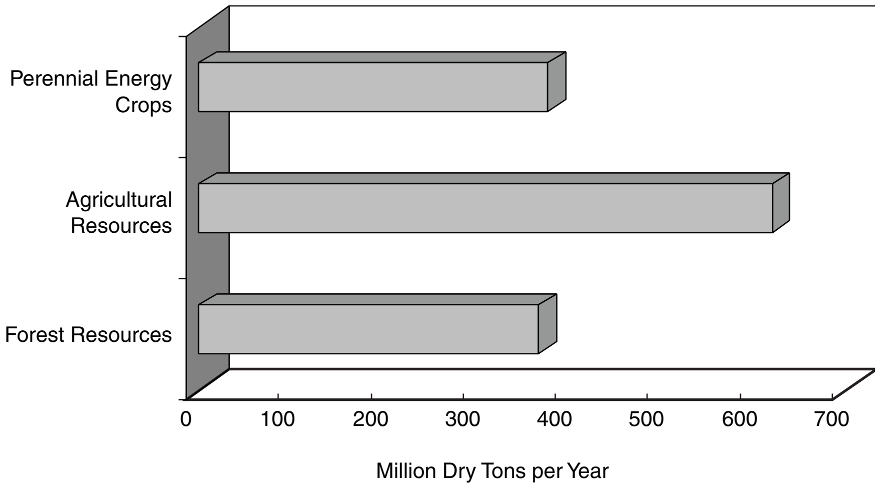


FIGURE 1 U.S. land can supply 1.3 billion tons of biomass for biofuels and still meet other needs. Source: Adapted from Perlack et al., 2005.

produced by fermentation of biomass-derived syngas, and biodiesel produced from algae grown in huge saltwater ponds. Algae are the exception in relying on a new "crop" and are also by far the earliest in development. Each is being actively pursued for commercialization; limited space prevents further discussion of these alternatives.

CHEMICAL COMPOSITION OF BIOMASS AND THEORETICAL ETHANOL YIELDS

Carbohydrates are the only portion of the plant that can be fermented to ethanol. In fibrous biomass, carbohydrates are mostly present in the plant cell walls and are in the form of cellulose and hemicellulose. Cellulose can be converted to glucose and hemicellulose to a mixture of sugars, the composition of which varies with the source of biomass. Herbaceous hemicellulose contains mostly xylose and significant amounts of arabinose and glucoronic acids (Dien et al., 2005). Other minor sugars include galactose and ribose. While glucose is a hexose and has six carbons, arabinose and xylose contain five carbons and are termed "pentoses." The major significance of this is that distillers' yeast (*Saccharomyces*) cannot ferment pentose sugars. This precludes the use of commercial yeast in converting lignocellulose to ethanol.

Plants are approximately 60 percent wt./wt. carbohydrates, of which hemicellulose accounts for about a third (Wiselogel et al., 1996). When fermented to ethanol, one CO_2 is produced for each ethanol, so the theoretical yield for neutral sugars is 0.51 g of ethanol per 1 g of sugar. For herbaceous biomass the theoretical yield of ethanol is 100-110 gal. per dry ton; which compares with 124 gal. per dry ton for corn (for conversion calculator see *www1.eere.energy.gov/biomass/ethanol_yield_calculator.html*). As a practical matter most experts use conversion factors of 60-90 gal. of ethanol per dry ton of biomass, with the lower range having been demonstrated and the upper limit extrapolated from current research.

BIOCHEMICAL CONVERSION

Many processes have been conceptualized for converting fibrous biomass to ethanol. All have common aspects, so I will first discuss a conceptual design for a dilute acid pretreatment (Figure 2) (Aden et al., 2002). Dry biomass that arrives at the facility is first cleaned and milled. The biomass is mixed with a dilute mineral acid solution to a solids consistency of 30-40 percent wt./wt. The biomass is conveyed to a steam explosion reactor where it is heated to 180-220ºC for 0.5-5.0 minutes before being quickly (and explosively) cooled by the sudden release of the reactor pressure (Schell et al., 2003). Treating in this manner physically reduces particle sizes and changes the consistency of the product from a damp fiber to sludge. It also breaks down the plant cell wall, removing the hemicellulose to the syrup, displacing the lignin, and swelling the tightly (highly crystalline)

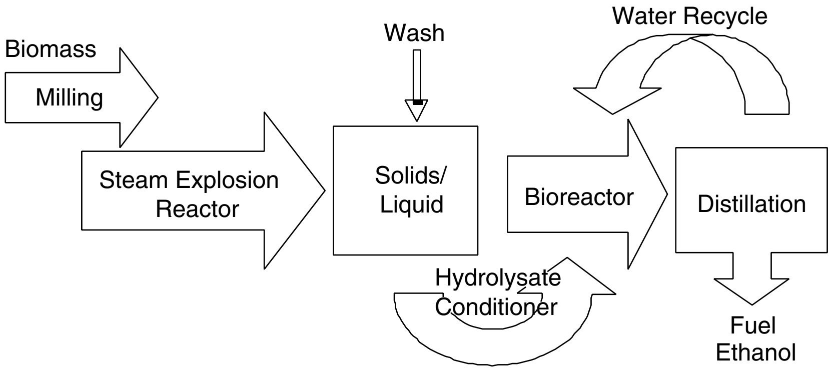


FIGURE 2 Conceptual flow diagram of process for converting lignocellulose to ethanol (see text for details).

arrayed cellulose fibers. To recap, following the steam explosion the solids consist largely of lignin and cellulose, and the syrup contains most of the hemicellulose carbohydrates, water extractables, a little glucose, and minor amounts of released lignin products.

Following pretreatment the solids are recovered and washed, possibly using a press. The syrup and wash water are mixed and the residual sulfuric acid neutralized by adding lime. Pretreatment produces a wide variety of soluble side products, some of which are quite toxic to microbes. Therefore, the syrup often needs to be conditioned to reduce its toxicity prior to fermentation. Following this the syrup is remixed with the solids. At this point the biomass is too thick to ferment directly, so enzymes are added to thin the slurry and to begin saccharifying the cellulose to glucose. For current commercial enzymes the temperature is held at 50-55ºC for 18-24 hours. Next, the biocatalyst is added, which begins to ferment the released sugars to ethanol. The fermentation temperature will generally be lower than 50ºC, but its specific set point will depend upon the choice of microorganism. At the same time the fermentation is occurring the enzymes continue to release sugars for fermentation. As mentioned above, a special microbe needs to be used that is capable of fermenting the pentose sugars in addition to the glucose. A number of microbes are now available that ferment either xylose or both xylose and L-arabinose in addition to glucose. The fermentation could theoretically last up to 7 days, but is usually ended after 3 days. The ethanol is stripped out of the beer, distilled, and finished by passing through a molecular sieve to remove the last of the water. The solids (e.g., largely lignin) are recovered from the stillage by centrifugation and combusted to generate steam for the overall process. The recovered liquid is (hopefully) treated and recycled in the process.

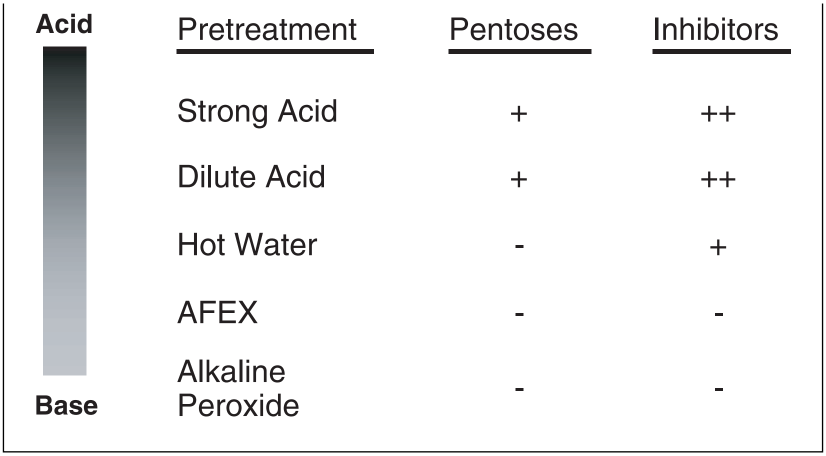
There are many process variations for converting biomass to ethanol. In a simultaneous saccharification and fermentation (SSF), the microbe is added with the enzyme (Takagi et al., 1977; Emert and Katzen, 1980). The key advantage of co-adding them is that the microbe ferments glucose immediately to ethanol, thereby avoiding any buildup of glucose in the culture. Maintaining a low glucose concentration has the advantage of avoiding end-product inhibition of the enzyme and helps minimize the risk for contamination. Fermentations are run as open processes, and contamination is always a concern. Detailed technoeconomic models for SSF of poplar wood and corn stover have been developed by the National Renewable Energy Laboratory (Aden et al., 2002; Wooley et al., 1999). Enzymes are a major cost of processing biomass to ethanol. If microbes were used that produce some of their own enzymes, it would be possible to eliminate a large expense item. Using microbes that produce their own carbohydrolytic enzymes is the central theme of consolidated bioprocessing (CBP) (Den Haan et al., 2007; Katahira et al., 2006; Lynd et al., 2005). An alternative to SSF would be to completely hydrolyze the carbohydrates and remove the solids prior to fermentation, which is referred to as SHF. This is the process used by Iogen Corp. for its demonstration plant (Tolan, 1999). It has the advantages of making the fermentation faster because it is not enzyme limited, eliminates solids from the bioreactor, supplies a cleaner burning lignin, and may (theoretically) allow for recovering and recycling of enzymes. However, end-product inhibition of the cellulases is a major concern.

UNIT OPERATIONS

The major processing steps for converting biomass to ethanol are pretreatment, enzymatic saccharification, fermentation, and recovery. Lignocellulose contains primarily structural carbohydrates, which are highly resistant to enzymatic conversion to monosaccharides. Thermochemical pretreatment is needed to deconstruct the cell wall structure, allowing enzymes access to the carbohydrate polymers. The cell wall has been compared with reinforced concrete, where hemicellulose is the concrete, lignin the hydrophobic sealant, and cellulose microfibrils are the reinforcing bars (Bidlack et al., 1992). Specifically, pretreatment is needed to reduce particle size, dissolve the xylan, displace the lignin, and create broken ends in and swell the cellulose microfibers. There are numerous pretreatments available, some of which are summarized in Figure 3 (reviews: Dien et al., 2005; Mosier et al., 2005).

There are three major categories of enzymes for converting pretreated biomass into fermentable sugars: cellulases, xylanases along with auxiliary enzymes for debranching xylan, and ligninases. A list of these enzymes is presented in Table 1.

Cellulases are by far the most important, because they are used to convert



Pretreatment	Pentoses	Inhibitors
Strong Acid	+	++
Dilute Acid	+	++
Hot Water	-	+
AFEX	-	-
Alkaline Peroxide	-	-

FIGURE 3 Selected pretreatments for lignocellulose. Strong acid and dilute acid have the advantage of producing monosaccharides while the neutral/alkaline pretreatments only solublized xylan. However, the acidic pretreatments also produce more inhibitors that are problematic for fermentation.

TABLE 1 Biomass-Related Enzymes

Cellulases
endo-1,4-β-D-glucanase (EC-3.2.1.4), exo-1,4-β-glucanase (exocellobiohydrolase, EC-3.2.1.91) and β-D-glucosidase (β-D-glucoside glucanhydrolase, EC-3.2.1.21)

Xylanases
endo-1,4-β-D-xylanase, EC-3.2.1.8), β-xylosidase (EC-3.2.1.37)

Xylan Debranching Enzymes
α-L-arabinofuranosidase (EC-3.2.1.55), β-glucuronidase (EC-3.2.1.31), acetylxylan esterase (EC-3.1.1.72), feruloyl esterase (EC-3.1.1.73), p-coumaroyl esterase (EC-3.1.173), others

Ligninases
Lignin peroxidase (LiP, EC-1.11.1.7), manganese peroxidase (MnP, EC-1.11.1.13) and laccase (EC-1.10.3.2)

cellulose into glucose (Zhang and Lynd, 2004). Dilute acid pretreatment converts the hemicellulose carbohydrates directly to monosaccharides and therefore only requires cellulase blends; commercial blends containing xylanase activity can in some cases improve conversion efficiency. Other pretreatments will solubilize the xylan but require additional enzymes (hemicellulases) (Saha, 2003) to saccharify it completely to fermentable sugars. Ligninases have not been widely applied to biomass bioconversion as yet.

The bioethanol industry is dependent upon *S. cerevisiae* for fermentation of glucose. Unfortunately *S. cerevisiae* does not ferment pentose sugars, and these sugars are too abundant in lignocellulose to ignore. The bacterium *Zymomonas mobilis* also selectively produces ethanol and has been offered as a substitute for *S. cerevisiae*. But like *Saccharomyces*, *Z. mobilis* does not ferment pentoses. Therefore, researchers have had to depend upon molecular methods for developing new biocatalysts for converting pentoses (and especially xylose) into ethanol. Two approaches have been taken to solve this problem: (1) engineering *S. cerevisiae* and *Z. mobilis* to ferment xylose and in the case of *Z. mobilis* also arabinose or (2) engineering Gram-negative bacteria that use a wide variety of sugars to selectively produce ethanol under anaerobic conditions. There has been considerable work on using the latter strategy to develop Gram-positive bacteria, but while progress is being made, full success has been elusive. Table 2 reviews the microorganisms available for fermenting xylose (Dien et al., 2003; Jeffries and Jin, 2004).

TABLE 2 Comparison of Batch Fermentations with Xylose for Ethanologenic Strains

Strain	Host	Xylose (g/l)	Max. EtOH (g/l)	EtOH Eff[a] (%)	Reference
E. coli	K011[b]	90	41.0	89	Yomano et al. (1998)
	FBR5	95	41.5	90	Dien et al. (2000)
	LY01	140	63.2	88	Yomano et al. (1998)
K. oxytoca	*M5A1(pLOI555)*	100	46.0	95	Ohta et al. (1991)
Z. mobilis	*CP4:pZB5*	60	23.0	94	Lawford et al. (1999)
S. cerevisiae	*TMB 3400*[c]	50	13.3	67	Karhumaa et al. (2007)
	RWB 218[c]	20	8.4	85	Kuyper et al. (2005)
	RE700A(pKDR)[d]	45	19.0	74	Sedlak and Ho (2004)

[a] Ethanol efficiency: % yield of theoretical based upon 51 g ethanol per 100 g of xylose present.
[b] On 140 g/l xylose, strain K011 produced 59.5 g/l ethanol in 120 hr (Yomano et al., 1998).
[c] Cultured on mineral medium.
[d] Cultured on rich medium, estimated from Fig. 8 of Sedlak and Ho (2004).

FUTURE TRENDS

Ethanol production from wood dates back before World War II. Modern technology has allowed the potential of much higher yields with a smaller environmental footprint. However, challenges remain, and further research will be needed to make lingocellulosic ethanol cost-competitive. Efforts will continue toward producing more robust pentose-fermenting microorganisms with higher productivity and more efficient, less-expensive enzymes. More work will also be directed at understanding the cell wall and the sources of biomass recalcitrance. Simultaneously there should be increased efforts to engineer plants for easier conversion to sugars (e.g., less lignin or altered cell wall structures) or that produce some of the enzymes needed for breaking down cell walls in situ. The U.S. Department of Energy has recently announced that it will fund three institutions for 5 years with this goal in mind (*http://www.genomicsgtl.energy.gov/centers/*).

In an attempt to jump start a lignocellulose ethanol industry the U.S. Department of Energy has announced that it will fund six commercial efforts for up to a total of $385 million (news release at *http://www.energy.gov/news/4827.htm*). Abengoa Bioenergy Biomass, Poet Companies, and Iogen Biorefinery will focus on biochemical conversion of herbaceous biomasses, including corn cobs and fiber, switchgrass, and wheat straw. Alico Inc. will use a hybrid process whereby the biomass is converted to syngas, and the syngas is then fermented to ethanol. Range Fuels will apply a strictly thermochemical approach. Blue Fire Ethanol Inc. will utilize the Arkenol process, which relies on strong acid hydrolysis (*http://www.bluefireethanol.com*). It is hoped by those working in the field that the combination of strong political and industrial interests will help to unlock lignocellulose as a commercially successful feedstock for ethanol.

REFERENCES

Aden, A., M. Ruth, K. Ibsen, J. Jerchura, K. Neeves, J. Sheehan, B. Wallace, L. Montague, A. Slayton, and J. Lukas. 2002. Lignocellulosic Biomass to Ethanol Process Design and Economics Utilizing Co-Current Dilute Acid Prehydrolysis and Enzymatic Hydrolysis for Corn Stover. Rep. NREL/TP-510-32438. Golden, CO: National Renewable Energy Laboratory.

Bidlack, J., M. Malone, and R. Benson. 1992. Molecular structure and component integration of secondary cell walls in plants. Proceedings of the Oklahoma Academy of Sciences 72:51-56.

Den Haan, R., J. E. McBride, D. C. L. Grange, L. R. Lynd, and W. H. Van Zyl. 2007. Functional expression of cellobiohydrolases in *Saccharomyces cerevisiae* towards one-step conversion of cellulose to ethanol. Enzyme and Microbial Technology 40:1291-1299.

Dien, B. S., M. A. Cotta, and T. W. Jeffries. 2003. Bacteria engineered for fuel ethanol production: Current status. Applied Microbiology and Biotechnology 63:258-266.

Dien, B. S., L. Iten, and C. D. Skory. 2005. Converting herbaceous energy crops to bioethanol: A review with emphasis on pretreatment processes. Pp. 1-11 in Handbook of Industrial Biocatalysis. Boca Raton, FL: Taylor and Francis Group.

Dien, B. S., N. Nichols, P. J. O'Bryan, and R. J. Bothast. 2000. Development of new ethanologenic escherichia coli strains for fermentation of lignocellulosic biomass. Applied Biochemistry and Biotechnology 84(6):181-196.

Emert, G. H., and R. Katzen. 1980. Gulf's cellulose-to-ethanol process. CHEMTECH 10:610-615.

Farrell, A. E., R. J. Plevin, B. T. Turner, A. D. Jones, M. O'Hare, and D. M. Kammen. 2006. Ethanol can contribute to energy and environmental goals. Science 311:506-508.

Jeffries, T. W., and Y. S. Jin. 2004. Metabolic engineering for improved fermentation of pentoses by yeasts. Applied Microbiology and Biotechnology 63:495-509.

Karhumaa, K., R. Sanchez, L. B. Hahn-Hagerdal, and M. F. Gorwa-Grauslund. 2007. Comparison of the xylose reductase-xylitol dehydrogenase and the xylose isomerase pathways for xylose fermentation by recombinant *Saccharomyces cerevisiae*. Microbial Cell Factories 6:5.

Katahira, S., A. Mizuike, H. Fukuda, and A. Kondo. 2006. Ethanol fermentation from lignocellulosic hydrolysate by a recombinant xylose- and cellooligosaccharide-assimilating yeast strain. Applied Microbiology and Biotechnology 72:1136-1143.

Kuyper, M., M. P. Hartog, M. J. Toirkens, M. J. H. Almering, A. A. Winkler, J. P. van Dijken, and J. T. Pronk. 2005. Metabolic engineering of a xylose-isomerase-expressing *Saccharomyces cerevisiae* strain for rapid anaerobic xylose fermentation. FEMS Yeast Research 5:399-409.

Lawford, H. G., J. D. Rousseau, A. Mohagheghi, J. D. McMillan. 1999. Fermentation performance characteristics of a prehydrolyzateadapted xylose-fermenting recombinant *Zymomonas* in batch and continuous fermentations. Applied Biochemistry and Biotechnology 77:191-204.

Lynd, L. R., W. H. Van Zyl, J. E. McBride, and M. Laser. 2005. Consolidated bioprocessing of cellulosic biomass: An update. Current Opinion in Biotechnology 16:577-583.

Mosier, N., C. Wyman, B. Dale, R. Elander, Y. Lee, M. Holtzapple, and M. Ladisch. 2005. Features of promising technologies for pretreatment of lignocellulosic biomass. Bioresource Technology 96:673-686.

Ohta, K., D. S. Beall, J. P. Mejia, K. T. Shanmugam, and L. O. Ingram. 1991. Metabolic engineering of klebsiella oxytoca M5A1 for ethanol production from xylose and glucose. Applied and Environmental Microbiology 57:2810-2815.

Perlack, R. D., L. L. Wright, A. F. Turhollow, R. L. Graham, B. J. Strokes, and D. C. Erbach. 2005. Biomass as Feedstock for a Bioenergy and Bioproducts Industry: The Technical Feasibility of a Billion-Ton Annual Supply. Report DOE/GO-102995-2135. Washington, DC: U.S. Department of Energy and U.S. Department of Agriculture.

Saha, B. C. 2003. Hemicellulose conversion. Journal of Industrial Microbiology and Biotechnology 30:279-291.

Schell, D. J., J. Farmer, M. Newman, and J. D. McMillan. 2003. Dilute-sulfuric acid pretreatment of corn stover in pilot-scale reactor: Investigation of yields, kinetics, and enzymatic digestibilities of solids. Applied Biochemistry and Biotechnology - Part A Enzyme Engineering and Biotechnology 108(1-3):69-86.

Sedlak, M., and N. W. Y. Ho. 2004. Production of ethanol from cellulosic biomass hydrolysates using genetically engineered *Saccharomyces* yeast capable of cofermenting glucose and xylose. Applied Biochemistry and Biotechnology 113-116:403-416.

Takagi, M., S. Abe, G. H. Suzuki, G. H. Emert, and N. Yata. 1977. A method for production of alcohol direct from cellulose using cellulase and yeast. Pp. 55-571 in Proceedings of the Bioconversion Symposium. New Delhi: Indian Institute of Technology.

Tolan, J. S. 1999. Alcohol production from cellulosic biomass: The iogen process, a model system in operation. Pp. 117-127 in The Alcohol Textbook Third Edition. Nottingham, UK: Nottingham University Press.

Westcott, P. C. 2007. Ethanol Expansion in the United States. Rep. FDS-07D-01. Washington, DC: U.S. Department of Agriculture.

Wiselogel, A., S. Tyson, and D. Johnson. 1996. Biomass feedstock resources and composition. Pp. 105-119 in Handbook on Bioethanol: Production and Utilization. Washington, DC: Taylor & Francis Inc.

Wooley, R., M. Ruth, J. Sheehan, K. Ibsen, H. Majdeski, and A. Galvez. 1999. Lignocellulosic Biomass to Ethanol Process Design and Economics Utilizing Co-current Dilute Acid Prehydrolysis and Enzymatic Hydrolysis Current and Future Scenarios. Rep. TP-580-26157. Golden, CO: National Renewable Energy Lab.

Yomano, L. P., S. W. York, and L. O. Ingram. 1998. Isolation and characterization of ethanol-tolerant mutants of *Escherichia coli* KO11 for fuel ethanol production. Journal of Industrial Microbiology and Biotechnology 20:132-138.

Zhang, Y. H. P., and L. R. Lynd. 2004. Toward an aggregated understanding of enzymatic hydrolysis of cellulose: Noncomplexed cellulase systems. Biotechnology and Bioengineering 88:797-824.

Sustainable Biorefineries

Carina Maria Alles and Robin Jenkins
DuPont Engineering Research & Technology
Wilmington, Delaware

SUSTAINABLE BIOFUELS

Opportunities and Challenges

As the global demand for energy continues to rise, biofuels hold the promise of providing a renewable alternative to fossil fuels. High oil prices and the threat of negative climate change impacts have sparked a surge in research and business activities. National security concerns and the desire to increase farm incomes have prompted governments around the world to enact policies in favor of biofuels. Developing countries hope to increase access to inexpensive fuel in isolated areas. Biofuels are also seen as an option to reduce air pollution in urban areas. However, the current boom has sparked controversy as well. Opponents accuse biofuels of requiring more energy inputs and causing more greenhouse gas emissions than their fossil counterparts. Many concerns center on possible environmental degradation: erosion, deterioration of soil health, depletion of aquifers, and losses in biodiversity. Large-scale biofuels production is often regarded as a threat to food production and conservation efforts (Hill et al., 2006; Worldwatch Institute, 2006).

Sustainability assessments aimed at quantifying the economic, environmental, and societal impacts can help move the often heated debates over biofuels to a more factual level. Science-based methods like life cycle assessment (LCA), a

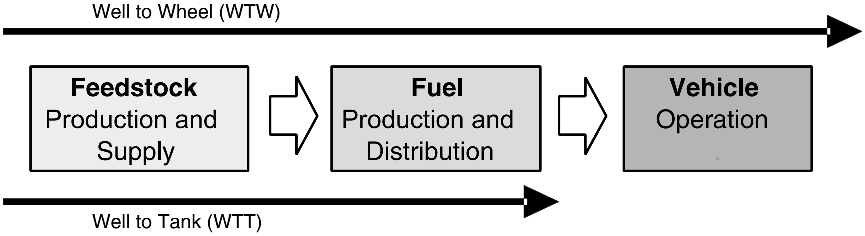

Well to Wheel (WTW)
Feedstock Production and Supply
Fuel Production and Distribution
Vehicle Operation
Well to Tank (WTT)

FIGURE 1 Fuel value chain. Source: DuPont.

holistic approach to quantify environmental impacts throughout the value chain of a product (ISO, 2006), have gained acceptance among decision makers.

Biofuels Sustainability in Well-to-Wheel Perspective

To assess the sustainability of (bio)fuels all stages of the fuel value chain have to be considered (Figure 1): feedstock production and supply, fuel production and distribution, and vehicle operation. An assessment of the entire value chain (i.e., well-to-wheel [WTW]) allows a fair comparison of different fuels, whereas a well-to-tank (WTT) or cradle-to-gate (CTG) (here: extraction of raw materials from the ground to fuel production) approach is adequate to compare different technologies to make the same fuel.

CASE STUDY: BIOREFINERY TECHNOLOGY DEVELOPMENT

The Integrated Corn-based Biorefinery

The DuPont integrated corn-based biorefinery (ICBR) process will use innovative technology to convert corn grain and stover into fermentable sugars for the parallel production of value-added chemicals and fuel ethanol (DOE, 2006). Current R&D efforts in the ICBR program focus on optimized process design for the conversion of lignocellulosic biomass to ethanol. From the beginning, economic evaluation and LCA have been used side by side to guide researchers to the most sustainable process alternative (Figure 2).

In the ICBR program, stakeholder engagement was critical to setting relevant sustainability goals. Stakeholder interests are represented by an external advisory panel of subject matter experts from government agencies, academia, industry, and nongovernmental organizations (NGOs), who helped prioritize topics, translate them into quantifiable metrics, and formulate specific targets for the environmental performance of the ICBR technology. The panel's independent, critical review of life cycle methodology and results has been invaluable to the ICBR program.

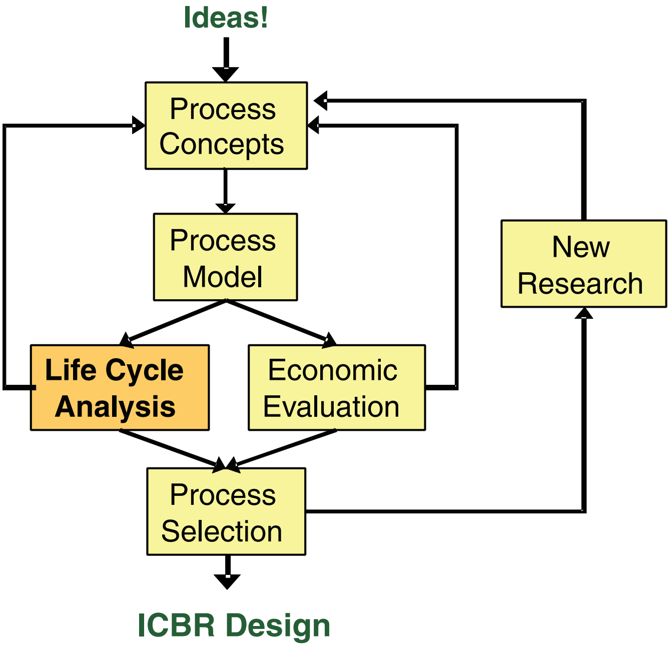


FIGURE 2 Integration of LCA into process development. Source: DuPont.

ICBR Life Cycle Analysis

Life cycle assessment sheds light on relative or directional changes in sustainability resulting from ICBR technology choices. The LCA model follows all natural resources extracted from the environment and all material releases to the environment that cross the system boundaries shown as the dotted line in Figure 3. Many of these flows are then aggregated into impact assessments (e.g., fossil energy consumption, which includes petroleum, natural gas, coal, and lignite use) or greenhouse gas (GHG) emissions, where gases like anthropogenic CO_2, CH_4, and N_2O are weighted by their global warming potential.

A process model of ethanol production from corn stover in an ICBR facility forms the core of the LCA model. The environmental impacts of material and energy inputs to the biorefinery are tracked back to ground, using LCA databases and publicly available information to describe upstream processes. Michigan State University provided a rigorous LCA model of corn farming, including agrochemicals manufacture and the production of fuels used in corn farming (Kim et al., unpublished manuscript). Most ICBR designs assume that nonfermentable

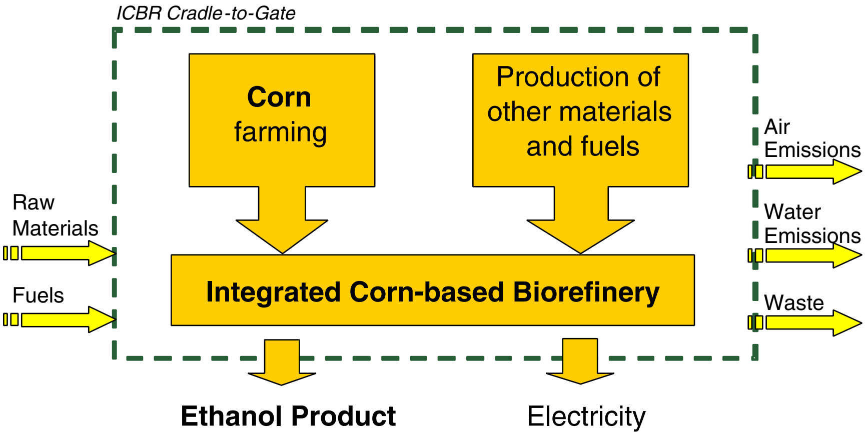


FIGURE 3 Cradle-to-gate LCA model of the ICBR. Source: DuPont.

biomass provides fuel for an onsite cogeneration facility, with the option to sell excess electricity to the local grid. The LCA credit for electricity sales covers the environmental burden of electricity generation all the way back to primary energy sources (Kim and Dale, 2005).

As the ICBR process development progressed the LCA model was continuously updated. Multiple scenario and sensitivity analyses helped identify favorable design options and optimized process parameter settings.

Comparison versus Benchmarks

Eventually ICBR technology will have to be competitive against other technologies to produce ethanol from corn. The LCA model has also been used to benchmark the environmental performance of the ICBR versus the incumbent technology (i.e., ethanol from corn grain) (Graboski, 2002) and potential alternative routes to produce ethanol from corn stover (Sheehan et al., 2004). Publicly available benchmark data were carefully reviewed when the environmental targets for the ICBR were defined. When necessary, critical assumptions and data sources were aligned to ensure a fair comparison.

Results for cradle-to-gate fossil energy consumption, a metric commonly used to address the energy balance of biofuels, are shown in Figure 4. In all three scenarios the CTG fossil energy footprint to produce a gallon of bioethanol is well below the lower heat value (LHV) of the fuel (i.e., the energy ultimately delivered to a vehicle). The ICBR compares favorably against both benchmarks.

Greenhouse gas emissions have emerged as a priority metric in policy making. To benchmark biofuels against the incumbent fossil fuel, differences in the energy content of the fuels need to be considered. Figure 5 shows well-to-wheel greenhouse gas emissions normalized to an equal amount of fuel energy for U.S. gasoline and the three biofuel alternatives discussed above.

All three biofuels cause fewer GHG emissions along the fuel value chain than gasoline. Cellulosic ethanol offers significant advantages in GHG reductions, since corn stover has a smaller GHG footprint than corn grain (Kim et al., unpublished manuscript) and provides renewable fuel to the ethanol production facility, whereas conventional dry mills rely on natural gas or coal (Graboski, 2002). The ICBR ethanol presents the largest GHG reduction potential of all cases shown, thanks to advanced pretreatment and fermentation technologies and optimized process integration. A breakthrough in stover-to-ethanol technology could pave the way for ethanol from other lignocellulosic feedstocks, such as perennial grasses and agricultural or forestry residues.

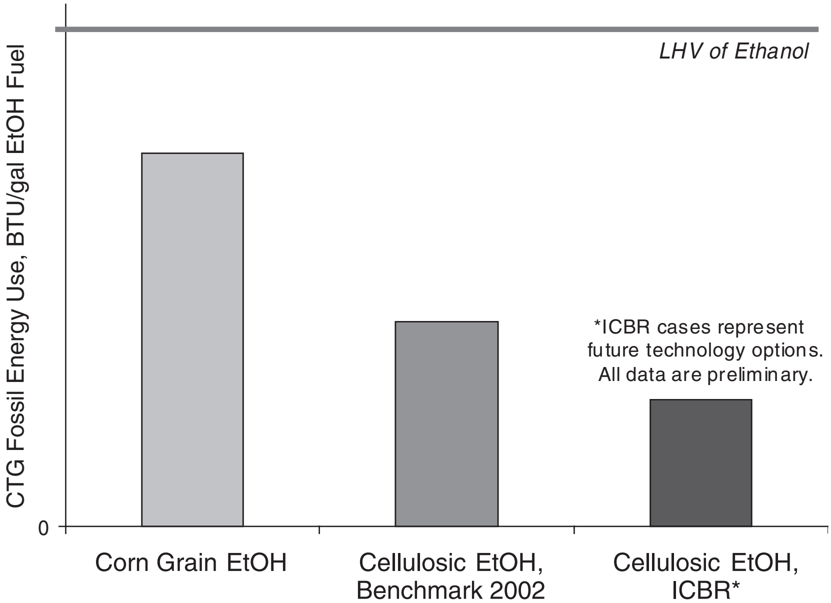


FIGURE 4 Cradle-to-gate fossil energy use of selected biofuels. Source: DuPont.

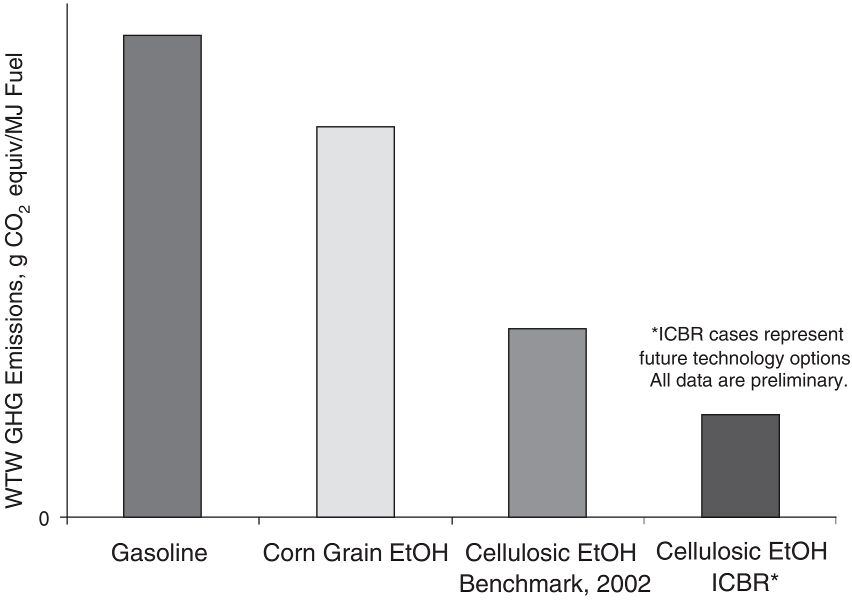


FIGURE 5 Well-to-wheel greenhouse gas emissions of selected fuels. Source: DuPont.

ATTRIBUTES OF SUSTAINABLE BIOREFINERIES

Fundamental learnings from the in-depth analysis of a broad spectrum of ICBR technology options can be applied to biorefineries in general.

Feedstock Selection

The cradle-to-refinery-gate footprint of the primary biorefinery feedstock is a significant contributor to the overall footprint of a biofuel. Generally, lignocellulosic biomass has a lower environmental burden than food crops. In LCA terms, waste materials from industrial processes or municipal collection would come free of burden if using them in biorefinery processing would result in the avoidance of disposing of these materials. Similarly the footprint of processing chemicals used at the biorefinery needs to be considered, especially in the pretreatment of cellulosic feedstocks.

Biorefinery Process Efficiency

As a rule of thumb, cost-efficient measures to improve process efficiency will have a positive impact on both the economic and the environmental performance of biorefineries. The common engineering goal to maximize product yield usually translates into reductions in feedstock footprint, waste generation, and energy consumption. Increasing the effective concentration of solids, intermediates, and products in aqueous process streams not only reduces the amount of water taken from the watershed but also reduces the burden on separation steps. Measures to lower the energy consumption within biorefinery battery limits (e.g., by heat integration or equipment design) have a positive impact on the overall biofuel energy balance, whether they reduce the need to bring supplemental fuel to the biorefinery or enable additional energy exports.

Industrial Ecology of Biorefineries

Typically the biofuel is not the only material output of a biorefinery. If more outputs find their ways into beneficial uses, then the economic and environmental burdens of the biorefinery operation can be allocated among a wider range of coproducts, effectively reducing both the production cost and the cradle-to-gate footprint of the biofuel. For example, lignin and other nonfermentable components of biomass feedstocks need not be disposed as wastes, but can be used as fuels to generate thermal or electrical energy onsite or offsite. Other potential coproducts include animal feed, fertilizer, and intermediates for specialty chemicals.

The colocation of biorefineries with other facilities processing agricultural and forestry feedstocks, livestock feedlots, or power plants in industrial parks will facilitate the beneficial exchange of material and energy flows to a great extent.

REFLECTIONS: SUSTAINABILIY IN TECHNOLOGY DEVELOPMENT

In the development of sustainable fuels, materials, and services for the future, researchers are challenged to find innovative technical solutions without losing sight of the economic, societal, and environmental impacts of their work. The ICBR program demonstrates how the early integration of sustainability analysis into the creative process of technology development enables a holistic approach to research guidance, where economic and environmental metrics are used alongside product performance standards to define and monitor success.

ACKNOWLEDGMENTS

Sustainability analysis in the ICBR program (DOE 1435-04-03-CA-70224) was a team effort, involving partners from Michigan State University, the National Renewable Energy Laboratory, Diversa Corporation, and DuPont. Interactions with researchers and leadership have been very rewarding. The program also benefited immensely from the perspectives of external stakeholders, in particular, members of the ICBR LCA advisory panel.

REFERENCES

DOE (U.S. Department of Energy). 2006. Biomass program—Integrated corn-based bio-refinery. Online. Available at <http://www1.eere.energy.gov/biomass/pdfs/corn-based_biorefinery.pdf>. Accessed July 2007.

Graboski, M. S. 2002. Fossil energy use in the manufacture of corn ethanol. National Corn Growers Association. Online. Available at <http://www.ncga.com/ethanol/pdfs/energy_balance_report_final_R1.PDF>. Accessed July 2007.

Hill, J., E. Nelson, D. Tilman, S. Polasky, and D. Tiffany. 2006. Environmental, economic, and energetic costs and benefits of biodiesel and ethanol fuels. Proceedings of the National Academy of Sciences 103:11206-11210.

ISO (International Organization for Standardization). 2006. Environmental management—Life cycle assessment—Principles and framework. ISO 14040 Ed. 2. Geneva, Switzerland.

Kim, S., and B. Dale. 2005. Life cycle inventory information of the United States electricity system. International Journal of Life Cycle Assessment 10:294-304.

Kim, S., B. Dale, and R. Jenkins. Life Cycle Assessment of Corn Grain and Corn Stover. Unpublished manuscript.

Sheehan, J., A. Aden, K. Paustian, K. Killian, J. Brenner, M. Walsh, and R. Nelson. 2004. The energy and environmental aspects of using corn stover for fuel ethanol. Journal of Industrial Ecology 7(3-4):117-146.

Worldwatch Institute. 2006. Biofuels for transportation: Global potential and implications for sustainable agriculture and energy in the 21st century. Online. Available at <http://www.worldwatch.org/node/4078>. Accessed July 2007.

ADDITIONAL READING

Argonne National Laboratory: Transportation Technology R&D Center. 2007. GREET is "gold standard" for well-to-wheel analyses of vehicle/fuel systems. Online. Available at <http://www.transportation.anl.gov/software/GREET/publications.html>. Accessed July 2007.

DOE (U.S. Department of Energy). 2007. Ethanol. Online. Available at <http://www1.eere.energy.gov/biomass/ethanol.html>. Accessed July 2007. *Note: Contains many further links on biofuels.*

Ecole Polytechnique Federale de Lausanne (EPFL). 2007. Roundtable on sustainable biofuels announces first draft principles for stakeholder discussion. Online. Available at <http://Energy-Center.epfl.ch/Biofuels>. Accessed July 2007. *Note: Global stakeholder discussion, access via website; website also contains further links on emerging biofuels policy standards worldwide.*

EPA (U.S. Environmental Protection Agency). 2007. Life-cycle assessment (LCA). Online. Available at <http://www.epa.gov/ORD/NRMRL/lcaccess/>. Accessed July 2007. *Note: Contains useful background information and many further links on LCA, including "LCA 101."*

Farrell, A. E., R. J. Plevin, B. T. Turner, A. D. Jones, M. O'Hare, and D. M. Kammen. 2006. Ethanol can contribute to energy and environmental goals. Science 27(311):506-508. *Note: Contains a meta-study of recent ethanol LCAs.*

National Resources Defense Council. 2004. Growing energy: How biofuels can help end America's oil dependence. Online. Available at <http://www.nrdc.org/air/transportation/biofuels/contents.asp>. Accessed July 2007.

National Resources Defense Council. 2007. Getting biofuels right: Eight steps for reaping real environmental benefits from biofuels. Online. Available at <http://www.nrdc.org/air/transportation/biofuels/contents.asp>. Accessed July 2007.

Shapouri, H. 2004. The 2001 net energy balance of corn ethanol. Paper presented at the Corn Utilization and Technology Conference, Indianapolis, IN, June. Online. Available at <http://www.usda.gov/oce/reports/energy/index.htm>. Accessed July 2007. *Note: This USDA website also contains reports on the economics of bioethanol.*

Modeling and Simulating Human Behavior

Introduction

Christian Lebiere
Carnegie Mellon University
Pittsburgh, Pennsylvania

Robert Wray
Soar Technology
Ann Arbor, Michigan

Mature science and engineering disciplines have long converged on levels of description that best and most efficiently capture the set of phenomena on which they are focused. This convergence has led to the evolution of frameworks that allowed each discipline to build on the results of the others while focusing at different levels of abstraction. Obvious examples in the basic sciences are the different levels of focus in biology, chemistry, and physics. Such consensus has escaped up to now the study of human behavior, with various levels of description fighting for supremacy rather than building upon each other. However, recent developments in the cognitive sciences have started to lead toward such a convergence. Detailed study of neural anatomy and physiology reveal details of the computational properties of individual neurons and neural circuits. Functional brain imaging techniques are generating a wealth of data about brain organization. Integrated computational cognitive architectures have enabled precise modeling of quantitative human performance data of increasingly complex tasks in dynamic environments, including social interactions with other entities. Artificial intelligence techniques inspired by human cognition are being used to control robots and provide realistic partners and opponents in virtual reality training. Thus, these disciplines are starting to converge to create an empirical and computational framework in which to understand, model, and simulate human behavior at multiple levels of abstraction, from detailed neural simulations to high-fidelity cognitive models of behavior, to complex intelligent agents acting and interacting in dynamic environments.

Computational Cognitive Neuroscience and Its Applications

Laurent Itti
University of Southern California
Los Angeles, California

A number of tasks that can be effortlessly achieved by humans and other animals have until recently remained seemingly intractable for computers. Striking examples of such tasks exist in the visual and auditory domains. These include recognizing the face of a friend in a photograph; understanding whether a new, never-seen object is a car or an airplane; quickly finding objects or persons of interest like one's child in a crowd; understanding fluent speech while also deciphering the emotional state of the speaker; or reading cursive handwriting. In fact, several of these tasks have become the hallmark of human intelligence, while other, seemingly more complex and more cognitively involved tasks, such as playing checkers or chess, solving differential equations, or proving theorems, have been mastered by machines to a reasonable degree (Samuel, 1959; Newell and Simon, 1972). An everyday demonstration of this state of affairs is the use of simple image, character, or sound recognition in CAPTCHA tests (completely automated public Turing tests to tell computers and humans apart; see von Ahn et al., 2004) used by many websites to ensure that a human, rather than a software robot, is accessing the site. (CAPTCHA tests, for example, are used by websites that provide free e-mail accounts to registered users, and are a simple yet imperfect way to prevent spammers from opening thousands of e-mail accounts using an automated script.)

To some extent these machine-intractable tasks are the cause for our falling short on the early promises made in the 1950s by the founders of artificial intelligence, computer vision, and robotics (Minsky, 1961). Although tremendous progress has been made in just a half century, and one is beginning to see cars that can drive on their own or robots that vacuum the floor without human supervision, such machines have not yet reached mainstream adoption and remain highly limited in their ability to interact with the real world. Although in the early years one could blame the poor performance of machines on limitations in computing resources, rapid advances in microelectronics have now rendered such excuses less credible. The core of the problem is not only how many computing cycles one may have to perform a task but also how those cycles are used and in what kind of algorithm and computing paradigm.

For biological systems, interacting with the visual world (in particular through vision, audition, and other senses) is key to survival. Essential tasks like locating and identifying potential prey, predators, or mates must be performed quickly and reliably if an animal is to stay alive. Thus, biological visual systems may have evolved to become particularly attuned to some visual features or events that may indicate possible dangers or rewards. Taking inspiration from nature, recent work in computational neuroscience has hence started to devise a new breed of algorithms, which can be more flexible, robust, and adaptive when confronted with the complexities of the real world. I focus here on describing recent progress with a few simple examples of such algorithms, concerned with directing attention toward interesting locations in a visual scene, so as to concentrate the deployment of computing resources primarily onto these locations.

MODELING VISUAL ATTENTION

Positively identifying any and all interesting targets in one's visual field has prohibitive computational complexity, making it a daunting task even for the most sophisticated biological brains (Tsotsos, 1991). One solution, adopted by primates and many other animals, is to break down the visual analysis of the entire field of view into smaller regions, each of which is easier to analyze and can be processed in turn. This serialization of visual scene analysis is operationalized through mechanisms of *visual attention*. A common (although somewhat inaccurate) metaphor for visual attention is that of a virtual "spotlight," shifting toward and highlighting different subregions of the visual world, so that one region at a time can be subjected to more detailed visual analysis (Treisman and Gelade, 1980; Crick, 1984; Weichselgartner and Sperling, 1987). The central problem in attention research then becomes how best to direct this spotlight toward the most interesting and behaviorally relevant visual locations. Simple strategies, like constantly scanning the visual field from left to right and from top to bottom, like many computer algorithms do, may be too slow for situations where survival depends on reacting quickly. Recent progress in computational neuroscience has proposed a number of new, biologically inspired algorithms that implement more efficient strategies. These algorithms usually distinguish between a bottom-up drive of attention toward conspicuous, or salient, locations in the visual field, and a volitional and task-dependent top-down drive of attention toward behaviorally relevant scene elements (Desimone and Duncan, 1995; Itti and Koch, 2001). It is important to note here how we separate the notion of salience from that of relevance, although both have often been colloquially associated. Simple examples of bottom-up salient and top-down relevant items are shown in Figure 1.

Koch and Ullman (1985) introduced the idea of a saliency map to accomplish attentive selection in the primate brain. This is an explicit two-dimensional map that encodes the saliency of objects in the visual environment. Competition among neurons in this map gives rise to a single winning location that corresponds to the most salient object, which constitutes the next target. If this location is subsequently inhibited, the system automatically shifts to the next most salient location, endowing the search process with internal dynamics.

Later research has further elucidated the basic principles behind computing salience (Figure 2). One important principle is the detection of locations whose local visual statistics significantly differ from the surrounding image statistics, along some dimension or combination of dimensions that are currently relevant to the subjective observer (Itti et al., 1998; Itti and Koch, 2001). This significant difference could be in a number of simple visual feature dimensions that are believed to be represented in the early stages of cortical visual processing: color, edge orientation, luminance, or motion direction (Treisman and Gelade, 1980; Itti and Koch, 2001). Wolfe and Horowitz (2004) provide a very nice review of which elementary visual features may strongly contribute to visual salience and guide visual search.

Two mathematical constructs can be derived from electrophysiological recordings in living brains, which shed light onto how this detection of statistical odd man out may be carried out. First, early visual neurons often have center-surround receptive field structures, by which the neuron's view of the world consists of two antagonistic subregions of the visual field, a small central region that drives the neuron in one direction (e.g., excites the neuron when a bright pattern is presented) and a larger, concentric surround region that drives the neuron in the opposite direction (e.g., inhibits the neuron when a bright pattern is presented) (Kuffler, 1953). In later processing stages this simple concentric center-surround receptive field structure is replaced by more complex types of differential operators, such as Gabor receptive fields sensitive to the presence of oriented line segments (Hubel and Wiesel, 1962). In addition, more recent research has unraveled how neurons may react with similar stimulus preferences but be sensitive to different locations in the visual field. In particular, neurons with sensitivities to similar patterns tend to inhibit each other, a phenomenon known as nonclassical surround inhibition (Allman et al., 1985; Cannon and Fullenkamp, 1991; Sillito et al., 1995). Taken together these basic principles suggest ways by which detecting an odd man out, or a significantly different item in a display, can be achieved in a very economical (in terms of neural hardware) yet robust manner.

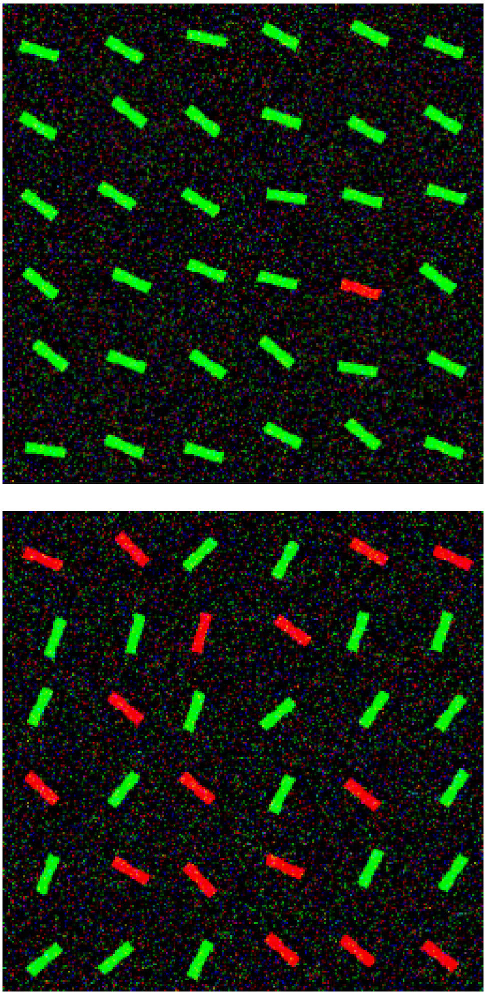


FIGURE 1 (Top) Example where one item (a dark-red and roughly horizontal bar[1]) in an array of items (bright-green bars) is highly salient and immediately and effortlessly grabs visual attention in a bottom-up, image-driven manner. (Bottom) Example where a similar item (a dark-red and roughly vertical bar) is not salient but may be behaviorally relevant if your task is to find it as quickly as possible; top-down, volition-driven mechanisms must be deployed to initiate a search for the item. (Also see Treisman and Gelade, 1980.) Source: Itti and Koch, 2000. Reprinted with permission from Elsevier.

[1] To see figures in color, go to the 2007 U.S. FOE Presentations page at <http://www.nae.edu/frontiers>.

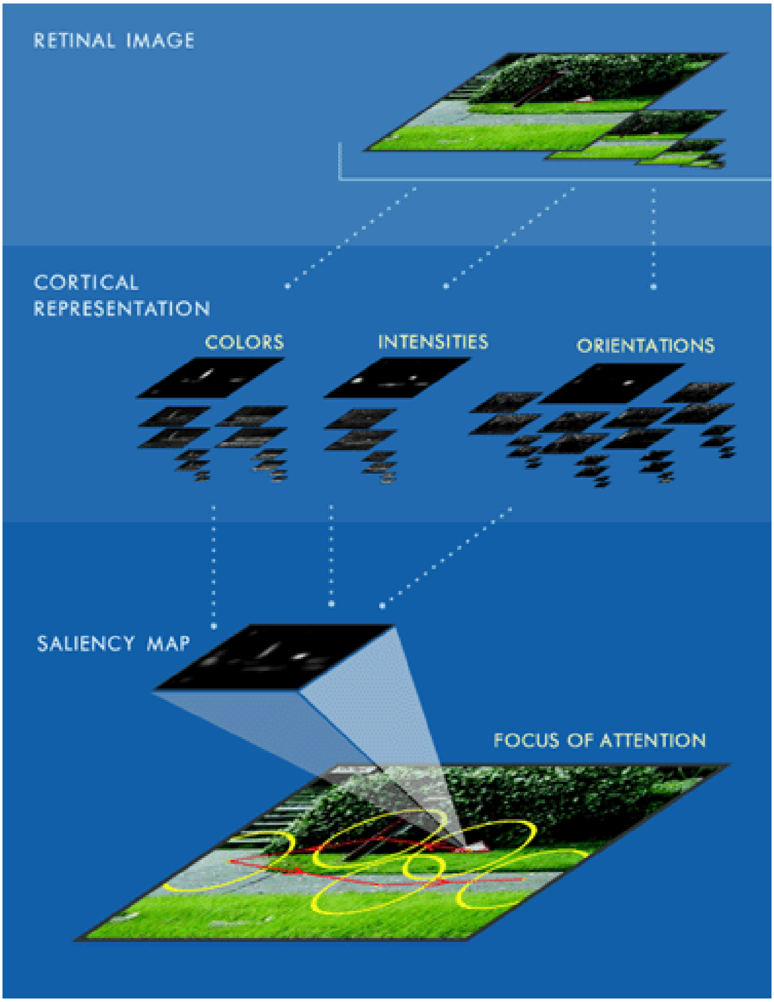


FIGURE 2 Architecture for computing saliency and directing attention toward conspicuous locations in a scene. Incoming visual input (top) is processed at several spatial scales along a number of basic feature dimensions, including color, luminance intensity, and local edge orientation. Center-surround operations as well as nonclassical surround inhibition within each resulting feature map enhance the neural representation of locations that significantly stand out from their neighbors. All feature maps feed into a single saliency map that topographically represents salience irrespectively of features. Attention is first directed to the most salient location in the image, and subsequently to less salient locations. Source: <http://iLab.usc.edu>.

Many computational models of human visual search have embraced the idea of a saliency map under different guises (Treisman, 1988; Wolfe, 1994; Niebur and Koch, 1996; Itti and Koch, 2000). The appeal of an explicit saliency map is the relatively straightforward manner it allows the input from multiple, quasi-independent feature maps to be combined and to give rise to a single output: the next location to be attended. Related formulations of this basic principle have been expressed in slightly different terms, including defining salient locations as those that contain spatial outliers (Rosenholtz, 1999), which may be more informative in Shannon's sense (Bruce and Tsotsos, 2006). Similarly, a more general formulation has been proposed by which salient locations may be more surprising in a Bayesian sense (Itti and Baldi, 2006). Electrophysiological evidence points to the existence of several neuronal maps, in the pulvinar, the superior colliculus, and the intraparietal sulcus, which appear to encode specifically for the saliency of a visual stimulus (Robinson and Petersen, 1992; Gottlieb et al., 1998; Colby and Goldberg, 1999). Presumably, these different brain areas encode for different aspects of salience, from purely bottom-up representations to more elaborated representations that may incorporate top-down influences and behavioral goals.

Fully implemented computational models of attention, although relatively recent, have spawned a number of exciting new technological applications, including

- automatic target detection (e.g., finding traffic signs along the road or military vehicles in a savanna) (Itti and Koch, 2000);
- robotics (using salient objects in the environment as navigation landmarks) (Frintrop et al., 2006; Siagian and Itti, 2007);
- image and video compression (e.g., giving higher quality to salient objects at the expense of degrading background clutter) (Osberger and Maeder, 1998; Itti, 2004);
- automatic cropping or centering of images for display on small portable screens (Le Meur et al., 2006); and
- medical image processing (e.g., finding tumors in mammograms) (Hong and Brady, 2003).

FROM ATTENTION TO VISUAL SCENE UNDERSTANDING

How can cognitive control of attention be integrated with the simple framework described so far? One hypothesis is that another map exists in the brain that encodes for top-down relevance of scene locations (a task relevance map, or TRM) (Navalpakkam and Itti, 2005). In the TRM a highlighted location might have a particular relevance because, for example, it is the estimated target of a flying object; it is the next logical place to look at given a sequence of action, recognition, and reasoning steps, like in a puzzle video game; or one just guesses there might be something worth looking at there. Clearly, building computational models in which a TRM is populated with sensible information is a very difficult problem, tantamount to solving most of the problems in vision and visual scene understanding.

Recent work has demonstrated how one can implement quite simple algorithms that deliver highly simplified but useful task relevance maps. In particular, Peters and Itti (2007) (Figure 3) recently proposed a model of spatial attention that (1) can be applied to arbitrary static and dynamic image sequences with interactive tasks and (2) combines a general computational implementation of both bottom-up (BU) saliency and dynamic top-down (TD) task relevance (Figure 4). The novelty lies in the combination of these elements and in the fully automated nature of the model. The BU component computes a saliency map from 12 low-level, multiscale visual features, similar to the process described in the previous section. The TD component computes a low-level signature of the entire image (the "gist" of the scene) (Torralba, 2003; Peters and Itti, 2007) and learns to associate different classes of signatures with the different gaze patterns recorded from human subjects performing a task of interest. It is important to note that while we call this component top-down, it does not necessarily imply that it is high level—in fact, while the learning phase certainly uses high-level information about the scenes, that becomes summarized into the learned associations between scene signatures and expected behavior in the form of TD gaze position maps. It is also important to realize how, while the computation of image signatures relies on a visual processing algorithm, the association between signatures and predicted gaze positions is learned and thus forms a descriptive behavioral model.

We measured (Peters and Itti, 2007) the ability of this model to predict the eye movements of people playing contemporary video games. We found that the TD model alone predicts where humans look about twice as well as does the BU model alone; in addition, a combined BU×TD model performs significantly better than either individual component. Qualitatively the combined model predicts some easy-to-describe but hard-to-compute aspects of attentional selection, such as shifting attention leftward when approaching a left turn along a racing track. Thus, our study demonstrates the advantages of integrating bottom-up factors derived from a saliency map and top-down factors learned from image and task contexts in predicting where humans look while performing complex visually guided behavior. In continuing work we are exploring ways of introducing additional domain knowledge into the top-down component of our attentional system.

DISCUSSION AND CONCLUSIONS

Visual processing of complex natural environments requires animals to combine, in a highly dynamic and adaptive manner, sensory signals that originate from the environment (bottom-up) with behavioral goals and priorities dictated by the task at hand (top-down). In the visual domain, bottom-up and top-down guidance

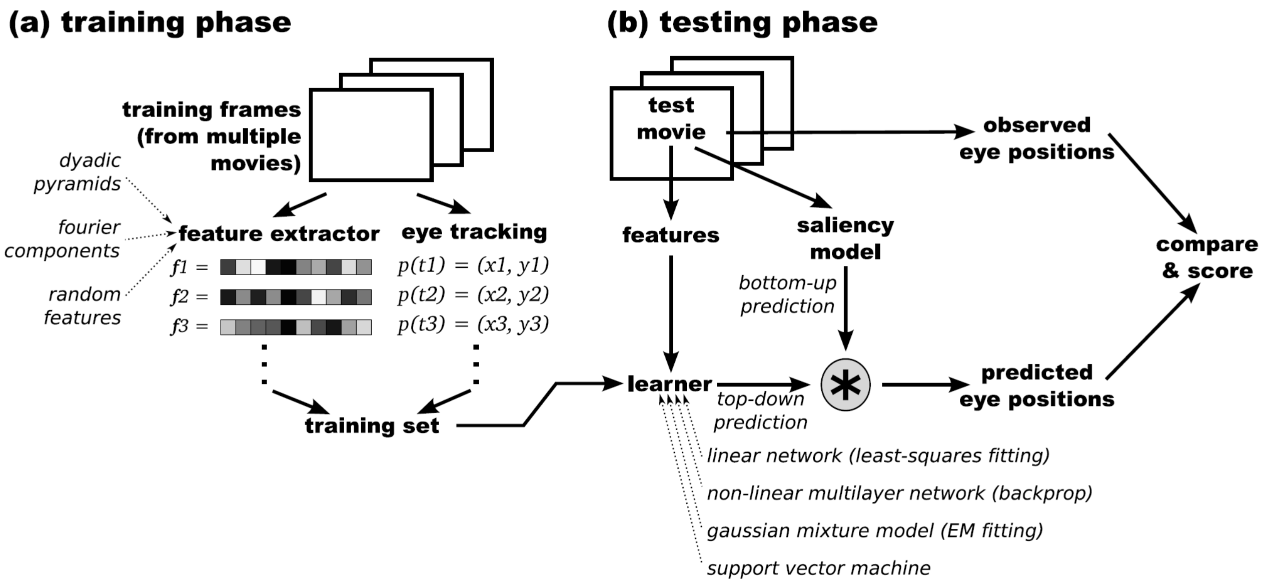


FIGURE 3 Schematic illustration of our model for learning task-dependent, top-down influences on eye position. First, in (a), the training phase, we compile a training set containing feature vectors and eye positions corresponding to individual frames from several video game clips that were recorded while observers interactively played the games. The feature vectors may be derived from either the Fourier transform of the image luminance; or dyadic pyramids for luminance, color, and orientation; or as a control condition, a random distribution. The training set is then passed to a machine-learning algorithm to learn a mapping between feature vectors and eye positions. Then, in (b), the testing phase, we use a different video game clip to test the model. Frames from the test clip are passed in parallel to a bottom-up saliency model, as well as to the top-down feature extractor, which generates a feature vector that is used to generate a top-down eye position prediction map. The bottom-up and top-down prediction maps can be combined with pointwise multiplication, and the individual and combined maps can be compared with the actual observed eye position. Source: Peters and Itti, 2007. Reprinted with permission. © IEEE.

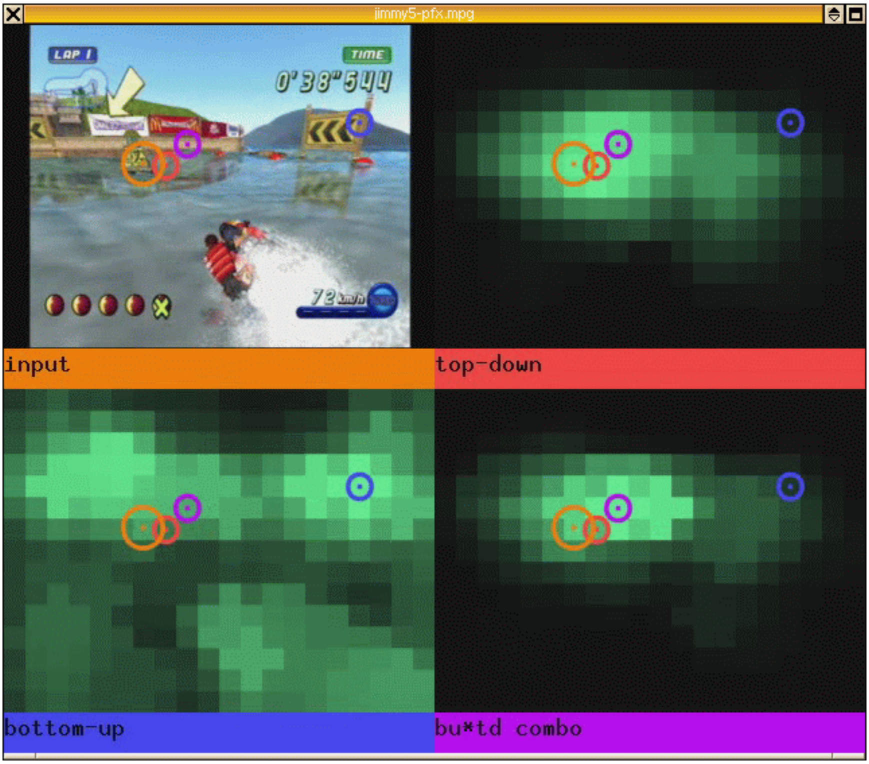


FIGURE 4 Combined bottom-up plus top-down model and comparison with human eye movements. While a subject was playing this jet-ski racing game (upper left), his gaze was recorded (large orange circle in all four quadrants[2]). Simultaneously, the top-down map (upper right) was computed from previously learned associations between scene gist and human gaze (location of maximum top-down activity indicated by small red circle). The bottom-up saliency map was also computed from the video input (lower left, maximum at small blue circle). Both maps were combined, here by taking a pointwise product (lower right, maximum at small purple circle). This figure exemplifies a situation where top-down dominates gaze allocation (several objects in the scene are more bottom-up salient than the one being looked at). Source: Peters and Itti, 2007. Reprinted with permission. © IEEE.

of attention toward salient or behaviorally relevant targets have been studied and modeled extensively, as has more recently the interaction between bottom-up and top-down control of attention. Thus, in recent years a number of neurally inspired

[2] To see figures in color, go to the 2007 U.S. FOE Presentations page at <http://www.nae.edu/frontiers>.

computational models have emerged that demonstrate unparalleled performance, flexibility, and adaptability in coping with real-world inputs. In the visual domain, in particular, such models are achieving great strides in tasks, including focusing attention onto the most important locations in a scene, recognizing attended objects, computing contextual information in the form of the gist of the scene, and planning or executing visually guided motor actions.

Together these models present great promise for future integration with more conventional artificial intelligence techniques. Symbolic models from artificial intelligence have reached significant maturity in their cognitive reasoning and top-down abilities, but the worlds in which they can operate have been necessarily simplified (e.g., a chess board, a virtual maze). Very exciting opportunities exist to attempt to bridge the gap between the two disciplines of neural modeling and artificial intelligence.

ACKNOWLEDGMENTS

This work is supported by the Human Frontier Science Program, the National Science Foundation, the National Geospatial-Intelligence Agency, and the Defense Advanced Research Projects Agency. The author affirms that the views expressed herein are solely his own, and do not represent the views of the United States government or any agency thereof.

REFERENCES

Allman, J., F. Miezin, and E. McGuinness. 1985. Stimulus specific responses from beyond the classical receptive field: Neurophysiological mechanisms for local-global comparisons in visual neurons. Annual Review of Neuroscience 8:407-430.

Bruce, N., and J. K. Tsotsos. 2006. Saliency based on information maximization. Advances in Neural Information Processing Systems 18:155-162.

Cannon, M. W., and S. C. Fullenkamp. 1991. Spatial interactions in apparent contrast: Inhibitory effects among grating patterns of different spatial frequencies, spatial positions and orientations. Vision Research 31:1985-1998.

Colby, C. L., and M. E. Goldberg. 1999. Space and attention in parietal cortex. Annual Review of Neuroscience 22:319-349.

Crick, F. 1984. Function of the thalamic reticular complex: The searchlight hypothesis. Proceedings of the National Academy of Sciences 81(14):4586-4590.

Desimone, R., and J. Duncan. 1995. Neural mechanisms of selective visual attention. Annual Review of Neuroscience 18:193-222.

Frintrop, S., P. Jensfelt, and H. Christensen. 2006. Attentional landmark selection for visual SLAM. Pp. 2582-2587 in Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06). Piscataway, NJ: IEEE.

Gottlieb, J. P., M. Kusunoki, and M. E. Goldberg. 1998. The representation of visual salience in monkey parietal cortex. Nature 391:481-484.

Hong, B.-W., and M. Brady. 2003. A topographic representation for mammogram segmentation. Lecture Notes in Computer Science 2879:730-737.

Hubel, D. H., and T. N. Wiesel. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. Journal of Physiology 160:106-154.

Itti, L. 2004. Automatic foveation for video compression using a neurobiological model of visual attention. IEEE Transactions on Image Processing 13:1304-1318.

Itti, L., and P. Baldi. 2006. Bayesian surprise attracts human attention. Pp. 1-8 in Advances in Neural Information Processing Systems, Vol. 19 (NIPS 2005). Cambridge, MA: MIT Press.

Itti, L., and C. Koch. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research 40(10-12):1489-1506.

Itti, L., and C. Koch. 2001. Computational modeling of visual attention. Nature Reviews Neuroscience 2(3):194-203.

Itti, L., C. Koch, and E. Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20:1254-1259.

Koch, C., and S. Ullman. 1985. Shifts in selective visual attention: Towards the underlying neural circuitry. Human Neurobiology 4:219-227.

Kuffler, S. W. 1953. Discharge patterns and functional organization of mammalian retina. Journal of Neurophysiology 16(1):37-68.

Le Meur, O., P. Le Callet, D. Barba, and D. Thoreau. 2006. A coherent computational approach to model bottom-up visual attention. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(5):802-817.

Minsky, M. 1961. Steps toward artificial intelligence. Proceedings of the Institute of Radio Engineers 49:8-30.

Navalpakkam, V., and L. Itti. 2005. Modeling the influence of task on attention. Vision Research 45(2):205-231.

Newell, A., and H. Simon. 1972. Human Problem Solving. Englewood Cliffs, NJ: Prentice-Hall.

Niebur, E., and C. Koch. 1996. Control of selective visual attention: Modeling the "where" pathway. Neural Information Processing Systems 8:802-808.

Osberger, W., and A. J. Maeder. 1998. Automatic identification of perceptually important regions in an image using a model of the human visual system. Pp. 701-704 in Proceedings of the 14th International Conference on Pattern Recognition, Vol. 1. Piscataway, NJ: IEEE.

Peters, R. J., and L. Itti. 2007. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. Pp. 1-8 in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE.

Robinson, D. L., and S. E. Petersen. 1992. The pulvinar and visual salience. Trends in Neuroscience 15:127-132.

Rosenholtz, R. 1999. A simple saliency model predicts a number of motion popout phenomena. Vision Research 39:3157-3163.

Samuel, A. 1959. Some studies in machine learning using the game of checkers. IBM Journal 3(3):210-229.

Siagian, C., and L. Itti. 2007. Biologically-inspired robotics vision Monte-Carlo localization in the outdoor environment. Pp. 1723-1730 in Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07). Piscataway, NJ: IEEE.

Sillito, A. M., K. L. Grieve, H. E. Jones, J. Cudeiro, and J. Davis. 1995. Visual cortical mechanisms detecting focal orientation discontinuities. Nature 378:492-496.

Torralba, A. 2003. Modeling global scene factors in attention. Journal of the Optical Society of America A - Optics Image Science and Vision 20(7):1407-1418.

Treisman, A. 1988. Features and objects: The fourteenth Bartlett memorial lecture. Quarterly Journal of Experimental Psychology 40:201-237.

Treisman, A., and G. Gelade. 1980. A feature integration theory of attention. Cognitive Psychology 12:97-136.

Tsotsos, J. K. 1991. Is complexity theory appropriate for analyzing biological systems? Behavioral and Brain Sciences 14(4):770-773.

von Ahn, L., M. Blum, and E. Langford. 2004. How lazy cryptographers do AI. Communications of the ACM 47(2):57-60.

Weichselgartner, E., and G. Sperling. 1987. Dynamics of automatic and controlled visual attention. Science 238:778-780.

Wolfe, J. M. 1994. Guided search 2.0: A revised model of visual search. Psychonomic Bulletin and Review 1(2):202-238.

Wolfe, J. M., and T. S. Horowitz. 2004. What attributes guide the deployment of visual attention and how do they do it? Nature Reviews Neuroscience 5:1-7.

ADDITIONAL READING

Wolfe, J. M. 1998. Visual search. Pp. 13-74 in Attention, H. Pashler, ed. London: University College.

Barriers, Bridges, and Progress in Cognitive Modeling for Military Applications

Kevin A. Gluck
Air Force Research Laboratory
Mesa, Arizona

The role of the Air Force Research Laboratory (AFRL), like the other service laboratories, is to conduct the basic and applied research and advanced technology development necessary to create future technology options for the Department of Defense. At the Warfighter Readiness Research Division of AFRL's Human Effectiveness Directorate we have initiated a research program focused on mathematical and computational cognitive process modeling for replicating, understanding, and predicting human performance and learning. This research will lead to new technology options in the form of human-level synthetic teammates, cognitive readiness analysis tools, and predictive and prescriptive knowledge-tracing algorithms. Creating a future in which these objectives become realities requires tightly coupled, multidisciplinary, collaborative interaction among scientists and engineers dedicated to overcoming the myriad challenges standing between current reality and our future vision.

BARRIERS AND BRIDGES

There are many barriers to progress in cognitive science in general and to computational cognitive process modeling in particular. I will emphasize just two of them here. The first is a domain barrier. There exists an infinite variety of domains in which humans learn and perform, and in order to simulate human performance and learning in a particular domain, we must provide relevant domain knowledge to the simulated human. Transfer from one domain to the next is largely a function of the degree to which the knowledge in the two domains overlaps. The reason this is problematic for scientific progress is that the domains typically used to study human cognitive functioning in the laboratory are very different from the domains of application in the real world. Laboratory domains are mostly simple, abstract, and of short duration, whereas real-world application domains are complex, situated, and of long duration. Thus, in the field of cognitive science we must look for ways to build bridges between laboratory and applied contexts.

The second barrier I will emphasize here is a disciplinary barrier. Cognitive science is a field of study comprising seven subdisciplines: anthropology, artificial intelligence, education, linguistics, neuroscience, philosophy, and psychology. These subdisciplines involve very different methods, frameworks, and theories, and it is both challenging and exciting to make progress at disciplinary intersections. For instance, there is a powerful *zeitgeist* currently associated with neuroscience-based explanations of everything from attentional, perceptual, and related cognitive phenomena (leading to the creation of a field known as computational cognitive neuroscience—see Itti's paper in this volume) to complex economic decision making (leading to the creation of a field known as neuroeconomics—see Glimcher, 2003). This has led people in some circles to speculate that there ought to be ways to improve the readiness of our military personnel by capitalizing on the tools, methods, empirical results, and theories of neuroscience. Simultaneously there is interest in bringing together the subdisciplines of anthropology, artificial intelligence, and psychology in order to better understand and prepare for multicultural interaction (see the paper by van Lent and colleagues in this volume). Making scientific progress across these disciplinary boundaries requires that we build bridges among the neural, cognitive, and social bands of human experience (Newell, 1990). Anderson and Gluck (2001) noted that the same challenge exists in connecting neuroscience and educational practice and proposed that cognitive architectures are an appropriate formalism for building such bridges. I propose that cognitive architectures also are an appropriate formalism for building bridges from neuroscience to the military's cognitive readiness applications, using cognitive phenomena and models.

THE SOLUTION: COGNITIVE ARCHITECTURES

The purpose of all scientific disciplines is to identify invariant features and explanatory mechanisms for the purpose of understanding the phenomena of interest in the respective disciplines. Within the cognitive science community there is an approximately 50-year history of empirical research that involves using carefully constructed (usually simple and abstract) laboratory tests to isolate components of the human cognitive system in order to model and understand them. Sometimes optimistically referred to as "divide and conquer," this approach has led to comprehensive empirical documentation and sophisticated theories of hundreds of phenomena (e.g., fan effect, framing effect, Stroop effect) and functional components (e.g., attention, perception, memory, cognition, motor movement). A subset of the cognitive science community has become concerned that this divide and conquer approach is not leading to a unified understanding of human cognitive functioning, and has proposed cognitive architectures as the solution to that problem (Newell, 1973). Thus, cognitive architectures are intended to serve an integrative, cumulative role within the cognitive science community. They are where the fractionated theories come together in a unifying account not only of the computational functionality of the component processes but also of the architectural control structures that define the relationships among those components, and of the representation of knowledge content that is used by cognition. Gray (2007) explains how these three theoretical spaces (components, control structures, and knowledge) interact and provides numerous case studies of each. Ultimately it is at the intersection of these theories that cognitive architectures exist.

ONGOING COGNITIVE MODELING RESEARCH

Our cognitive modeling research program at the Air Force Research Laboratory's Mesa Research Site is organized around a set of methodological strategies with associated benefits. First, we are using and improving on the ACT-R (Adaptive Control of Thought—Rational) cognitive architecture (Anderson et al., 2004), because it provides a priori theoretical constraints on the models we develop; facilitates model reuse among members of the ACT-R research community; and serves the integrating, unifying role described earlier. Second, we use the architecture, or equations and algorithms inspired by it, to make quantitative predictions in order to facilitate eventual transition to applications that make accurate, precise predictions about human performance and learning. Third, we develop models in both abstract, simplified laboratory tasks and in more realistic, complex synthetic task environments in order to begin constructing those bridges between the laboratory and the real world. Fourth, we compare the predictions of our models to human-subject data, in order to evaluate the necessity and sufficiency of the computational mechanisms and parameters that are driving those predictions and in order to evaluate the validity of the models. We are pursuing this research strategy in several lines of research, which I briefly describe next.

Knowledge tracing. This is our only research line that is entirely mathematical modeling and does not involve a computational modeling component. The current approach is an extension and (we think) improvement to the general performance equation proposed by Anderson and Schunn (2000); thus, it derives from the computational implementation of learning and forgetting processes in ACT-R. The new equation allows us to make performance predictions or prescribe the timing and frequency of training, both of which will enable tailored training experiences at individual and team levels of analysis (Jastrzembski et al., 2006).

Communication. One of the barriers standing between us and human-level synthetic teammates is that we don't have a valid computational implementation of natural language, verbal or otherwise. This is critical because good teammates adapt their communications in order to facilitate accomplishing the shared mission. Our research in natural language modeling involves extending the double R computational cognitive linguistic theory to knowledge-rich, time-pressured team performance environments similar to those encountered in real-world situations, such as unmanned air vehicle reconnaissance missions (Ball et al., 2007).

Spatial competence. Spatial cognition has long been a subspecialization within the cognitive science community, but typically individual scientists or research groups adopt particular phenomena to study without worrying about how the pieces of the spatial cognitive system come back together to create a more general competence. It turns out there is no comprehensive theory of the mechanisms and processes that allow for spatial competence. Our research in this area is pushing the field and the ACT-R architecture in the direction of a neurofunctional and architectural view of how spatial competence is realized in the brain and the mind (Gunzelmann and Lyon, 2008).

Fatigue. There is a rich history of sleep-related fatigue research conducted in and sponsored by the military laboratories. We are adding a new twist to that tradition by implementing new architectural mechanisms and processes that allow us to replicate the effects of sleepiness on the cognitive system. The process models are then combined with biomathematical models of the circadian and sleep homeostat systems to create the capacity to predict what the precise effects of sleep deprivation or long-term sleep restriction will be in a given performance context (Gunzelmann et al., 2007).

High-performance and volunteer computing. As our cognitive modeling research expanded in breadth and depth and our scientific and technical objectives grew more ambitious, we began to exceed the capacity of our local computing resources. In the search first for more resources and subsequently for more intelligent and efficient use of available resources, we have begun to use both high-performance computing and volunteer computing as platforms for processor horsepower. We have demonstrated that such platforms can indeed be used productively for faster progress in cognitive modeling (Gluck et al., 2007) and are investing in additional software improvements for facilitating the use of these resources.

AN IMPORTANT DIRECTION FOR THE RESEARCH COMMUNITY

I close by mentioning an important research direction for the cognitive modeling community: overcoming the knowledge engineering bottleneck. The key here is not the development of tools for doing manual knowledge engineering more efficiently, although that is a perfectly fine idea in the interim. Instead, I believe it is critical that we develop the ability for our modeling architectures to acquire their own knowledge without direct human assistance. This will require a variety of learning mechanisms based on a combination of cognitive psychology, machine learning, and Internet search algorithms.

REFERENCES

Anderson, J. R., and K. A. Gluck. 2001. What role do cognitive architectures play in intelligent tutoring systems? Pp. 227-261 in Cognition and Instruction: 25 Years of Progress, D. Klahr and S. M. Carver, eds. Mahwah, NJ: Lawrence Erlbaum Associates.

Anderson, J. R., and C. D. Schunn. 2000. Implications of the ACT-R learning theory: No magic bullets. Pp. 1-34 in Advances in Instructional Psychology: Educational Design and Cognitive Science, Vol. 5., R. Glaser, ed. Mahwah, NJ: Lawrence Erlbaum Associates.

Anderson, J. R., D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin. 2004. An integrated theory of the mind. Psychological Review 111:1036-1060.

Ball, J., A. Heiberg, and R. Silber. 2007. Toward a large-scale model of language comprehension in ACT-R 6. Pp. 163-168 in Proceedings of the 8th International Conference on Cognitive Modeling. New York: Psychology Press.

Glimcher, P. W. 2003. Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics. Cambridge, MA: MIT Press.

Gluck, K. A., M. Scheutz, G. Gunzelmann, J. Harris, and J. Kershner. 2007. Combinatorics meets processing power: Large-scale computational resources for BRIMS. Pp. 73-83 in Proceedings of the Sixteenth Conference on Behavior Representation in Modeling and Simulation. Orlando, FL: Simulation Interoperability Standards Organization.

Gray, W. D., ed. 2007. Integrated Models of Cognitive Systems. New York: Oxford University Press.

Gunzelmann, G., and D. R. Lyon. 2008. Mechanisms of human spatial competence. Pp. 288-307 in Spatial Cognition V: Lecture Notes in Artificial Intelligence #4387, T. Barkowsky, C. Freksa, M. Knauff, B. Krieg-Bruckner, and B. Nebel, eds. Berlin, Germany: Springer-Verlag.

Gunzelmann, G., K. Gluck, J. Kershner, H. P. A. Van Dongen, and D. F. Dinges. 2007. Understanding decrements in knowledge access resulting from increased fatigue. Pp. 329-334 in Proceedings of the 29th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

Jastrzembski, T. S., K. A. Gluck, and G. Gunzelmann. 2006. Knowledge tracing and prediction of future trainee performance. Pp. 1498-1508 in Proceedings of the Interservice/Industry Training, Simulation, and Education Conference. Orlando, FL: National Training Systems Association.

Newell, A. 1973. You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. Pp. 283-310 in Visual Information Processing, W. C. Chase, ed. New York: Academic Press.

Newell, A. 1990. Unified Theories of Cognition. Cambridge, MA: Harvard University Press.

Modeling of Culturally Affected Human Behavior

Michael van Lent, Mark Core, Steve Solomon,
Milton Rosenberg, Ryan McAlinden, and Paul Carpenter
Institute for Creative Technologies
University of Southern California
Los Angeles, California

One major textbook divides artificial intelligence (AI) systems into two categories, those that are meant to think and act rationally and those that are meant to think and act like people (Russell and Norvig, 2003). Developers of systems that focus on rational behavior seek to maximize expected rewards based on the agent's goals; they do not attempt to account for the many social or cultural factors that influence human behavior. Even systems with human-like behavior as their goal, such as ACT-R (Anderson and Lebiere, 1998) and SOAR (Laird et al., 1987), generally focus either on passing the Turing test or simulating detailed aspects of cognition and/or neural physiology. These models are becoming more like intelligent software agents and are advancing our understanding of the neural and cognitive layers of human thought. However, very few AI systems explore the sociocultural layer of human thought.

For decades, the so-called "soft sciences" (psychology, sociology, and anthropology) have focused on the behavioral phenomena that "make us human," such as emotion, personality, and culture. Not surprisingly, however, these investigations have generated many competing, hotly debated theories but few clear answers.

A few brave AI researchers have waded into the fray by selecting individual theories and attempting to model them computationally in combination with agent-based AI methods. For example, the PsychSim system is a multi-agent-based computer program that lets social psychologists and sociologists run experiments in simulation (Pynadath and Marsella, 2005). PsychSim draws on theory of mind from the social sciences and combines it with standard AI approaches, such as decision theory and partially observable Markov decision processes, to create a system of agents that can model and reason about each other.

Similarly, the emotion-modeling and adaptation, or EMA, model (Marsella and Gratch, 2006) combines appraisal theory (Scherer et al., 2001) with standard AI plan representations to model relationships between an agent's goals and emotions. Finally, the cultural-cognitive architecture (Taylor et al., 2007) builds on schema theory (D'Andrade, 1992) and uses the SOAR architecture to model culturally specific behavior schemas or scripts.

Culture is one of the more complex and interesting elements of human social behavior. Even the definition of the term "culture" is hotly debated. Geert Hofstede (1994), a widely known researcher of culture, defines it as "the collective programming of the mind that separates one group of people from another." Following in this vein, we define culture more specifically as the aspects of physical appearance, internal knowledge, and external behavior common to a cultural group. A cultural group is defined as a group of people who identify with the group through a shared trait, such as gender, race, ethnicity, religion, nationality, "regionality," age, economic status, social class, education, or occupation.

Individuals belong to many different cultural groups simultaneously. Current theories of culture propose that, in any given situation, an individual selects a dominant cultural identity trait, which plays a primary role in influencing his or her behavior (DiMaggio, 1997). In the Culturally Affected Behavior (CAB) Project at the University of Southern California, we distinguish between culture and personality. Culture denotes the aspects of appearance, reasoning, and behavior that are common to a group; personality denotes the aspects of appearance, reasoning, and behavior that are specific to an individual and by which that individual defines his or her identity within the group.

Previous research on culture in the fields of psychology, sociology, and anthropology can be divided into two general categories. In one category, researchers (e.g., Hofstede) attempt to identify the high-level cultural parameters (such as power distance, individualism, masculinity, uncertainty avoidance, long-term orientation) that characterize a culture. This is the cultural equivalent of the Myers-Briggs personality test (Myers, 1962). Researchers in the second category focus on detailed aspects of culturally influenced behavior (such as greetings and polite or impolite gestures) (D'Andrade, 1992; DiMaggio, 1997).

Unfortunately, it is not feasible to derive low-level details of culturally influenced behavior solely from high-level cultural parameters. Hofstede's five dimensions of culture do not provide enough information for us to derive, for example, that in Muslim cultures women should not initiate a handshake when greeting a man. However, Hofstede's masculinity dimension (which is generally very high in Muslim cultures) could be used as an indicator that culturally important details are involved in a woman greeting a man. The masculinity dimension might also suggest that Hindu cultures, which have similar masculinity values, might share many details in this area with Muslim cultures.

Thus high-level theories can provide useful indicators and parallels that could make it less difficult to create cultural-behavior "modules" that encode the details of culturally influenced behavior. At the very least, high-level theories indicate areas of behavior that are likely to have culturally specific aspects and/or suggest commonalities among cultures, suggesting that detailed cultural information might be reusable.

THE CULTURALLY AFFECTED BEHAVIOR PROJECT

At the CAB Project, we draw on the schema theory and theory of mind (mentioned above), plus shared symbol theory (D'Andrade, 1984), to develop a computational approach for representing, encoding, and using cultural knowledge at the individual and aggregate levels. A number of aspects of this problem are described in more detail below.

Computational. Unlike most research in sociology, anthropology, and psychology, we are attempting to develop cultural models and representations that can be integrated into an AI system that operates in an educational game. To be implementable in a computer simulation, the approach to cultural modeling must be computational. Although this approach must be fairly formal, models can still have significant uncertainties and approximations.

Approach. The goal of the CAP Project is not to develop a specific software implementation but to develop a conceptual approach that can be implemented in a variety of ways using general programming languages or AI.

Representing. One of our primary challenges is to create a representation that is not only easy to author and modify, but is also capable of supporting changes to an AI character's cultural model without requiring that the character's entire behavior set be re-authored.

Encoding. In addition to defining the representation of cultural knowledge, we will explore how these cultural representations can be created. One potential advantage of the modular approach to cultural representation is that it supports easier authoring of cultural models. In addition to direct authoring, encoding includes acquiring cultural knowledge from other sources, such as machine learning and data mining from databases compiled by sociologists and anthropologists.

Using. Exploring the representation and encoding aspects of cultural models requires considering how these models might be used to affect the appearance, reasoning, and behavior of AI agents. As part of the CAB Project, we are investigating a variety of ways modularized cultural models can affect agent behavior, including (1) selecting which aspect of an agent's cultural identity should be most important in a given situation, (2) filling in details about another agent or human avatar's cultural identity, (3) establishing opinions and attitudes toward others, (4) influencing the selection of goals and objectives, and (5) selecting culturally specific behaviors and actions for achieving those goals.

THE CAB APPROACH

Our goal is to combine established theories from the social sciences with computational methods from AI. However, we also draw inspiration from many social-science theories and focus on swappable culture "modules" that allow a virtual character in an educational computer game to look, sound, and act differently based on the currently loaded culture.

A primary hypothesis of the CAB Project is that it is possible to separate cultural knowledge from task or domain knowledge. This means an AI character's culture can be changed without changing the character's entire knowledge base. We are trying to modularize the cultural model into a "chunk" of knowledge that affects the AI agent's appearance, reasoning, and behavior but is as separate from the rest of the agent's behavior model as possible.

MAJOR CHALLENGES

A central challenge is to determine the extent to which cultural knowledge, which obviously has a pervasive influence on behavior, can be modularized. Some aspects of a game character, such as appearance, can be changed very easily by switching the three-dimentional character model and texture or "skin" on that model. Slightly more challenging is modifying a character's animations (how the character moves) and voice model (how the character sounds, including accent).

The greatest challenge, however, is modifying how a character thinks and reacts in interpersonal situations. We model a set of interlinked "sociocultural norms," such as "is-observant-of-Islam," "avoids-alcohol," and "is-not-corrupt," each of which has (1) a culturally dependent weight based on the importance of that norm to the target culture and (2) a situation-dependent degree based on how much the agent feels the current situation supports or challenges that norm. Many sociocultural norms represent "shared symbols" in that they reflect common attitudes toward specific objects, gestures, and concepts in the character's environment.

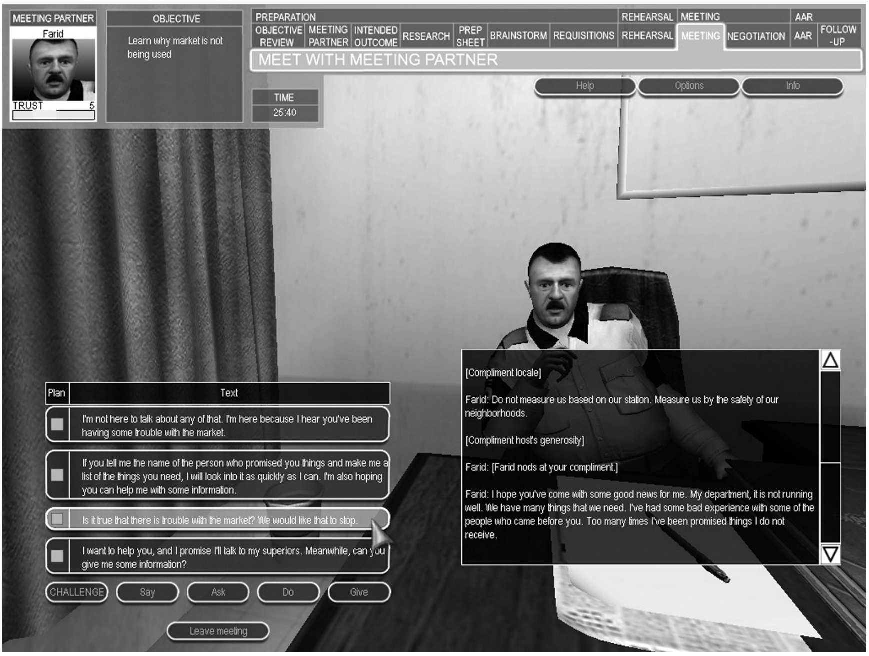
In any given situation, the weighted sum of degrees across all sociocultural norms represents the character's "sociocultural comfort level" with that situation. For example, Farid, the Iraqi police officer, might have a high weight for the "is-observant-of-Islam" sociocultural norm, which necessitates a high weight for the "avoids-alcohol" norm. If the human player were to offer Farid alcohol as a gift, the weight of the "avoids-alcohol" norm would decrease, and Farid would be less comfortable.

However, Fritz, the German police officer, might have a zero weight for both the "is-observant-of-Islam" norm and the "avoids-alcohol" norm. Thus offering a gift of alcohol to Fritz might seem to be a positive relationship-building action. However, if Fritz has a high weight for the "is-not-corrupt" norm, offering him any gift might be interpreted as an attempted bribe, which would challenge his "is-not-corrupt" norm and consequently decrease his sociocultural comfort level.

By weighting norms, every character, no matter what his or her culture, can include the entire set of sociocultural norms. If a norm is not relevant to a character's specific culture, it can be "turned off" by setting the weight to zero (as with the "is-observant-of-Islam" norm for Fritz). Therefore, changing a character's culture is simply a matter of changing the weights on the sociocultural norms rather than significantly restructuring the character's knowledge base.

APPLICATIONS

The ELECT BiLAT immersive-training application shows the advantages of the CAB cultural-behavior modules. The ELECT BiLAT application (see Figure 1) will give students an opportunity to prepare for and conduct meetings and negotiations in cross-cultural settings. At present, the cultural behaviors being developed for the ELECT project are specific to Iraqi culture. Thus the elements of Iraqi culture are not structured as a swappable module in the system but are dispersed throughout the system in both explicit and implicit ways. As a result, moving the ELECT BiLAT application to a new culture will require re-authoring



Learn why market is not being used

- □ I'm not here to talk about any of that. I'm here because I hear you've been having some trouble with the market.
- □ If you tell me the name of the person who promised you things and make me a list of the things you need, I will look into it as quickly as I can. I'm also hoping you can help me with some information.
- □ Is it true that there is trouble with the market? We would like that to stop.
- □ I want to help you, and I promise I'll talk to my superiors. Meanwhile, can you give me some information?

[Compliment locale]

Farid: Do not measure us based on our station. Measure us by the safety of our neighborhoods.

[Compliment host's generosity]

Farid: [Farid nods at your compliment.]

Farid: I hope you've come with some good news for me. My department, it is not running well. We have many things that we need. I've had some bad experience with some of the people who came before you. Too many times I've been promised things I do not receive.

FIGURE 1 A screen shot from the ELECT BiLAT application, which can help students prepare for conducting meetings and negotiating in cross-cultural settings.

much of the system rather than just re-authoring the parts specific to culturally affected behavior.

Nevertheless, by exploring the challenges to representing and authoring cultural behavior modules, the CAB Project might make it possible to adapt systems like ELECT BiLAT to a wide variety of cultures without requiring the creation of completely new databases of behavior knowledge. We are currently working on a modified version of ELECT BiLAT that can support a meeting with either Farid, the Iraqi police officer, or Fritz, the German police officer, by modifying the weightings of the character's sociocultural norms and changing the character's appearance, animations, and voice model.

This example shows how cultural information might influence the behavior of a single entity. The approach works well for modeling the influence of culture at the individual level and for small groups of individuals and enables users to interact with those groups or individuals in real time. However, at the macro level, which involves the behavior of large groups and populations, the models must apply not to individual entities but to trends in behavior across interacting cultural groups. An example of culturally affected behavior at this level is the social structure of traditional tribes, which varies from culture to culture. So far, our efforts have been focused primarily on cultural behavior at the individual level, but we will also investigate macro-level cultural behaviors.

METRICS FOR EVALUATING THE SYSTEM

The first challenge to defining metrics for success in the modeling of culturally affected behavior is determining exactly what should be measured. We have attempted to evaluate four aspects of the system. First, to ensure that the system can support an immersive user experience and that it acts in a culturally appropriate way, we must evaluate the believability (also called observational fidelity) of the generated behavior. Second, to ensure stability, we evaluate the functional validity of the system (i.e., its ability to run without bugs or crashes).

Third, to ensure that new cultures can be rapidly encoded, we measure the "authorability" of the cultural knowledge modules. Finally, we evaluate the explanatory fidelity of the system (i.e., its ability to explain why the system is behaving in a given way). Metrics for believability, authorability, and explanatory fidelity will be based on acceptance tests with users. Functional validity can be measured through unit and functional testing.

CONCLUSION

The development of computational models of the social layer of human behavior is a challenging goal. Projects like CAB require teams of researchers and authorities in disciplines that have not traditionally worked together. Fortunately, we can draw on the experiences of previous generations or researchers who have created new fields of study on the boundaries between established scientific disciplines. For example, cognitive modeling, the computational modeling of human cognition, requires expertise in cognitive psychology, computer science, and computational linguistics (the study of language using computational models). Following in the footsteps of these pioneers, computational social science has the potential to grow into a mature field of study that supports further research investigations and has important practical applications.

REFERENCES

Anderson, J., and C. Lebiere. 1998. The Atomic Components of Thought. Mahwah, NJ: Lawrence Erlbaum Associates.

D'Andrade, R. G. 1984. Cultural meaning systems. Pp. 88-119 in Culture Theory: Essays on Mind, Self, and Emotion, R. A. Shweder and R. LeVine, eds. Cambridge, UK: Cambridge University Press.

D'Andrade, R. G. 1992. Schemas and motivation. Pp. 23-44 in Human Motives and Cultural Models, R. G. D'Andrade and C. Strauss, eds. Cambridge, UK: Cambridge University Press.

DiMaggio, P. 1997. Culture and cognition. Annual Review of Sociology 23:263-287.

Hofstede, G. 1994. Cultures and Organizations: Software of the Mind: Intercultural Cooperation and Its Importance for Survival. London: HarperCollins.

Laird, J. E., A. Newell, and P. S. Rosenbloom. 1987. SOAR: An architecture for general intelligence. Artificial Intelligence 33(1):1-64.

Marsella, S., and J. Gratch. 2006. EMA: A computational model of appraisal dynamics. Pp. 601-606 in Cybernetics and Systems 2006, R. Trappl, ed. Vienna: Austrian Society for Cybernetic Studies.

Myers, J. B. 1962. The Myers-Briggs Type Indicator Manual. Princeton, NJ: Educational Testing Service.

Pynadath, D. V., and S. C. Marsella. 2005. PsychSim: Modeling theory of mind with decision-theoretic agents. Proceedings of the 19th International Joint Conference of Artificial Intelligence. Edinburgh, Scotland, July 30-August 5. Online. Available at <http://www.ijcai.org/proceedings05.php>.

Russell, S., and P. Norvig. 2003. Artificial Intelligence: A Modern Approach, 2nd ed. New York: Prentice Hall.

Scherer, K. R., A. Schorr, and T. Johnstone, eds. 2001. Appraisal Processes in Emotion: Theory, Methods, Research. London: Oxford University Press.

Taylor, G., M. Quist, S. Furtwangler, and K. Knudsen. 2007. Toward a hybrid cultural cognitive architecture. Paper presented at the Workshop on Cognition and Culture, 29th Annual Cognitive Science Society, Nashville, Tennessee, August 1-4. CD-ROM. Available from Lawrence Erlbaum Associates.

Safe Water Technologies

Introduction

Carol R. Rego
CDM
Cambridge, Massachusetts

Paul K. Westerhoff
Arizona State University
Tempe, Arizona

Water supply and distribution of safe drinking water were acclaimed as the fourth most significant engineering accomplishment of the 20th century by the National Society of Professional Engineers. Across America and worldwide not only are water resources becoming scarce but access to unpolluted water is also extremely difficult. Technologies to provide safe drinking water must be relatively inexpensive and capable of treating enormous volumes of water 24/7. In the United States alone approximately 40 billion gallons of water are treated to drinking water standards each day. The application of new technologies is often driven by new or more stringent regulations rather than pure technological advancements, although increasing attention to consumer perceptions is becoming important criteria for design. One example of this latter trend is how the water industry is struggling with addressing emerging contaminants, such as trace levels of pharmaceuticals, that are now ubiquitous in the environment. While these may pose an insignificant health concern, the public's perception is becoming increasingly important.

This session will focus on technologies that have undergone significant growth in their application as drinking water treatments over the last decade. A common trend in new technologies is that while they have a smaller physical footprint, they are more energy intensive (e.g., pressurized membrane filtration, ultraviolet irradiation). This can be contrasted to the use of lower energy-consuming biological processes. This session will conclude by addressing how water quality changes after it is treated at a centralized facility and distributed through networks of underground pipes to consumers. Society will continue to demand high-quality tap water over the next century, and new technologies and management approaches need to be developed.

In summary, the concept for this session will be, first, to feature the latest advancements in three technologies for producing high-quality water, and then conclude with a presentation on the distribution system, which is the final frontier that must be explored to holistically provide safe water to consumers.

Ultraviolet Irradiation: An Age-Old Emerging Technology for Water Treatment

Karl G. Linden
University of Colorado at Boulder

Ultraviolet (UV) irradiation is considered an emerging technology for public health protection even though it is an ageless process that was instrumental in bringing about life on Earth and was harnessed into a technology for human use in treating water in the late 1800s (Downes and Blunt, 1877). This paper will review why there is now such an emerging interest in UV, the basics of UV technology, why UV works, the use of UV for disinfection and oxidation of pollutants in water, and the frontiers of UV technologies.

TRADITIONAL USE OF UV TECHNOLOGIES IN WATER TREATMENT

Since the 1970s, there has been a growing understanding that chlorine application for disinfection of water has adverse effects in both natural waters and drinking water. In wastewater, chlorine was used to disinfect the sewage before it was discharged into natural water bodies. In drinking water, chlorine has been used for over 100 years to disinfect water for potable uses. In North America, UV first found wide application for treatment of wastewater. Because chlorine residual is toxic to aquatic life, the discharge of wastewater into waterways had to cease by either dechlorinating the water or changing the disinfectant. UV was a natural application because it was effective, does not require any storage or extended contact time, and does not leave any residual. Until recently UV was not acceptable for disinfection of drinking water for the very reason it was accepted for wastewater: no residual. However, growing concern over chlorinated disinfection by-products associated with more stringent regulations (EPA, 1996), coupled with the ineffectiveness of chlorine against hardy protozoan pathogens, has led to a search for disinfection alternatives since the 1970s. In 1998 and soon thereafter it was discovered that UV was in fact very effective against many chlorine-resistant pathogens, including *Cryptosporidium* (Clancy et al., 1998) and *Giardia* (Linden et al., 2002), which propelled UV into the forefront of disinfection alternatives and allowed the U.S. Environment Protection Agency (EPA) to establish more stringent regulations for the control of *Cryptosporidium* and *Giardia* in drinking water. Now UV is a fast-growing disinfection process being incorporated all over the world and is encouraged as an effective water disinfection process in the most recent EPA drinking water regulations (EPA, 2006).

HOW DOES UV WORK FOR DISINFECTION AND OXIDATION?

UV is a physical process, harnessing the power and energy of photons to effect destruction of pathogenic microorganisms and break apart chemical bonds of pollutants. In order for UV to be effective, the photons both have to be absorbed by the target pathogen or chemical (first law of photochemistry) and pack enough energy to cause a lasting photochemical effect.

In disinfection applications the photon target is the DNA of a microorganism. The absorbance spectrum of the target is one determinant of the wavelengths that will be most effective. UV irradiation covers the electromagnetic spectrum from 100 nm to 400 nm (100-200 nm is the vacuum UV range), where the most important wavelengths for disinfection are between 240 nm and 280 nm, the peak wavelengths of DNA absorbance (Figure 1). In chemical destruction applications, the important factors are the absorbance features of the target pollutant and the efficiency of the photonic transformation of chemical bonds (known as the quantum yield). Two examples of target pollutants are also presented in Figure 1.

Engineered UV systems were first developed at the turn of the 20th century with the mercury arc lamp. Current conventional UV lamps include mercury vapor lamps of the low-pressure (LP) and medium-pressure (MP) variety, indicating the internal mercury vapor pressure, which dictates the emission spectrum illustrated in Figure 2. LP lamps are low in power, high in UV efficiency, and generate near monochromatic emission at 253.7 nm, near the peak DNA absorbance. MP lamps are higher in power, lower in UV efficiency, and generate a broadband polychromatic emission with characteristic peaks between 200 nm and 400 nm. These differences translate into engineering decisions and opportunities for each lamp, specific to the application considered. On the frontiers of UV lamp technology are nonmercury-based sources including high-energy, pulsed UV-based lamps, often with xenon gas, and UV-based light-emitting diodes (LEDs) currently under development. An example of the types of spectra emitted from a pulsed UV source is presented in Figure 2. These new UV sources may soon radically

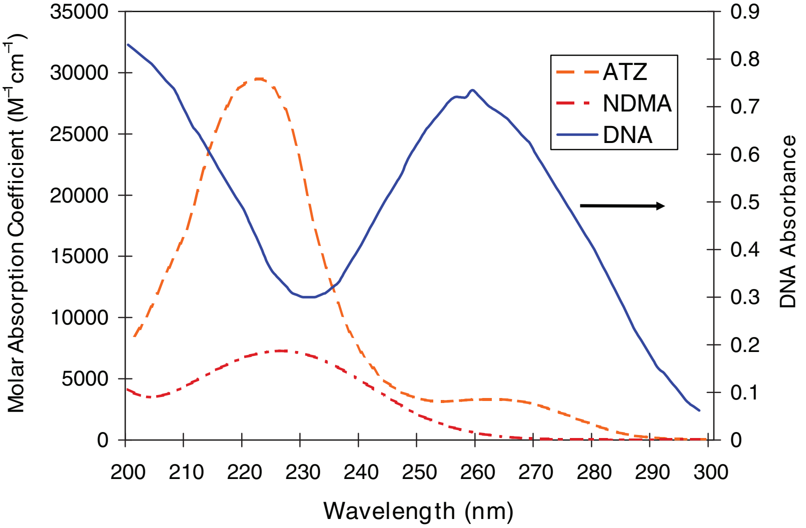


FIGURE 1 Absorbance spectra of chemical pollutants atrazine (ATZ) and N-nitrosodimethylamine (NDMA), compared with the absorbance features of DNA.

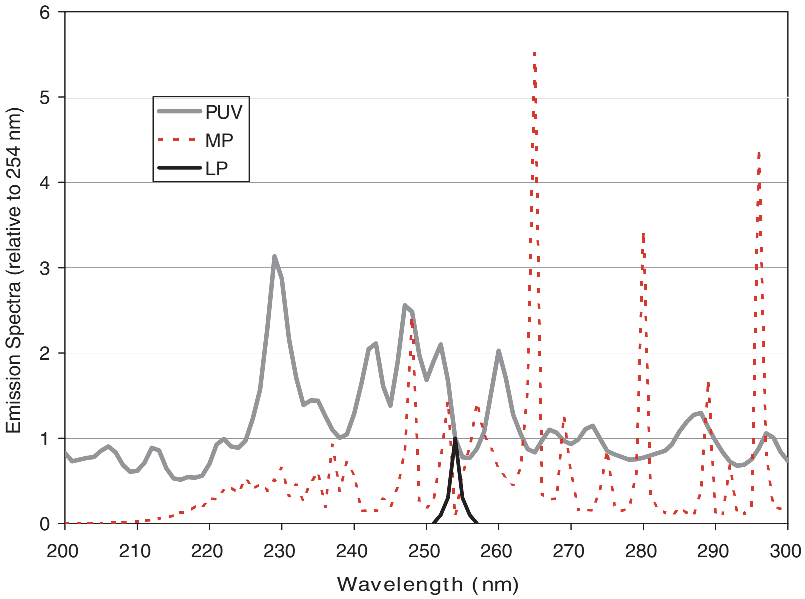


FIGURE 2 Emission spectra of conventional low-pressure (LP) and medium-pressure (MP) mercury vapor lamps, compared with a new surface discharge pulsed-UV source. (Pulsed source developed by Phoenix Science and Technology, Chelmsford, MA.)

change the engineering design of UV systems, offering more flexibility in UV reactor configuration.

UV FOR DISINFECTION OF PATHOGENS IN WATER

Interestingly, the efficacy of UV against pathogens varies greatly, despite the ubiquitous nature of the nucleic acid target. UV doses are measured in units of millijoules per centimeter squared (mJ/cm^2) or energy per unit area. The UV dose required for 99.99 percent inactivation of various pathogens is displayed in Figure 3. UV disinfection systems are commonly designed around a UV dose of 40 mJ/cm^2. Clearly, both bacterial and protozoan pathogens are very easy to inactivate compared with viruses and adenoviruses in particular. However, recent research shows that MP UV is much more effective than LP UV, specifically for some hard-to-disinfect viruses (Linden et al., 2007). It is hypothesized that MP can cause a diversity of damage to the viral pathogen, and because of the complexity of the infection process for this virus, multiple types of damage are required for inactivation.

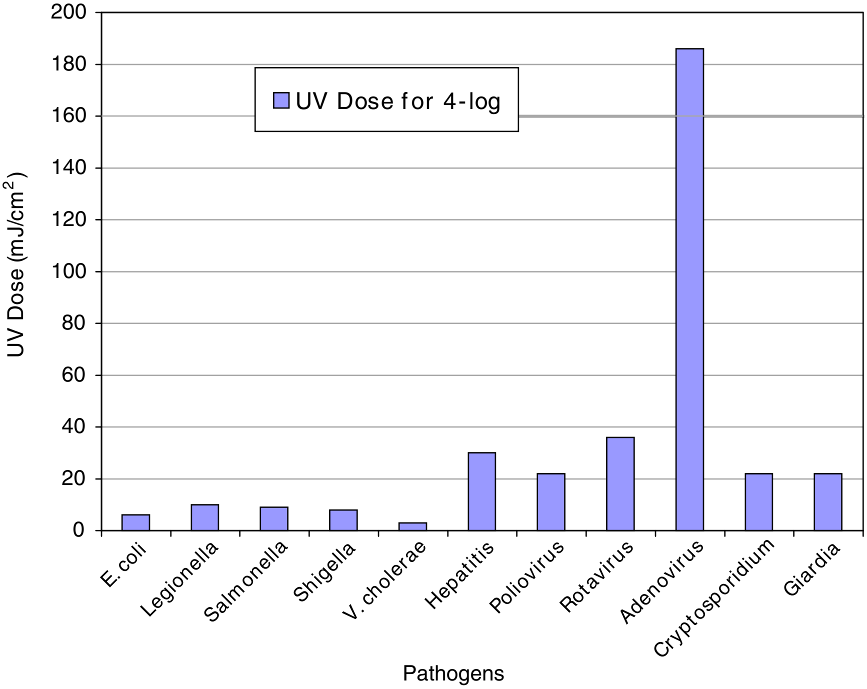


FIGURE 3 UV doses required for 4-log (99.99 percent) disinfection of various pathogens in water.

One caveat with disinfection is that many microbes have the ability to repair their UV-induced DNA damage through both light- and dark-based repair processes (Harm, 1980). Light-based repair involves an enzyme (photolyase) that gets activated during exposure to near UV and visible light wavelengths (350-450 nm). Dark repair typically involves excision repair and does not need light activation. These repair processes attack the pyrimidine dimers formed from UV exposure that leads to inactivation. At high enough doses (two to three times that needed for disinfection) the repair processes are not effective, and under typical drinking water conditions light-based repair is not a concern thanks to covered finished water reservoirs and the water distribution network.

UV FOR OXIDATION OF POLLUTANTS IN WATER

Another application for UV irradiation is the destruction of chemical pollutants in water (Oppenländer, 2003). This chemical destruction can take place through direct absorption of UV photons (photolysis) or indirectly by formation of highly reactive oxidative radicals. UV photolysis requires very high levels of UV energy, up to 100 times higher than what is required for disinfection. When hydrogen peroxide (H_2O_2) is added to water in the presence of UV photons in the germicidal range, the H_2O_2 is split into two hydroxyl radicals (•OH) by the photons and the radicals react rapidly (less than microseconds) with organic molecules in the water, thus degrading pollutants.

$$H_2O_2 + h\nu(UV_{200\text{-}300}) \Rightarrow 2{\bullet}OH$$

The UV-H_2O_2 process is termed an advanced oxidation process (AOP) because it accelerates the natural oxidation that occurs in the environment. It is a rapidly growing treatment technology and uses either LP or MP UV technologies or other UV sources with emissions overlapping that of H_2O_2 for destroying a variety of pollutants in water, including taste and odor-causing compounds, endocrine-disrupting compounds, pharmaceuticals, personal care products, and agricultural chemicals. Advanced oxidation processes for water treatment are a growing area of research because the public is getting more concerned about these emerging contaminants.

OTHER UV APPLICATIONS

Another area of growing interest is the use of UV irradiation for treatment of water in lesser developed countries. A number of student groups working through Engineers Without Borders and other groups have installed LP UV systems in Africa, Latin America, and Asia. These systems can be constructed with local materials and run using a low-power solar panel. Solar UV disinfection, using the sun's UV rays, is also an area of high interest among students: water is placed in discarded plastic bottles and exposed to sunlight over 4 to 6 hours to disinfect it by light- and heat-based processes. These applications illustrate the practicality of UV technologies for providing safe water.

THE FRONTIERS OF UV

Although UV itself is an emerging technology on the frontier of water treatment, there are many new engineering and scientific advances that continue to push and improve the technology and help develop a better understanding of the fundamentals of how UV works, leading to improved process design. Advances in UV lamp technologies coupled with concerns about mercury in UV lamps have led to lamp development, including the new pulsed-UV lamps with instant-on capabilities and radiation intensities up to 10,000 times greater than LP UV sources. UV-based LEDs are also emerging, as this technology is being pushed to lower wavelengths that are more germicidally effective. New technologies offer the possibility of radical changes in UV reactor design and improvements in effectiveness for water treatment, at lower energy costs, with safe materials for the environment. As UV technologies become accepted for drinking water, they are also being explored in water reuse applications, an important source of new water, as our freshwater resources are rapidly being either compromised or drained. Of great interest is the UV-AOP process in water reuse for combined disinfection and oxidation of contaminants. Advances in molecular biology and analytical chemistry have all contributed to furthering the knowledge of how UV works for improving water quality and have led to new developments in UV technologies. In summary, UV technology is another example of harnessing the power of nature for use by humans in our modern society. Natural UV from the sun has been harvested, engineered, and channeled to serve water-quality needs and produce safe water.

REFERENCES

Clancy, J. L., T. M. Hargy, M. Marshall, and J. Dyksen. 1998. Inactivation of oocysts of *Cryptosporidium parvum* in water using ultraviolet light. Journal of American Water Works Association 90:92-102.

Downes, A., and T. P. Blunt. 1977. Research on the effect of light upon bacteria and other organisms. Proceedings of the Royal Society of London 26:488-500.

EPA (U.S. Environmental Protection Agency). 1996. Fact Sheet-Key Provisions of Safe Drinking Water Act 1986 Amendments: Outlines Provisions of the 1986 Amendments to the 1974 Act, Safe Drinking Water Act Amendments of 1996—General Guide to Provisions [EPA 810-S-96-001]. Washington, DC: EPA.

EPA. 2006. Federal Register 71(3). Washington, DC: EPA.

Harm, W. 1980. Biological Effects of Ultraviolet Radiation. Cambridge, U.K.: Cambridge University Press.

Linden, K. G., J. Thurston, R. Schaefer, and J. P. Malley, Jr. 2007. Enhanced inactivation adenovirus types 2 and 40 under medium pressure and pulsed UV disinfection systems. Applied and Environmental Microbiology 73(23):7571-7574.

Linden, K. G., G. A. Shin, G. Faubert, W. Cairns, and M. D. Sobsey. 2002. UV disinfection of *Giardia lamblia* in water. Environmental Science and Technology 36:2519-2522.

Oppenländer, T. 2003. Photochemical Purification of Water and Air: Advanced Oxidation Processes (AOPs): Principles, Reaction Mechanisms, Reactor Concepts. New York: Wiley-VCH.

ADDITIONAL READING

Darby, J. L., K. E. Snider, and G. Tchobanoglous. 1993. Ultraviolet disinfection for wastewater reclamation and reuse subject to restrictive standards. Water Environment Research 65(2):169-180.

EPA (U.S. Environmental Protection Agency). 1986. Office of Research and Development. Design Manual: Municipal Wastewater Disinfection. EPA 625/1-86/021. Cincinnati, OH: EPA.

EPA. 2006. Ultraviolet Disinfection Guidance Manual. Office of Water, EPA 815-R-06-007. Washington, DC: EPA.

Jagger, J. H. 1967. Introduction to Research in UV Photobiology. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Linden, K. G., and J. L. Darby. 1997. Estimating effective germicidal dose from medium pressure UV lamps. ASCE Journal of Environmental Engineering 123(11):612-619.

Qualls, R. G., and J. D. Johnson. 1983. Bioassay and dose measurement in UV disinfection. Applied Environmental Microbiology 45:872-877.

Wolfe, R. L. 1990. Ultraviolet disinfection of potable water. Environmental Science and Technology 24(6):768-773.

Membrane Processes to Address the Global Challenge of Desalination

Amy E. Childress
University of Nevada, Reno

INCREASING SALINITY: A MAJOR ENVIRONMENTAL CONCERN OF THE 21ST CENTURY

Salinity refers to the presence of soluble salts in soils and water, including surface waters and groundwater. The main chemical species contributing to salinity are sodium, calcium, magnesium, chloride, carbonate, bicarbonate, sulfate, and silica. In the oceans, which contain 97 percent of Earth's water, salinity ranges from 35,000 mg/L to 41,000 mg/L of salts (usually referred to as total dissolved solids, or TDS). Inland waters have salinities ranging from less than 1,500 mg/L for freshwater to 1,500-20,000 mg/L for brackish water. The salinity in inland regions may be due to the presence of soils formed from marine sediments left behind by retreating seas, the weathering of rock minerals, or the deposition of oceanic salt onto the land by wind or rain. Some inland regions have naturally elevated salinity levels due to environmental features, such as mineral springs, carbonate deposits, salt deposits, and seawater intrusion. Other inland regions have anthropogenically elevated salinity levels. Like many environmental degradation processes, increasing salinity is exacerbated by human activities. Urban runoff, land use practices (clearing, irrigation, industrial farming, overextraction), septic systems, chemicals used in water treatment processes, and road de-icing can substantially increase the extent of the problem. Unlike many other environmental contaminants, salts do not naturally degrade over time. Salts that are deposited (naturally or anthropogenically) into a body of water accumulate until they are intentionally removed.

Increasing salinity is one of the most important environmental issues of the 21st century. Rising salinity levels have both environmental and economic costs: reduction in agricultural productivity, loss of biodiversity in both terrestrial and aquatic ecosystems, and decline in water quality for drinking, irrigation, industrial reuse, and recreation purposes (Tsiourtis, 2001; Ettouney et al., 2002; USBOR and SNL, 2003).

In terms of drinking water treatment, the need for additional potable water supplies has led to increased numbers of seawater and brackish water desalination facilities being constructed worldwide. To be fit for human consumption drinking water should contain less than 1,500 mg/L of salts. In the United States the Environmental Protection Agency has set a secondary standard for TDS in drinking water at 500 mg/L. In terms of wastewater treatment, TDS discharge standards vary for different facilities but are consistently becoming more stringent for all facilities. This has substantial implications for conventional wastewater treatment facilities that were not designed for salinity removal. Many facilities must retrofit with additional processes specifically for TDS removal. Thus, desalination has developed from a concept related to seawater applications in coastal regions to a concept applying to a range of applications across all regions. Furthermore, as concerns about water scarcity and quality continue to grow, the research and engineering of economic and environmentally friendly desalination systems will continue to grow at a tremendous rate.

DESALINATION: FROM ITS ANCIENT ROOTS TO ITS CURRENT PROMINENCE

The roots of desalination can be traced back to the fourth century BC, when Aristotle and Hippocrates described the process of evaporating saltwater to produce freshwater (Service, 2006). The first actual practice of desalination, in AD 200, is attributed to sailors who distilled seawater by collecting freshwater steam from the boilers on their ships (Schirber, 2007). Modern seawater desalination began in the early 20th century; the first distillation facility was put into operation on the island of Curacao in the Netherlands Antilles in 1928. Modern reverse osmosis (RO) began in the 1950s when Reid and Breton (1959) found that a cellulose acetate film rejected sodium chloride from aqueous solution, and Loeb and Sourirajan (1962) developed a high flux modified cellulose acetate membrane. The first commercial RO facility began operation in 1965 in Coalinga, California.

Today some form of desalination is used in approximately 130 countries; as of 2005 more than 10,000 desalting units larger than 0.3 million gallons per day (MGD) (nominal) had been installed or contracted worldwide (Cooley et al., 2006). The two processes at the core of conventional desalination facilities continue to be distillation and RO. RO generally has lower capital costs and requires less energy than distillation. A typical seawater RO plant requires 1.5-2.5 kilowatt-hours (kWh) of electricity to produce 1 m^3 of water (Service, 2006), and a typical distillation plant requires 10 times that amount (Ettouney et al., 2002; Cooley et al., 2006). On the other hand, distillation typically produces water with much lower salt content than RO systems. Distillation systems typically produce water with less than 25 mg/L, and RO systems typically produce water with less than 500 mg/L (Cooley et al., 2006; Service, 2006).

Although substantial improvements have been made to RO and distillation processes in the past decades, there is a strong need for improved desalination systems, especially systems that will achieve a greater rate of production of treated water at lower energy expenditure and cost. Several systems being investigated include dewvaporation, capacitive deionization, chemical precipitation, membrane distillation (MD), and forward osmosis (FO). This paper focuses on the viability, efficiency, and practical application of MD and FO for brackish water, seawater, and brine desalination at low energy expenditure and with minimal impact on the environment.

MEMBRANE DISTILLATION AND FORWARD OSMOSIS: INNOVATIVE TECHNOLOGIES FOR DESALINATION APPLICATIONS

MD is a temperature-driven separation process involving the transport of mass and heat through a hydrophobic, microporous membrane. The driving force for mass transfer in direct contact MD (DCMD) is the vapor pressure gradient across the membrane. In a configuration referred to as vacuum-enhanced DCMD, warmer feed water is in contact with the feed side of the membrane and a cooler water stream, under vacuum, is in direct contact with the opposite side of the membrane (Figure 1). The temperature difference between the streams and the vacuum applied on the permeate side of the membrane induce the vapor pressure gradient for mass transfer. Compared with conventional distillation methods, MD requires only small temperature differences—temperature differences achievable through the use of low-grade or waste heat sources. Compared with RO, the driving force in MD is not reduced by osmotic pressure (Figure 2) and thus, MD can provide enhanced recovery through brine desalination (Cath et al., 2004). For these reasons MD may have substantial energy and recovery advantages. MD has shown to be an effective process for desalination of brackish water and seawater (Lawson and Lloyd, 1997; El-Bourawi et al., 2006), removal of urea and endocrine-disrupting chemicals (Cartinella et al., 2006), and concentration of brines (Martinez-Diez and Florido-Diez, 2001; Childress and Cath, 2006).

FO is an osmotically driven separation process involving the diffusion of water through a semipermeable membrane. Water diffuses from a stream of low solute concentration to a draw solution (DS) stream of high osmotic pressure (Figure 3). The driving force for mass transport is the difference in osmotic pres-

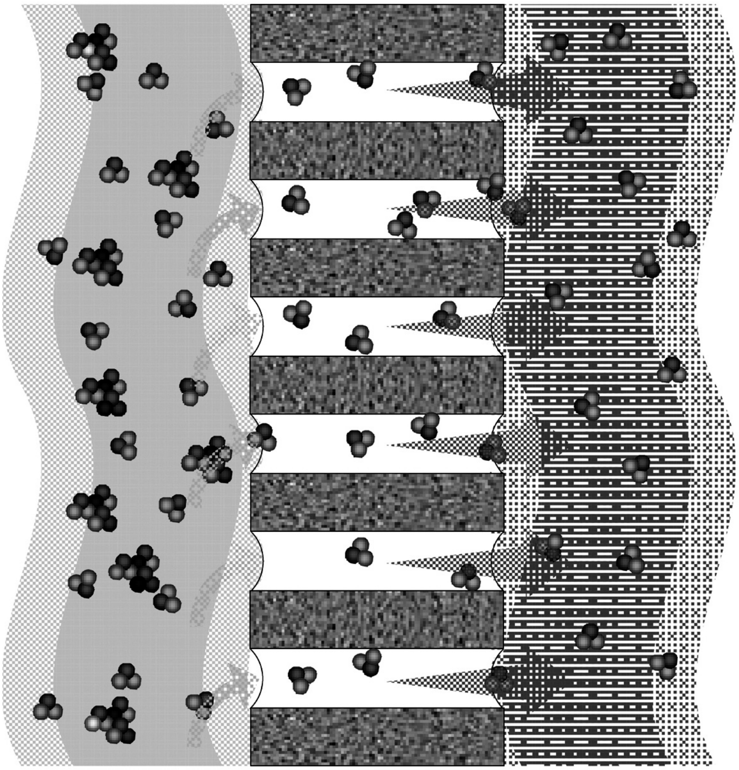


FIGURE 1 Direct contact membrane distillation.

sure between the DS and the feed solution. The main advantages of using FO are that it operates at low or no hydraulic pressures, it has high rejection of a wide range of contaminants, and it may have a lower membrane-fouling propensity than pressure-driven membrane processes (Cath et al., 2006). Because the only pressure involved in the FO process is due to flow resistance in the membrane module (a few bars), the equipment used is very simple and membrane support is less of a problem (Cath et al., 2006). FO has been shown to be an effective pretreatment process for RO (Figure 4) and has been used to desalinate feed streams ranging from brackish water to brine (McCutcheon et al., 2005; Martinetti et al., 2007.

DIRECT POTABLE REUSE: THE FRONTIER

MD and FO have been investigated together in a hybrid system designed for direct potable reuse in long-term space missions. Long-term human missions in space require a continuous and self-sufficient supply of freshwater for consumption, hygiene, and maintenance. Long-range or long-duration missions, like

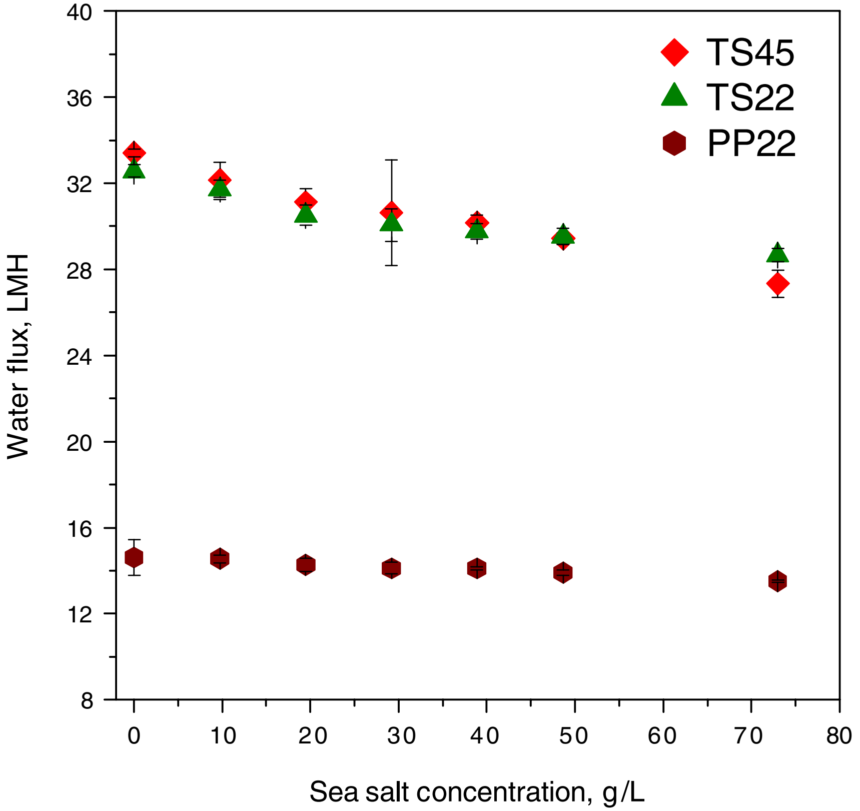


FIGURE 2 Flux as a function of seawater concentration in DCMD. Source: Cath et al., 2004. Reprinted with permission from Elsevier.

lunar missions or a human Mars exploration mission, depend on a water treatment system that recovers potable water from wastewater generated on board a spacecraft or in the planetary habitat. The three main sources of wastewater that can be reclaimed and reused in long-term space missions are hygiene wastewater, urine, and humidity condensate. The system to treat these wastewaters must be reliable, durable, capable of recovering a high percentage of the wastewater, and lightweight. Additionally, the system should operate autonomously with low maintenance, minimum power consumption, and minimal consumables (Wieland, 1994).

A pilot system referred to as the direct osmotic concentration (DOC) is one of several technologies that are being evaluated by the National Aeronautics and Space Administration (NASA) (Beaudry and Herron, 1997; Beaudry et al., 1999) to reclaim wastewater. In the original DOC concept three different membrane pro-

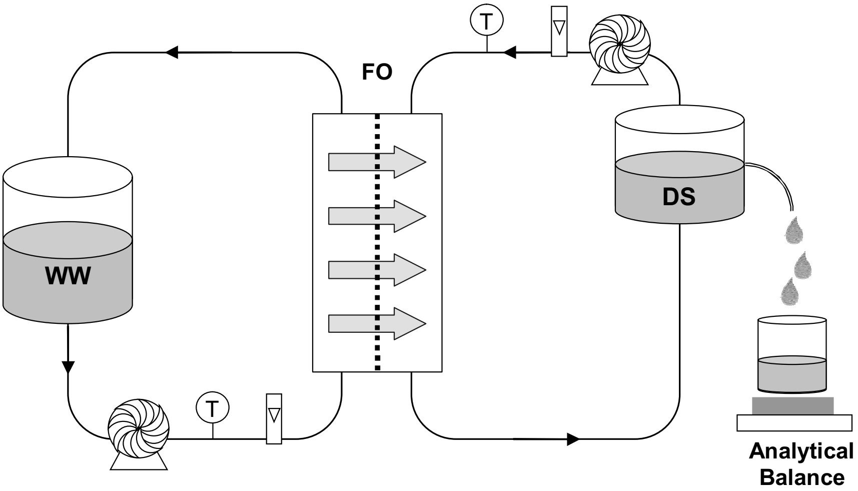


FIGURE 3 Schematics of forward osmosis concentration process. Source: Holloway et al., 2007. Reprinted with permission from Elsevier.

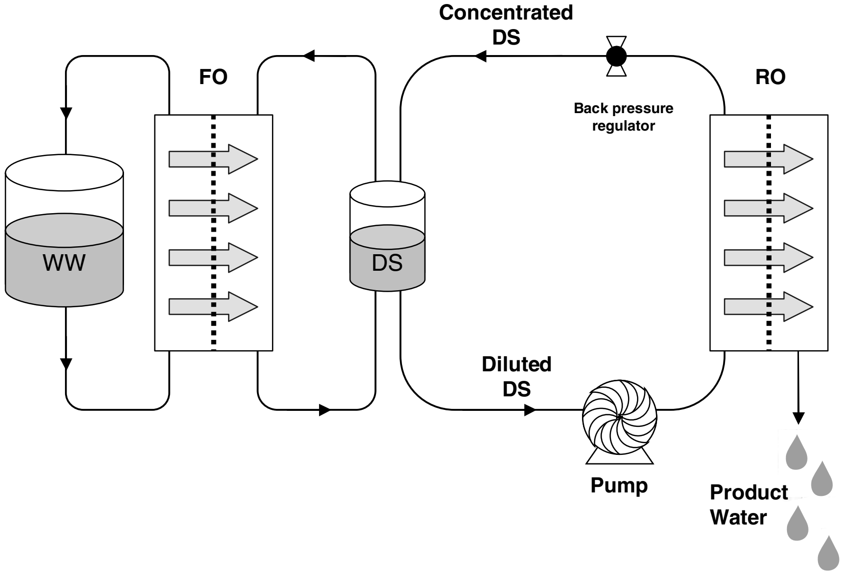


FIGURE 4 Combined RO-FO system for pretreatment before RO desalination. Source: Holloway et al., 2007. Reprinted with permission from Elsevier.

cesses (RO, FO, and osmotic distillation [OD]) were integrated into one system. Due to poor performance of the OD process the system design was revised. In the new design (Figure 5) DCMD is used to treat urine and to pretreat humidity condensate; FO is being used as pretreatment for RO for the hygiene wastewater. Separation of the feed streams in this hybrid system prevents the surfactants from hydrophilizing the hydrophobic MD membrane. This is critical because the MD membrane is required to remove the urea that would pass through an RO membrane. One important aspect of this investigation is the removal of trace contaminants. During long space missions, crewmembers will consume water that is continuously recycled. Thus, it is important to ensure that trace contaminants and particularly endocrine-disrupting chemicals (EDCs) be removed from the treated water. EDCs (e.g., estrone and estradiol) are excreted in urine by both males and females and conventional wastewater treatment is often insufficient for removing them (Belfroid et al., 1999; Johnson et al., 2000; Johnson and Sumpter, 2001).

Preliminary testing of the new DOC concept revealed that DCMD achieves greater than 99 percent estrone and estradiol removal and greater than 99.9 percent urea removal. Furthermore, the FO membrane provides an economical means to pretreat the hygiene wastewater and protects the RO membrane from excessive fouling. These results confirm that MD and FO processes can be combined with RO in a hybrid membrane configuration to treat complex liquid streams that cannot be treated with the individual processes. These results are not only pertinent to NASA's DOC system but they also have implications for terrestrial wastewater treatment applications.

Terrestrially, much less emphasis is currently being placed on wastewater reclamation than on seawater desalination for direct potable water treatment.

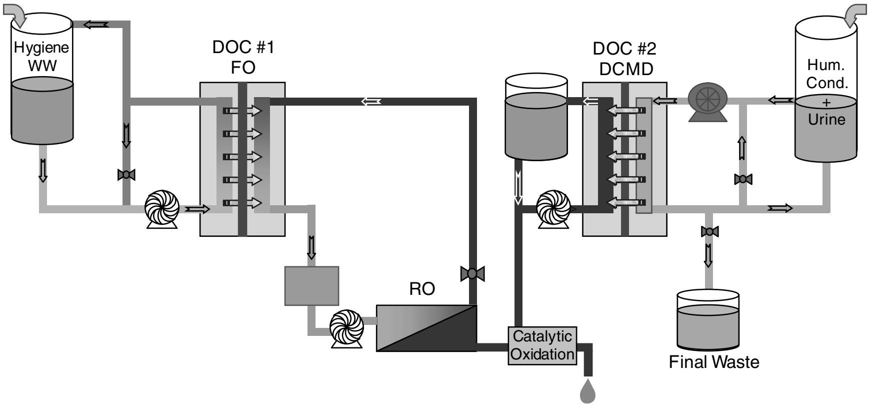


FIGURE 5 NASA DOC system.

However, wastewater reclamation costs are only a fraction of seawater desalination costs. Furthermore, wastewater (sewage) is the only guaranteed abundant source of water wherever there are people. Thus, for environmental sustainability the major barriers to direct potable reuse must be overcome. Ironically most of these barriers relate to public acceptability rather than technical capability. Further studies on systems such as the DOC will bring the concept of direct potable reuse closer to a reality.

SUSTAINABILITY: A GOAL FOR DEVELOPED AND DEVELOPING COUNTRIES

According to an Association of Environmental Engineering and Science Professors Research Frontiers Conference summary (Dentel and Cannon, 1999), there are five research directions that must be pursued in order to meet the demands of a sustainable environment. One of these directions, “Targeted or Tailored Separation and Destruction Systems,” includes membrane processes. More specifically, the summary refers to the development of membrane applications that employ new advances in membrane chemistry, membrane physical properties, and flow configurations. Furthermore, the Desalination and Water Purification Technology Roadmap has identified membrane technologies as an area of research necessary to develop cost-effective technological tools that can be used to help solve the nation’s water supply challenges. The research areas identified in the roadmap were selected to both speed the evolution of existing (current-generation) desalination and water purification technologies and lay the scientific and technical foundation for the development of advanced next-generation technologies (USBOR and SNL, 2003).

Water reuse and low energy consumption are key objectives to achieving sustainability in water and wastewater treatment. Because energy and water are inextricably bound, energy usage and production must be considered when evaluating water and wastewater treatment processes. Thus, a major objective of all desalination investigations should be to reduce energy costs and to explore opportunities for energy production. Consideration of these objectives has led to the development of desalination systems powered by renewable energy. One of the more well-studied of these systems is solar-powered MD (Hogan et al., 1991; Banat et al., 2002). In a similar type of application the combination of MD with a solar pond has also been considered (Walton et al., 1999). A major goal is to make these systems economically viable for remote regions.

CONCLUSION

MD and FO offer the potential for substantial energy and resource savings over conventional desalination processes. Use of these technologies may substantially mitigate the environmental impacts associated with current desalination practices. For developing countries a combination of these technologies with renewable energy may enable potable water production in remote, arid locations. For developed countries a combination of these technologies with existing desalination practices may bring the goal of direct potable reuse closer to a reality and make the goal of zero liquid discharge more cost-efficient.

REFERENCES

Banat, F. A., R. Jumah, and M. Garaibeh. 2002. Exploitation of solar energy collected by solar stills for desalination by membrane distillation. Renewable Energy 25(2):293-305.

Beaudry, E. G., and J. R. Herron. 1997. Direct osmosis for concentrating wastewater. Proceedings of the 27th International Conference on Environmental Systems (ICES). Lake Tahoe, NV, July 14-17.

Beaudry, E. G., J. R. Herron, and S. W. Peterson. 1999. Direct osmosis concentration of waste water—Final report. Corvallis, OR: Osmotek Inc.

Belfroid, A. C., A. V. D. Horst, A. D. Vethaak, A. J. Schäfer, G. B. J. Rijs, J. Wegener, and W. P. Cofino. 1999. Analysis and occurrence of estrogenic hormones and their glucuronides in surface water and waste water in the Netherlands. Science of the Total Environment 225(1-2):101-108.

Cartinella, J. L., T. Y. Cath, M. T. Flynn, G. C. Miller, K. W. Hunter and A. E. Childress. 2006. Removal of natural steroid hormones from wastewater using membrane contactor processes. Environmental Science and Technology 40:7381-7386.

Cath, T. Y., V. D. Adams, and A. E. Childress. 2004. Experimental study of desalination using direct contact membrane distillation: A new approach to flux enhancement. Journal of Membrane Science 228(1):5-16.

Cath, T. Y., A. E. Childress, and M. Elimelech. 2006. Forward osmosis: Principles, applications, and recent developments. Journal of Membrane Science 281(1-2):70-87.

Childress, A. E., and T. Y. Cath. 2006. Direct contact membrane distillation and forward osmosis for the concentration of primary and secondary RO brine—Final Report. Reno, NV: University of Nevada, Reno.

Cooley, H., P. H. Gleick, and G. Wolff. 2006. Desalination, With a Grain of Salt. Oakland, CA: Pacific Institute for Studies in Development, Environment, and Security.

Dentel, S., and F. Cannon. 1999. Speech at the Association of Environmental Engineering and Science Professors Research Frontiers Conference, University Park, PA, July 31-August 3.

El-Bourawi, M. S., Z. Ding, R. Ma, and M. Khayet. 2006. A framework for better understanding membrane distillation separation process. Journal of Membrane Science 285(1-2):4-29.

Ettouney, H. M., H. T. El-Dessouky, R. S. Faibish, and P. J. Gowin. 2002. Evaluating the economics of desalination. Chemical Engineering Progress 98(12):32-39.

Hogan, P. A., A. Sudjito, G. Fane, and G. L. Morrison. 1991. Desalination by solar heated membrane distillation. Desalination 81(1-3):81-90.

Holloway, R. W., A. E. Childress, K. E. Dennett, and T. Y. Cath. 2007. Forward osmosis for concentration of anaerobic digester centrate. Water Research 41(17):4005-4014.

Johnson, A. C., and J. P. Sumpter. 2001. Removal of endocrine-disrupting chemicals in activated sluge treatment works. Environmental Science and Technology 35:4697-4703.

Johnson, A. C., A. Belfroid, and A. D. Corcia. 2000. Estimating steroid estrogen inputs into activated sludge treatment works and observations on their removal from the effluent. Science of The Total Environment 256(2-3):163-173.

Lawson, K. W., and D. R. Lloyd. 1997. Review: Membrane distillation. Journal of Membrane Science 124(1):1-25.

Loeb, S., and S. Sourirajan. 1962. Sea water demineralization by means of an osmotic membrane. Advances in Chemistry Series 38:117-132.

Martinetti, C. R., T. Y. Cath, and A. E. Childress. 2007. Novel application of membrane distillation and forward osmosis for brackish water desalination. Paper presented at the 2007 AWWA Membrane Technology Conference and Exposition, Tampa, FL, March 18-21.

Martinez-Diez, L., and F. J. Florido-Diez. 2001. Desalination of brines by membrane distillation. Desalination 137:267-273.

McCutcheon, J. R., R. L. McGinnis, and M. Elimelech. 2005. Desalination by a novel ammonia-carbon dioxide forward osmosis process: Influence of draw and feed solution concentrations on process performance. Journal of Membrane Science 278(1-2):114-123.

Reid, C. E., and E. J. Breton. 1959. Water and ion flow across cellulosic membranes. Journal of Applied Polymer Science 1(2):133-143.

Schirber, M. 2007. Why desalination doesn't work (yet). Online. LiveScience. Available at <http://www.livescience.com/environment/070625_desalination_membranes.html>. Posted on June 25.

Service, R. F. 2006. Desalination freshens up. Science 313:1088-1090.

Tsiourtis, N. X. 2001. Desalination and the environment. Desalination 138(1-3):1.

USBOR (United States Bureau of Reclamation) and SNL (Sandia National Laboratory). 2003. Desalination and Water Purification Technology Roadmap—A Report of the Executive Committee. Denver, CO: Bureau of Reclamation.

Walton, J., S. Solis, H. Lu, C. Turner, and H. Hein. 1999. Membrane distillation desalination with a salinity gradient solar pond. Online. Available at <http://www.johncwalton.com/Projects/1999MembraneTechConf/1999-17thAnnualMembrneTechConf.htm>.

Wieland, P. O. 1994. Design for Human Presence in Space: An Introduction to Environmental Control and Life Support Systems. George C. Marshall Space Flight Center, AL: NASA.

Biological Treatments of Drinking Water

JESS C. BROWN
Carollo Engineers, P.C.
Phoenix, Arizona

Microbial biomass has been used since the early 1900s to degrade contaminants, nutrients, and organics in wastewater. Until recently, the biological treatment of drinking water was limited, particularly in the United States, but recent developments may mean that biological drinking water treatment may become more feasible and more likely to be accepted by the public. These developments include (1) the rising costs and increasing complexities of handling water treatment residuals (e.g., membrane concentrates); (2) the emergence of new contaminants that are particularly amenable to biological degradation (e.g., perchlorate); (3) the push for green technologies (i.e., processes that efficiently destroy contaminants instead of concentrating them); (4) regulations limiting the formation of disinfection by-products (DBPs); and (5) the emergence of membrane-based treatment systems, which are highly susceptible to biological fouling.

PROCESS FUNDAMENTALS

Bacteria gain energy and reproduce by mediating the transfer of electrons from reduced compounds (i.e., compounds that readily donate electrons) to oxidized compounds (i.e., compounds that readily accept electrons). Once electrons are donated by a reduced compound, they travel back and forth across a cell's mitochondrial membrane in a series of internal oxidation-reduction reactions. Ultimately, the electrons are donated to the terminal electron-accepting compound. This series of reactions, which is cumulatively known as the electron-transport chain, creates an electrochemical gradient across the cell membrane that bacteria use to generate adenosine triphosphate, also known as energy (Madigan et al., 1997).

As compounds gain or lose electrons, they are converted to different, often innocuous, forms that are thermodynamically more stable than the original compounds. The example below illustrates the microbially mediated oxidation-reduction reaction between acetate (an electron donor) and dissolved oxygen and nitrate (two environmental electron acceptors).

- $CH_3COO^- + 2O_2 \rightarrow 2HCO_3^- + H^+ \; \Delta G^{o\prime} = -844$ KJ/mol acetate
- $CH3COO^- + {}^{3}/_{5}NO_3^- + {}^{13}/_{5}H+ \rightarrow 2HCO_3^- + {}^{4}/_{5}H_2O + {}^{4}/_{5} N_2$
  $\Delta G^{o\prime} = -792$ KJ/mol acetate

Notice that nitrate, a common contaminant in drinking water, is converted to innocuous nitrogen gas. The Gibb's free-energy values for the overall reactions are shown below the equations (Rikken et al., 1996). The more negative the Gibb's free-energy value, the more thermodynamically unstable the reaction and the greater the energy yield for the bacteria mediating the reaction. Electron transfer in the overall reactions can be observed only by evaluating the oxidation states of individual atoms.

Biological drinking water treatment processes are based on the growth of bacterial communities capable of mediating oxidation-reduction reactions involving at least one target contaminant (Figure 1). Heterotrophic biological processes

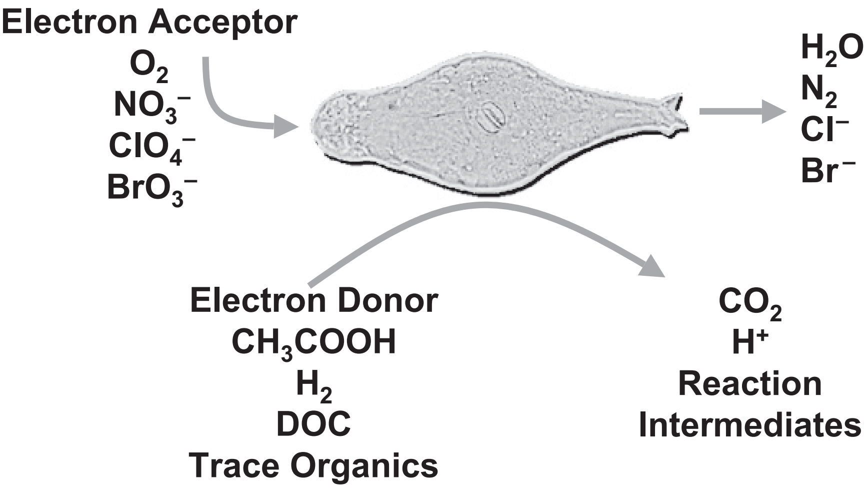


FIGURE 1 Microbially mediated oxidation-reduction reactions.

use an organic electron donor (e.g., acetic acid). Autotrophic biological processes use an inorganic electron donor (e.g., hydrogen).

CONTAMINANT APPLICATIONS

Biological processes can be used for a wide range of organic and inorganic contaminants (Table 1) in both surface water and groundwater.

TECHNOLOGY CONFIGURATIONS

Numerous forms and configurations of biological treatment processes are used to degrade contaminants in drinking water. Most are operated as fixed biofilm systems, meaning that the process includes a biogrowth support medium on which bacterial communities attach and grow (e.g., granular media). A smaller number of technologies operate as suspended growth systems, in which free-floating bacteria are hydraulically maintained within a reactor. Biological reactors can be inoculated with an enriched bacterial community or can simply be acclimated by the organisms indigenous to the water source being treated. Examples of biological treatment configurations are described below.

FIXED-BED PROCESSES

In fixed-bed (FXB) biological processes, biofilms develop on a stationary bed of media, such as sand, plastic, or granular activated carbon (Figure 2). The granular media bed can be contained in pressure vessels or open basins. In pressure-vessel systems, water is pumped up-flow or down-flow across the biological bed; in open-basin systems up-flow requires pumping, but down-flow occurs by gravity. As water is treated, biofilms increasingly restrict flow and cause head loss across the bed. If unchecked, the loss eventually exceeds the available driving pressure or causes short-circuiting through the bed. To avoid these complications, FXB systems are routinely taken off line and backwashed to remove excess biomass from the system (Brown et al., 2005; Kim and Logan, 2000). FXB is often coupled with pre-ozonation to improve the removal of organic material, which reduces regrowth potential and DBP formation in distribution systems.

FLUIDIZED-BED PROCESSES

Fluidized-bed reactors (FBRs) also use granular media to support biogrowth. Contaminated water is pumped up-flow through the reactor at a high rate to fluidize the granular media bed and reduce resistance to flow. Typically, the fluidization rate is controlled to maintain a 25 to 30 percent bed expansion over the resting bed height. Feed flow is supplemented with recycle flow to provide the appropriate up-flow velocity for fluidization (Green and Pitre, 1999; Guarini and

TABLE 1 Contaminants Amenable to Biological Treatment[1]

Contaminant Category	Removal Application	Description
Natural organic matter (NOM)	• Regrowth substrate • DBP precursors • Color • Membrane foulants	• The biological oxidation of carbonaceous organic matter to CO_2 can minimize potential regrowth in distribution systems, decrease the production of DBPs, remove color, and improve transmembrane fluxes without chemical additives. Ozone is often used before a biological process to increase removal of NOM.
Trace organics	• 2-methyl-isoborneol (MIB) • Geosmin • Algal toxins • Endocrine disruptors and pharmaceutically active compounds • Pesticides	• Biological oxidation to CO_2, often degraded as a secondary electron donor (i.e., does not yield the requisite energy to support cell maintenance and growth), requires the presence of a primary substrate, such as NOM.
	• Methyl tertiary-butyl ether (MTBE)	• Biological oxidation to CO_2.
Halogenated organics	• Perchloroethylene (PCE) • Trichloroethylene (TCE) • Dibromochloro-propane (DBCP) • Chloroform	• Biological reductive dechlorination produces innocuous ethane or CO_2.
Inorganics	• Perchlorate • Chlorate • Nitrate • Nitrite • Bromate	• Biological reduction produces innocuous end products (Cl^-, N_2, Br^-, H_2O), thus eliminating the generation of a contaminated concentrate stream.
	• Selenate • Chromate	• Biological reduction produces insoluble species that can be readily filtered or settled out of water, thus eliminating the need for chemical reduction.
	• Ammonia	• Biological oxidation of ammonia to nitrate provides an alternative to chemically intensive break-point chlorination.
	• Iron • Manganese	• Biological oxidation of soluble species (Fe^{2+}, Mn^{2+}) to insoluble species (Fe^{3+}, Mn^{4+}) eliminates the need for chemical oxidation prior to filtration or settling.

[1] Source: Adapted from personal experience and Bouwer and Crowe, 1988; Brown, 2006; Brown et al., 2005; Dahab and Woodbury, 1998; Herman and Frankenberger Jr., 1999; Kirisits et al., 2002; and Lauderdale et al., 2006.

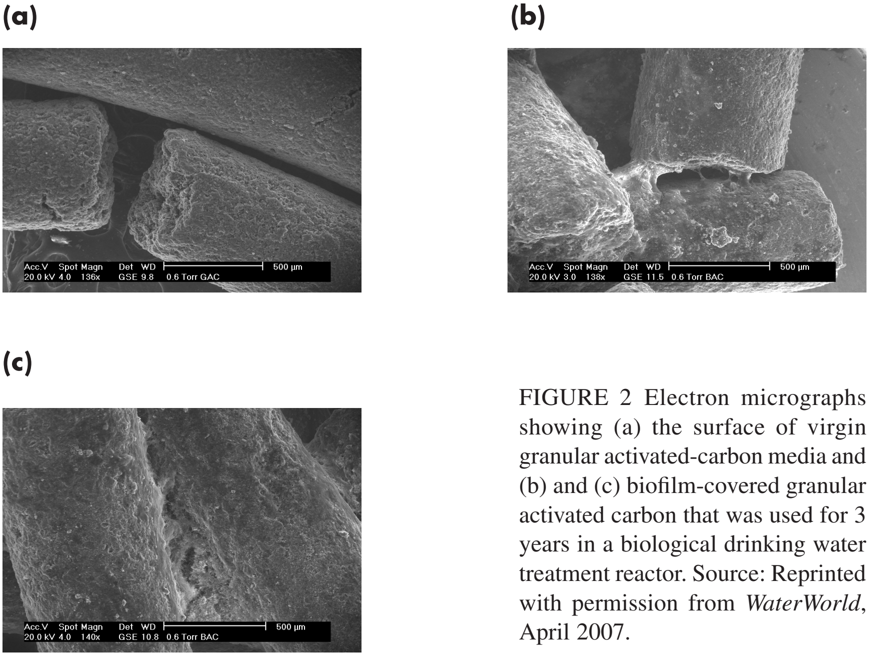

(a)
Acc.V Spot Magn Det WD 500 µm
20.0 kV 4.0 136x GSE 9.8 0.6 Torr GAC
(b)
Acc.V Spot Magn Det WD 500 µm
20.0 kV 3.0 138x GSE 11.5 0.6 Torr BAC
(c)
Acc.V Spot Magn Det WD 500 µm
20.0 kV 4.0 140x GSE 10.8 0.6 Torr BAC

FIGURE 2 Electron micrographs showing (a) the surface of virgin granular activated-carbon media and (b) and (c) biofilm-covered granular activated carbon that was used for 3 years in a biological drinking water treatment reactor. Source: Reprinted with permission from *WaterWorld*, April 2007.

Webster, 2004). Excess biomass is removed from FBR systems by (1) shear forces generated by the high feed-pumping rates and/or (2) in-line mechanical shearing devices. Therefore, although FBRs require higher feed-flow capacity, they do not require an off-line backwashing step.

MEMBRANE BIOREACTORS

Conventional Systems

Membranes can also be coupled with biological systems to improve the treatment of drinking water. In one approach, ultrafiltration membranes are submerged in a reactor basin that contains suspended biomass. The reactor basin provides the detention time necessary to achieve effective biological treatment. Treated water is drawn through the membranes by vacuum and pumped out to permeate pumps for further processing.

Airflow introduced at the bottom of the reactor basin performs several functions. First, it creates turbulence that scrubs and cleans the outside of the membranes and reduces the accumulation of solids on the membrane surface. Thus the membrane can operate for extended periods of time at high permeate fluxes. Second, the air has the beneficial side effect of oxidizing iron, manganese, and some organic compounds that may be present. Third, air ensures mixing in the process tank to maintain suspension of the biomass. Periodic backwashing of the membranes is done by passing permeate through the membranes in the reverse direction to dislodge solids from the membrane surface.

Biofilm Reactor Systems

A different approach to the "conventional" membrane bioreactor (MBR) uses hollow-fiber membranes to deliver hydrogen gas (an electron donor) to biofilms that grow on the outside of the fibers. When the hollow-fiber membranes are submerged in a reactor vessel through which contaminated water passes, contaminants diffuse from the bulk water into the biofilms and are degraded (Nerenburg et al., 2002). Occasionally, the membranes are chemically cleaned to remove excess biomass.

Ion-Exchange Membrane Systems

Yet another MBR method involves a reactor with two treatment chambers separated by an ion-exchange membrane. One chamber contains suspended biomass plus nutrients; the other chamber contains raw water. As raw water enters the system and moves through one chamber, ionic contaminants diffuse across the membrane into the biological treatment chamber where they are degraded. The objective of this approach is to separate the active biomass from the raw and treated water (Liu and Batista, 2000).

BANK FILTRATION SYSTEMS

Bank filtration wells, drilled near rivers and lakes, draw surface water through soil and aquifer material, which act as a passive treatment reactor. As the surface water moves through the aquifer, it is subject to filtration, dilution, sorption, and biodegradation processes (Gollnitz et al., 2003; Ray et al., 2002; Weiss et al., 2003a,b). Bank filtration, which has been used for more than 130 years in Europe, has aroused a great deal of global interest for use in reducing organic and particulate loads to drinking water treatment plants.

One of the oldest bank filtration systems draws water from the Rhine River in Germany and is part of the Düsseldorf Waterworks. This system has been in operation since 1870, and until about 1950, it was the only treatment process used at that facility.

PROCESS OPTIMIZATION

Various tools are available to facilitate the optimization of engineered biological treatment systems. Bench-scale reactors (Figure 3) and pilot-scale reactors (Figure 4) are often used in conjunction with mathematical models to isolate the impacts of various water-quality conditions and operating parameters on overall system performance.

Commercial Models

Available commercial models can be tailored to a specific treatment application and process configuration. Typically calibrated using results from bench- and/or pilot-scale testing, these models can simulate steady-state or dynamic conditions and account for hydraulic-flow regimes from plug-flow to complete mixing.

The models incorporate a wide range of parameters, such as feed-water quality and temperature, substrate and nutrient loading rates, contact time, biofilm thickness, specific surface area of reactor media, and biomass detachment. Not only can they predict bioreactor performance for a given set of environ-

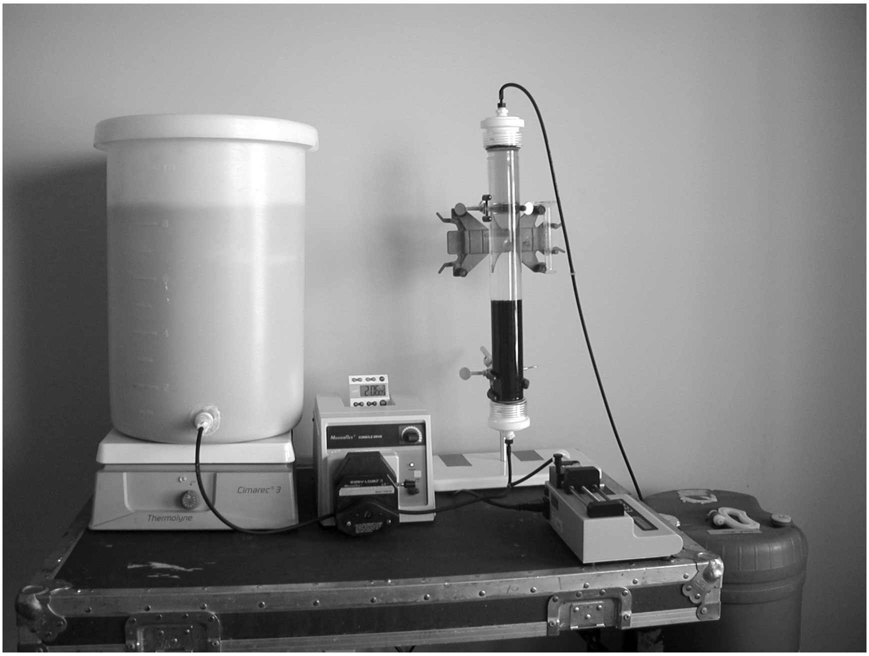


FIGURE 3 Bench-scale biological-treatment testing apparatus.

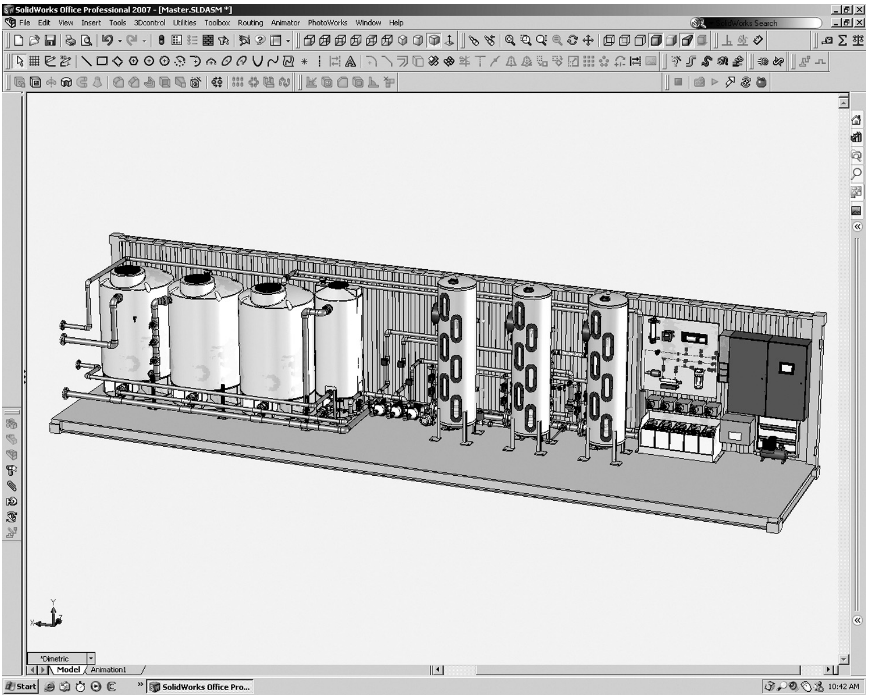


FIGURE 4 Illustration of a pilot-scale biological-treatment testing apparatus.

mental conditions, they can also elucidate observed phenomena in bioreactor systems. In other words, they can eliminate the “black box” perception of bioreactor processes.

Culture-Based Microbiological Analyses

Microbiological analyses provide another optimization tool. Using a targeted nutrient medium in conjunction with specific incubation conditions, pure cultures can be isolated from the mixed community of bacteria comprising a bioreactor. An enrichment of each pure culture can then be tested to identify optimal environmental conditions for that classification of bacteria. The information can then be used to tweak the operation or nutrient loading to a given bioreactor to favor the activity and growth of key contaminant-degrading microorganisms.

Molecular Microbiological Techniques

A complement to culture-based techniques, molecular microbiological techniques can be used to identify, quantify, locate, and track specific classes, families, genera, or species of bacteria. These techniques rely on the extraction, amplification, and sequencing (i.e., order of nucleic acid bases A, T, C, and G) of bacterial community DNA.

Once DNA sequences have been identified for a mixed community, they can be compared against large libraries of known bacterial DNA sequences to identify specific bacteria in a given treatment system to provide a "fingerprint" of the mixed microbial community. Nucleic acid probes, which are constructed using DNA sequence data to target specific bacteria, can then be used to quantify and track changes in a microbial community's fingerprint as a function of operational conditions or water-quality characteristics.

As environmental conditions change, the composition of a microbial community also changes. Molecular microbiological techniques can rapidly identify the environmental conditions that favor the growth and activity of the key contaminant-degrading bacteria in a bioreactor.

SUMMARY

Given that a key objective of treating drinking water is the inactivation or removal of microorganisms from raw water, using bacteria to help produce potable water would seem to fly in the face of conventional wisdom. However, biological drinking water treatment processes, which use indigenous, nonpathogenic bacteria, are always followed by downstream processes, such as final disinfection. Consequently, well-designed biological treatment systems pose no significant, inherent threats to the health or safety of distributed water. On the contrary, they can often provide an alternative to conventional processes that has several potential advantages:

- low operating costs
- high water-recovery rates
- destruction, rather than sequestration or concentration, of contaminants
- simultaneous removal of multiple contaminants
- minimal sludge production
- no hazardous waste streams
- minimal or no added chemicals
- robustness over a wide range of operating conditions and water qualities

Overall, biological drinking water treatment is highly efficient and environmentally sustainable. As green water treatment philosophies gain traction and as regulatory and residuals-handling constraints continue to tighten, the use of biological drinking water treatment technologies and processes will likely continue to expand around the globe.

REFERENCES

Bouwer, E. J., and P. B. Crowe. 1988. Biological processes in drinking water treatment. Journal of the American Water Works Association 80(9):82-93.

Brown, J. C. 2006. Simultaneous destruction of multiple drinking water contaminants using biological filtration. Proceedings of the 2006 AWWA Annual Conference and Exhibition, San Antonio, TX. CD available from AWWA, 6666 W. Quincy Avenue, Denver, CO 80235.

Brown, J. C., R. D. Anderson, J. H. Min, L. Boulos, D. Prasifka, and G. J. G. Juby. 2005. Fixed-bed biological treatment of perchlorate-contaminated drinking water. Journal of the American Water Works Association 97(9):70-81.

Dahab, M. F., and B. L. Woodbury. 1998. Biological options for nitrate removal from drinking water. Paper presented at the AWWA Inorganic Contaminants Workshop, San Antonio, TX; February 22-24. CD available from AWWA.

Gollnitz, W. D., J. L. Clancy, B. L. Whitteberry, and J. A. Vogt. 2003. RBF as a microbial treatment process. Journal of the American Water Works Association 95(12):56-66.

Green, M. R., and M. P. Pitre. 1999. Treatment of groundwater containing perchlorate using biological fluidized bed reactors with GAC or sand media. Pp. 241-257 in Perchlorate in the Environment, E.T. Urbansky, ed. New York: Kluwer Academic/Plenum.

Guarini, W. J., and T. Webster. 2004. Ex situ biological treatment of perchlorate using biological fluidized-bed reactors: An update. Proceedings of the East Valley Water District/AWWARF Water Quality Conference, Ontario, CA. CD available from AWWA.

Herman, D. C., and W. T. Frankenberger Jr. 1999. Bacterial reduction of perchlorate and nitrate in water. Journal of Environmental Quality 28(3):1018-1024.

Kim, K., and B. E. Logan. 2000. Fixed-bed bioreactor treating perchlorate-contaminated waters. Environmental Engineering Science 17(5):257-265.

Kirisits, M. J., V. L. Snoeyink, J. C. Chee-Sanford, B. J. Daugherty, J. C. Brown, and L. M. Raskin. 2002. Effects of operating conditions on bromate removal efficiency in BAC filters. Journal of the American Water Works Association 94(4):182-193.

Lauderdale, C. V., J. C. Brown, B. MacLeod, M. Simpson, and K. Petitte. 2006. A novel biological treatment approach for the removal of algal metabolites. Proceedings of the 2006 AWWA Water Quality and Technology Conference, Denver, CO. CD available from AWWA.

Liu, J., and J. Batista. 2000. A hybrid (membrane/biological) system to remove perchlorate from drinking waters. Paper presented at Perchlorate Treatment Technology Workshop, 5th Annual Joint Services Pollution Prevention and Hazardous Waste Management Conference and Exhibition, San Antonio, Texas, August 21-24.

Madigan, M. T., J. M. Martinko, and J. Parker. 1997. Brock Biology of Microorganisms, 8th ed. Upper Saddle River, NJ: Prentice Hall.

Nerenberg, R., B. E. Rittmann, and I. N. Najm. 2002. Perchlorate reduction in a hydrogen-based membrane-biofilm reactor. Journal of the American Water Works Association 94(11):103-114.

Ray, C., G. Melin, and R. B. Linsky. 2002. Riverbank Filtration: Improving Source-Water Quality. The Netherlands: Kluwer Academic Publishers.

Rikken, G. B., A. G. M. Kroon, and C. G. van Ginkel. 1996. Transformation of (per)chlorate into chloride by a newly isolated bacterium: Reduction and dismutation. Applied Microbiology and Biotechnology 45(3):420-426.

Weiss, W. J., E. J. Bouwer, W. P. Ball, C. R. O'Melia, H. Arora, and T. F. Speth. 2003a. Comparing RBF with bench-scale conventional treatment for precursor reduction. Journal of the American Water Works Association 95(12):67-80.

Weiss, W. J., E. J. Bouwer, W. P. Ball, C. R. O'Melia, M. W. LeChevallier, H. Arora, and T. F. Speth. 2003b. Riverbank filtration—fate of DBP precursors and selected microorganisms. Journal of the American Water Works Association 95(10):68-81.

ADDITIONAL READING

Rittmann, B. E., and P. L. McCarty. 2001. Environmental Biotechnology. Boston, MA: McGraw-Hill.

Distribution Systems: The Next Frontier

Vanessa L. Speight
Malcolm Pirnie Inc.
Cincinnati, Ohio

Drinking water treatment technology has advanced significantly over the past century. However, once the highly treated water leaves the treatment plant, it must travel to consumers through a distribution system. A variety of opportunities exist in the distribution to degrade water quality and affect the end product consumed by the public. As the last barrier in the treatment process, distribution systems are vital to the protection of public health. Because pipes are buried and outside the direct control of water utilities, managing distribution systems is one of the greatest challenges in the provision of safe drinking water.

Within the past few decades an increasing focus of research has been placed on distribution systems. It is a difficult balancing act to maintain a continuous water supply to customers, provide adequate fire flow at all parts of the system, maintain pressure, and ensure water quality. Distribution systems are built as cities grow and therefore contain a variety of pipe ages, materials, and quality of construction. Many systems contain pipes that are over 100 years old, most of those constructed of unlined cast iron. Replacement costs are rising, with estimated costs of replacing water system infrastructure in the United States ranging from $77 billion to $325 billion in the next 25 years (Deb et al., 2002). Increasing energy costs are also adversely affecting water utilities that rely on extensive pumping to maintain pressure and deliver water.

The National Academy of Sciences recently convened a panel to examine public health risks from distribution systems (NRC, 2007). This report focused on three areas related to distribution system integrity: physical, referring to the pipes as barriers; hydraulic, referring to the delivery of water at the desired quantity and pressure; and quality, referring to the maintenance of the water quality through its travel in the distribution system. All three areas of integrity must be addressed to ensure public health. High-priority areas for risk reduction identified by the panel included cross-connections, new and repaired water mains, and water storage (NRC, 2005).

Distribution systems represent the next frontier of research needs and technology challenges for the drinking water industry. In addition to the infrastructure replacement and management challenges, water-quality issues are highly scrutinized by the public and media. Tools are emerging to address these challenges but considerable work remains to fully address the issues.

DISTRIBUTION SYSTEM WATER QUALITY

One key to the protection of public health in the distribution system is the maintenance of a disinfectant residual, typically in the form of free chlorine or chloramine. As water travels through the distribution system the disinfectant oxidizes material in both the bulk water and at the pipe wall surface, thereby reducing the residual available to maintain disinfection (Figure 1). At the pipe wall the chlorine can react with corrosion products, sediments, and biofilms. Biofilms have been shown to grow on most common pipe materials, but the quantity of attached bacteria is several orders of magnitude higher on unlined cast iron pipes (Camper et al., 2003).

Drinking water regulations place limits on the minimum and maximum amounts of disinfectant allowable in the water. In addition, several classes of disinfection byproducts are regulated. These compounds, which form when chlorine

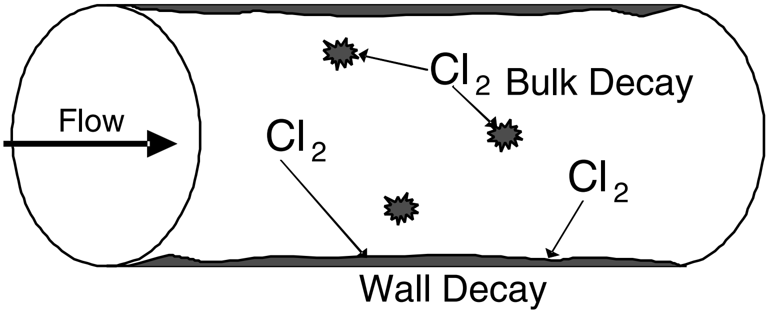


FIGURE 1 Schematic of chlorine decay reactions in a distribution system. Source: Adapted from Speight, 2003.

reacts with natural organic matter present in the water, are suspected carcinogens and have potential reproductive health effects (EPA, 2006). They continue to form in the presence of free chlorine as water travels to the consumer. Therefore, it is important to deliver the treated water as quickly as possible to the end users to reduce the potential for disinfection byproducts to form.

Microbial contamination in the distribution system is also a potential threat to public health (Craun and Calderon, 2001). Pathways for entry of microbial contaminants into the distribution system include survival of organisms in the treatment process, intrusion of contaminated groundwater from the outside of the pipe when pressure drops within the pipe, contamination during main installation or repair, and backflow from a nonpotable system connected to potable plumbing. The presence of pathogens within the biofilms or entering the distribution system is not well understood and considerable research is needed to fully understand the microbial interactions taking place throughout the system.

Storage of water is provided in elevated tanks, ground tanks, and sometimes in the pipes themselves. Depending on its design and operation, there is potential for water to remain within the tank for extended periods of time. The longer the retention time in the tank, the greater the potential for residual disinfectant decay, formation of disinfection byproducts, and microbial regrowth. Therefore, management of storage is a critical issue for distribution system operation. Furthermore, the need for fire flow storage often results in design of larger tanks than would be optimal for water-quality purposes. The design and operation of storage varies widely across the country, with regional preferences and architectural influences playing a large role. Tanks have also been shown to be the site of contaminant entry into the distribution system, either through broken hatches or sediment accumulation (Clark et al., 1996).

The distribution system is connected to every location where water is available to consumers. Typically the water utility's jurisdiction ends at the customer's meter and the remaining plumbing is the responsibility of the building owner. All the reactions that can degrade water quality in a distribution system can occur within household plumbing. Many incidents that gain media attention are linked to household plumbing issues, such as the problems with lead in Washington, D.C., in 2004 (EPA, 2007). Intentional contamination of the system is also a potential threat.

TOOLS TO ADDRESS DISTRIBUTION SYSTEM ISSUES

Because access to the distribution system itself is extremely limited, the water industry relies on a variety of tools to operate, maintain, and continually improve these systems. A variety of water-quality sampling is conducted daily to meet regulations and to inform decisions about operations. While collection of grab samples has been the traditional sampling method, the use of online monitors is increasing. These monitors can include measurement of simple water-quality parameters like pH and turbidity as well as more sophisticated analyses such as total organic carbon. Considerable research is ongoing to determine the optimal number of grab samples and online monitors and their placement, the best parameters to monitor for different objectives, and most accurate analytical methods (Speight et al., 2004). Along with the collection of online water-quality data a field of work is emerging in data analysis and event detection (Hart et al., 2007). While much of this work originated in the security arena, water utilities are seeking multiple benefits from this expensive equipment and are looking for basic operational information as well as an indication of a contamination event, intentional or otherwise (ASCE, 2004).

Hydraulic and water-quality modeling tools provide a way to link hydraulic and water-quality parameters in a simulation. Hydraulic models have been in use for several decades and can provide reliable simulations of flows and pressures when appropriately calibrated to real-world conditions (Ormsbee and Lingireddy, 1997). Figure 2 shows the input data required for hydraulic and water quality models, using chlorine modeling as an example.

A key challenge in hydraulic modeling is the determination of customer water usage at all points in the distribution system over time. Where customer meters

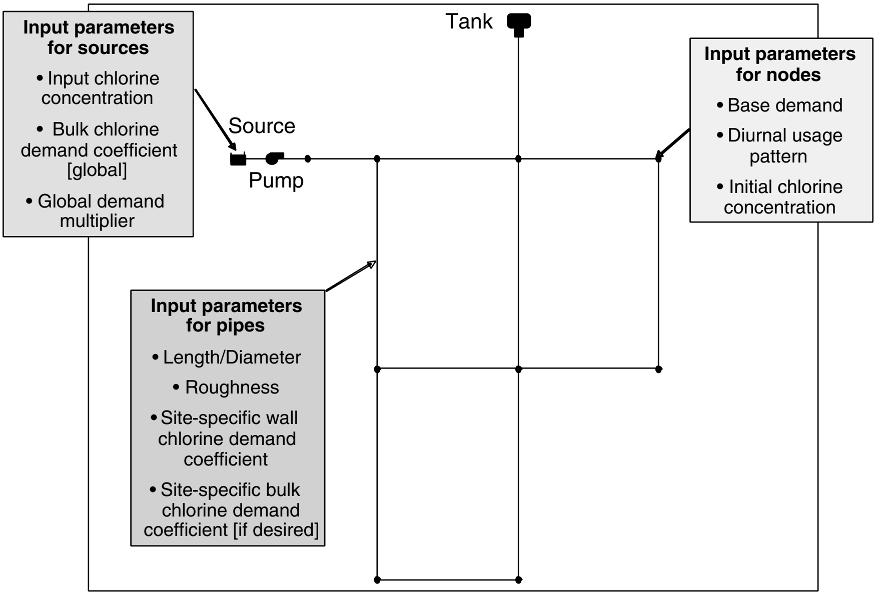

Tank
Input parameters for sources
• Input chlorine concentration
• Bulk chlorine demand coefficient [global]
• Global demand multiplier
Source
Pump
Input parameters for nodes
• Base demand
• Diurnal usage pattern
• Initial chlorine concentration
Input parameters for pipes
• Length/Diameter
• Roughness
• Site-specific wall chlorine demand coefficient
• Site-specific bulk chlorine demand coefficient [if desired]

FIGURE 2 Schematic of model input requirements to simulate chlorine in a distribution system. Source: Adapted from EPA, 2002.

exist, they are generally read on a monthly or quarterly basis and do not provide real-time data. Innovations in automated meter reading (AMR) may improve our ability to collect real-time customer usage data, but the data management challenges are significant. Very few utilities have implemented real-time modeling of their distribution systems by linking operational data with the model and running repeated simulations. Real-time models are primarily used for detection of situations that differ from baseline, such as main breaks or large fires, as well as energy management (Jentgen et al., 2003). The field has not yet advanced to the level of sophistication needed to fully automate the operation of a distribution system using a real-time model.

Water-quality modeling is still considered an emerging field but is increasingly used to assist in operational decisions and planning studies. Chlorine is the most commonly modeled parameter for water quality. A relatively simple first-order decay model for chlorine has been shown to provide good simulations of field data (Vasconcelos et al., 1997). Modeling of disinfection byproducts has been less successful using deterministic models (Speight et al., 2000). Probabilistic models of water quality hold some promise in that they allow the use of uncertain mechanistic formulations with appropriate accounting of uncertainty.

Modeling of microbial contaminants within the distribution system is limited by the lack of knowledge about biofilm processes, attachment or detachment of microbes to particles, growth mechanisms in low-nutrient environments, concentrations of microbes that enter the system through different pathways (e.g., low-pressure transient versus cross-connection with a sewer), and occurrence of pathogens. Particle transport and deposition modeling is particularly challenging for distribution systems because of the highly variable nature of flow, which in turn is caused by the highly variable nature of individual water usage.

The asset management field is developing tools to model infrastructure deterioration based on data input, such as pipe age, material, soil conditions, and number of breaks. The development of reliable, in situ condition assessment technologies for water mains is an emerging field as is the development of methods for quick rehabilitation that limit the time a water main is out of service. Given the large financial burden of replacing the deteriorating water distribution infrastructure in the United States, cost-effective approaches to asset management, rehabilitation, and replacement will be needed.

Alternative water delivery systems are also emerging, including dual distribution systems carrying potable water and nonpotable water separately. In certain parts of the country where water supply is scarce such as Florida and California, the use of reclaimed water (highly treated wastewater effluent) is commonplace and dual distribution systems are being installed throughout cities to provide an alternative for irrigation and fire fighting needs. The challenges associated with potable water distribution systems exist in parallel for reclaimed water distribution systems and include maintenance of the integrity of the buried infrastructure, maintenance of water quality, and delivery of adequate supply and pressure. Point-of-use treatment devices are also being considered in areas where achieving the desired water quality at the consumer's tap is not achievable.

CONCLUSIONS

Water distribution systems represent an exciting challenge for the engineering community. Solving the distribution system problems will require work in a variety of interrelated fields, including infrastructure materials, water treatment, hydraulics, water chemistry and microbiology, data management, computer modeling, human behavior, public health and public education, and risk management.

REFERENCES

ASCE (American Society of Civil Engineers). 2004. Interim voluntary guidelines for designing an online contaminant monitoring system. Online. Available at: <http://www.asce.org/static/1/wise.cfm>.

Camper, A. K., K. Brastrup, A. Sandvig, J. Clement, C. Spencer, and A. J. Capuzzi. 2003. Effect of distribution system materials on bacterial regrowth. Journal of the American Water Works Association, 95(7):107-121.

Clark, R. M., E. E. Geldreich, K. R. Fox, E. W. Rice, C. H. Johnson, J. A. Goodrich, J. A. Barnick, and F. Abdesaken. 1996. Tracking a *salmonella serovar typhimurium* outbreak in Gideon, Missouri: Role of contaminant propagation modeling. Journal of Water Supply Research and Technology—Aqua 45(4):171-183.

Craun, G. F., and R. L. Calderon. 2001. Waterborne disease outbreaks caused by distribution system deficiencies. Journal of American Water Works Association 93(9):64-75.

Deb, A. K., F. M. Grablutz, Y. J. Hasit, J. K. Snyder, G. V. Loganathan, and N. Agbenowsi. 2002. Prioritizing Water Main Rehabilitation and Replacement. Denver, CO: American Water Works Association Research Foundation.

EPA (U.S. Environmental Protection Agency). 2002. EPANET Program. For examples, see <http://www.epa.gov/nrmrl/wswrd/dw/epanet.html>.

EPA. 2006. National Primary Drinking Water Regulations: Stage 2 Disinfectants and Disinfection By-Products Rule; Final Rule. 40 CFR Parts 9, 141 and 142, Federal Register 71:2:388-493. Washington, DC: EPA.

EPA. 2007. Lead in DC Drinking Water. Washington, DC: EPA.

Hart, D., S. McKenna, K. Klise, V. Cruz, and M. Wilson. 2007. CANARY: A water quality event detection algorithm development tool. CD-ROM. World Environmental and Water Resources Congress 2007. Reston, VA: American Society of Civil Engineers.

Jentgen, L., S. Conrad, R. Riddle, E. W. Von Sacken, K. Stone, W. M. Grayman, and S. Ranade. 2003. Implementing a Prototype Energy and Water Quality Management System. Denver, CO: American Water Works Association Research Foundation.

NRC (National Research Council). 2005. Drinking Water Distribution Systems: Assessing and Reducing Risks—First Report. Washington, DC: National Academies Press.

NRC. 2007. Drinking Water Distribution Systems: Assessing and Reducing Risks. Washington, DC: National Academies Press.

Ormsbee, L. E., and S. Lingireddy. 1997. Calibrating hydraulic network models. Journal of American Water Works Association 89(2):42-50.

Speight, V.L. 2003. Development of a randomized sampling methodology for characterization of chlorine residual in drinking water distribution systems. Ph.D. Dissertation. University of North Carolina at Chapel Hill.

Speight, V. L., W. D. Kalsbeek, and F. A. DiGiano. 2004. Randomized stratified sampling methodology for water quality in distribution systems. Journal of Water Resources Planning and Management 130(4):330-338.

Speight, V. L., J. R. Nuckols, L. Rossman, A. M. Miles, and P. C. Singer. 2000. Disinfection by-product exposure assessment using distribution system modeling. Paper presented at AWWA Water Quality Technology Conference, Salt Lake City, UT, November 5-9.

Vasconcelos, J. J., L. A. Rossman, W. M. Grayman, P. F. Boulos, and R. M. Clark. 1997. Kinetics of chlorine decay. Journal of American Water Works Association 89(7):54-65.

ADDITIONAL READING

EPA (U.S. Environmental Protection Agency). Distribution System White Papers and Total Coliform Rule Issue Papers. See <http://www.epa.gov/safewater/disinfection/tcr/regulation_revisions.html>.

APPENDIXES

Breakout Sessions

On the first afternoon of the symposium, breakout sessions gave participants an opportunity to talk in smaller groups about public policy-oriented engineering issues. Some of the issues were discussed by more than one group. The discussions in each group are summarized below. The opinions and recommendations expressed are those of the participants and not necessarily those of the National Academy of Engineering (NAE).

TOPIC 1: WHAT DOES SUSTAINABILITY MEAN TO/FOR THE ENGINEERING COMMUNITY?

Group 1

The early part of the discussion was largely concerned with the meaning of the word "sustainability." In general, it was agreed that sustainability relates to preservation of ecological systems with a particular emphasis on global issues such as climate change and oil depletion. Sustainability also implies a "relative" measure. For instance, the gasoline produced today is said to be more sustainable compared to 10 years ago because of improved efficiencies and decreased energy consumption during the whole production cycle. Also, energy-saving compact fluorescent light bulbs might be considered more sustainable than incandescent ones.

One of the more common definitions of sustainability is that created by the Brundtland Commission, led by the former Norwegian Prime Minister Gro Harlem Brundtland. The Commission defined sustainable development as development that, "meets the needs of the present without compromising the ability of future generations to meet their own needs."

While there may be few regulations governing sustainability, it was agreed that clear quantitative definitions and means to measure and compare are important. This will allow corporations and individuals to set sustainability goals and measure progress against them. Life cycle analysis (LCA), which can provide a composite measure of sustainability, measures the environmental performance of products and services through all phases of their life cycle, starting from the extraction and processing of raw materials to manufacturing, transportation, use, recycling, and final disposal. LCA may refer to the impact of a product or service on specific measures such as greenhouse gas emissions (global warming), ozone layer depletion, and land use.

Various examples of sustainable practices at corporations represented by the breakout session participants were discussed, including water conservation during aluminum refining and the use of bio-based and biodegradable plastics.

Conclusions/Recommendations

- Sustainability through the preservation of natural resources for future generations should be an important consideration for all engineers and consequently for NAE.
- A rigorous definition of sustainability and its various facets is important and necessary.
- Standardized means to quantify and report sustainability are extremely important to encourage good practices and minimize fraudulent claims. This is an area on which NAE can exert a considerable influence, possibly through the commissioning of a task force composed of engineers, scientists, environmentalists, and policy makers. This task force may also explore ways to introduce regulations to encourage sustainable practices/processes by commercial enterprises.
- Awareness and education at the grassroots level are important in order for society to enjoy the benefits of sustainable practices for generations to come. Formal introduction to sustainability concepts in high school and college curricula will go a long way to achieving this. This is also an area where NAE can exert a strong influence through concerted efforts such as the preparation and distribution of appropriate literature, short course offerings, and teacher-training programs.

Group 2

This group spent the first part of its discussion trying to define sustainability. It was quickly discovered that this term means different things depending on the perspective and background of the individual. Sustainability can be examined in terms of energy, water, air, materials, and other environmental resources as well as financial resources.

In general, many people think of environmental sustainability as the key issue that we are facing, but the problem of addressing sustainability is much larger than this. There are social, political, and economic pressures that place constraints on the solutions that engineers develop. From the technological perspective, engineers can define the problem and develop solutions over a period of time; however, the broader questions that include society are:

- How can people live in a densely populated metro area yet maintain sustainability in terms of water, energy, air, and the environment? Is it possible to support large, dense population centers without much impact on the natural ecosystem? Is there a density limitation that we should consider?
- What role should governmental policy play in supporting or promoting sustainability in the broader society? How do engineers influence public policy, and are we appropriately trained to do so? This area is governed much more by public perception, economics, and law, which are outside of the typical training of engineers.
- Where do we invest our money for sustainability? Considering the questions above, a portion must go into developing policy and lobbying as well as R&D.
- Economic pressures will dictate changes in how we address sustainability. Is there sufficient communication between engineers and others to truly understand these pressures when developing technical solutions for sustainable systems?
- CO_2 production and the carbon footprint are now being touted as the metric for sustainability. Is this the best metric to use?
- While technical solutions to promote sustainability in a specific area can be developed, we must also consider life cycle analysis of products and devices. This includes everything in the production cycle from gathering the raw materials used in the device to methods of disposal.

The questions raised go far beyond technical solutions such as developing energy-efficient devices or renewable energy sources. Often there is little consideration of energy and environmental sustainability in developing countries or in domestic markets fueled by small businesses since the goal is to develop financial sustainability. Governmental policies in other countries adhere to different definitions of sustainability. This has influenced some shifts in how goods are produced and sold through globalization. While we need to address sustainability in all areas, the issue of economic growth and financial sustainability is difficult to deal with. This must be taken into account when developing solutions. Businesses must buy into sustainability as being good for them from the early stages of inception and growth. Sustainability should not be seen as needless regulation coming from environmentalists.

In addition to having business and industry buy into sustainability, we must also address the public's perception of sustainability. If the public does not understand the issues and challenges, it will be hard to get them to buy into sustainability concepts. A key example can be found in low-cost incandescent light bulbs vs. energy-efficient compact fluorescent light bulbs. For those who do not fully understand sustainability, the choice between the two is clear—the cheap incandescent bulb. The savings that consumers will see over a long period of time are not clear to them; most are interested in savings at the cash register. Thus, public education and awareness are keys to promoting the sustainable solutions that engineers develop, and they must learn how to articulate the challenges and solutions to the general public.

Finally, the role of politics and regulation should be to increase the pressures that lead us to sustainable solutions while keeping the public informed and providing an environment for financial sustainability of businesses and industries.

On the technical side, there are many challenges that engineers must still address. They must develop uniform metrics that provide fair and accurate comparisons of the impacts of various solutions to sustainability. Using the carbon footprint as a metric may not be the most appropriate method. Issues such as cogeneration and energy harvesting from the environment (e.g., waste heat, mechanical motions, etc.) must be considered to meet the needs of our growing society.

Recycling should also be promoted more aggressively, although herein is a hidden issue of economics. In many areas of the United States, PET plastic bottles used for drinks are simply dumped instead of being recycled because it is cheaper for the PET industry to make a new bottle from hydrocarbon feedstock than to recycle the PET. This problem should be addressed technically. Can we develop either a biodegradable plastic bottle or develop a bottle that can be easily recycled in order to reduce our impact on the environment? Some cities and regions are addressing this through public policy, for example, San Francisco's ban on plastic drinking bottles. We may need a combination of technical, societal, and policy pressures to resolve this issue.

There are issues in energy storage and energy production that will require engineers to develop creative solutions. While we have looked at energy efficiency and energy production through the years, energy storage solutions have lagged behind these other areas. The best device we have is the battery. Are there other solutions that improve the energy density of current battery technology? This is key since it would allow us to produce and store energy when it is cheap to do so (e.g., wind energy to electrical energy conversion during a storm) and discharge it when demands for energy are high.

We must also revisit the issue of nuclear energy. Nuclear energy production has a smaller environmental footprint than other forms of energy production. Yet, nuclear energy has a bad public image in the United States because no one wants a power plant or a nuclear waste repository near his home. To address future energy production needs, we may need to go to a decentralized form of energy production and develop a smart grid system for delivering power to local communities. This will also allow adoption of specific technologies that are best suited for a given region of the country to maximize energy production.

TOPIC 2: BALANCING CAREER AND FAMILY

Most of the group participants have families with children ranging in age from 3 months to 18 years old. Only one participant has one stay-at-home parent, while for the rest of the group both parents work full time in academia or in industry.

Issues discussed

Timing: The first issue that came up is whether the best time to have children is early or later in your career. Opinions were split. Some felt you should try to have children as soon as possible, even during graduate school. Others felt that having children later when one's career is well established provides more flexibility when it is needed and removes some of the pressure on proving yourself, which is more intense at the early stages of your career.

Academia vs. industry: The tenure-track process came up while discussing timing and led to comparisons between working in academia or in industry while having a family. The group agreed that academia provides significantly more flexibility. It was noted that the generation of women at the Frontiers of Engineering meeting is probably the first that had children before establishing tenure security and that more senior faculty members are adjusting to this new situation. Standards are only recently changing in industry as well. Some participants were the first to maintain their industry jobs after having children. Their example seems to be helping the next generation of engineers.

Productivity: The women in the group were in agreement that their productivity actually increased while expecting and after delivering their babies. This was attributed to: (1) strict deadlines that had to be met as imposed by the pregnancy and post-partum schedule; and (2) self- and work environment-imposed pressures of new parents proving to their peers that they are as productive as they were prior to having children.

Day care: Day care centers or in-house babysitting was used equally among the families with two parents working outside the home. All emphasized that good coordination between parents is necessary whether or not both have full-time jobs outside the home. Dropping off and picking up the children can be coordinated with other parents and in some cases between co-workers who live in close proximity. In a few cases, grandparents and extended family members who live nearby assist in day care.

Dealing with school schedules, after-school activities, and illness: Additional support is needed for the many holidays and days off in the school schedule that exceed the number of available vacation days. Suggestions for dealing with this included combining care with other parents to reduce the number of days off and being able to work from home as much as possible. The same applied to after-school activities.

Number of children: Some of the attendees have one child and asked the group how many children they think would be optimum. There was no consensus except for the realization that the amount of work required with more than one child increases exponentially rather than linearly! Everyone agreed that when children enter school this effort decreases somewhat.

Conclusions/Recommendations

The session was very informative and fun. Everyone was relieved to know that other professionals go through more or less the same problems, and generally all of us feel we are not always able to perform our best as both parents and as professionals with demanding careers. Everyone agreed on these issues and on recommendations regarding balancing family and career:

- *Outsource* as much as possible by hiring help for day care and time-consuming housework such as housecleaning and yard maintenance. It really improves quality of life and reduces stress levels.
- *On-site day care* is already available in some universities and corporate environments but with very long waiting lists. As a first step, employees can request on-site day care that would not necessarily impose a financial burden on the employer, i.e., organization of an onsite space that would minimize commute time and maximize employee efficiency.
- *Flexibility to work from home* is possible, especially after being established (or tenured) and if precedents were set by other colleagues.
- *A supportive spouse* is the key to making career and family work. Parents need to collaborate and take turns depending on their daily workload vs. family needs.

TOPIC 3: THE ROLE OF THE SOCIAL SCIENCES IN ENGINEERING

Each member raised one or more topics for discussion. These included:

- Computational modeling of the human role in engineered systems, specifically to explain and predict human performance in complex systems.
- The importance of human performance modeling for the armed services, where human physical and cognitive abilities are often stretched to the limit.
- Computational modeling of social psychology and systems for immersive training.

- The engineer as a social person, and also the untapped possibilities for collecting data on the Internet.
- The impact of technology on people and human needs going forward into the next century.
- The sociology of security technologies and issues of technology transfer from laboratory to products.
- Effects-based operations and human-computer interaction (HCI) issues.
- The role of education in technology transfer and especially the educational difference between people developing a technology and those who use it.
- Collaborative social environments and, in particular, HCI techniques to inform engineering processes and products.
- The difficulties of valuing social aspects of engineering, including economic barriers.
- Issues of teams and individuals in space, including modeling of distributed and individual expertise and training people in multiple disciplines.
- How different fields think about the same concepts, e.g., how statisticians think about data.

Different fields bring very different perspectives and methods to the same goals. In particular, there is a basic mismatch between engineering and the social sciences in expectations of precision. One participant drew an analogy between the current state of the social sciences and the early stages (the 1700s) of the physical sciences, e.g., it took 200 years to define a formal concept of "temperature." It was noted that social scientists are getting better at predicting, at least at aggregate levels. Engineers need to be able to quantify something, which can be part of the challenge of interacting with social scientists. There is a tendency in some fields, e.g., optimization or economics, to develop well-behaved mathematical models for the sake of abstraction and tractability that do not reflect reality. Conversely, it can be challenging to implement theories from social sciences because the process of translating verbal and conceptual theories into computer code is unclear. Simulations of social theories reveal a mismatch of levels of analysis between aggregate and individual behavior. There is a poor understanding of variations in social sciences, partly reflecting the impact of publication constraints, and an accompanying emphasis on small, reproducible effects over larger but more variable effects.

How do we get a sufficiently representative sample for theorizing about social and cognitive phenomena? A participant emphasized the importance of the Internet as a platform for social sciences research and modeling. Another mentioned a recent article in *The Economist* about the important role that online gaming can play as a research platform. In particular, massive online multiplayer games promise large and quantified sets of data. However, both the shortcoming and the advantage of field studies over laboratory studies is lack of control. It is particularly challenging to get data about human performance, decision making, and social interaction in extreme situations. One group member discussed using tragedy scenarios in his research, and another once did a study at Everest base camp. Although these examples have good face validity in the context of studying human performance under stress, even here it remains an open issue how much stress the participants in the studies really were under. A broadening array of measures, including skin conductance, EEG, and fMRI, are being used to improve our fundamental understanding of the nature of emotional and physical stressors and their relationship with human performance.

What is the relationship between believability and validity in social sciences research? Is believability of characters similar to tolerance in engineering? It was noted that people adapt over time. This is a challenge to the prospects for persistent predictive models in the social sciences. There is a tension between modeling expert knowledge, which is more predictable, vs. modeling human adaptation, which is more useful, and an accompanying tendency to throw out interesting data, e.g., training data.

How do we train people across traditional boundaries, such as between social sciences and engineering? Should formal education/training in social sciences be part of the process of developing engineers? Should exposure to physical, artificial, and engineering sciences be part of the process of developing social scientists? Some effort is being made at Princeton University at the undergraduate level to encourage engineering students to investigate educational opportunities in other domains in order to broaden their perspectives.

The session closed with this provocative question: What are the limits to the argument that we need to study the human as part of socio-technical systems, rather than investing in technologies that allow us to remove people from those systems?

TOPIC 4: BALANCING RIGOR AND CREATIVITY IN INFORMATION TECHNOLOGY, SCIENCE, ENGINEERING, AND DESIGN

Group 1

The group discussed whether rigor and creativity are always in conflict or only in some circumstances. Some alternatives to the "balance of rigor and creativity" dimension were suggested, such as low-risk vs. high-risk or incremental vs. creative. Two examples motivated the discussion of alternative dimensions. One was developing a particular sports car design where rigor is a necessary constant and creativity encouraged. The other was in a programming project where there was low creativity to begin with and rigor fell off rapidly. Both examples suggested that there isn't necessarily a conflict between rigor and creativity.

Discussion of a suggested tradeoff or conflict between creativity and rigor led to the question of whether there could be a synergistic relationship between rigor and creativity. One participant noted that opportunities to exercise creativity motivate increased rigor, possibly through increased enthusiasm by allowing engineers to buy into the process. Some participants thought that any synergy between rigor and creativity was conditioned on domain and task constraints. Several participants noted that they felt that tasks with constraints might allow for this synergy (e.g., the sports car design case); in contrast, in tasks with few constraints (e.g., brainstorming) one might see a conflict between creativity and rigor.

What are the impediments to creativity?

In education: There was a general consensus among the academics in the group that U.S. college students are behind in math and science compared to students from many other industrial countries. The sentiment seemed to be that this reflected a deficiency of rigor among U.S. students' education as a whole.

Some academics expressed dismay that engineering was not considered by prospective students to involve much creativity, despite the constructive and creative nature of the engineering enterprise. There was a related sentiment expressed among faculty that engineering-declared students might actually be driven away from engineering and science, in part from the lack of preparation that was expected at the college level and also as a result of "bait and switch." For example, playing video games and seeing the results of a creative process without the requisite rigor made explicit can lead to faulty expectations for those without suitable preparation.

Statements such as "we are teaching rigor, not creativity" and "students turn to clubs to exercise creativity" suggest that the educational system does not reward creativity. Reward systems, more generally, are addressed below.

In research: Participants noted that journals and conferences reward incremental advances and perhaps large creative endeavors that paid off. There was some sentiment that tenure criteria penalize dead ends and wrong turns, which are a necessary part of creative exploration.

There were distinctions raised between funding from industry and funding from government agencies. These differences focused on time scale: Was the funding source expecting results in 2, 3, 5, or 10 years, and does the time scale influence the creativity of the ideas? After some initial disagreement, there was agreement that industrial funding might be very short-term, inviting little creativity, or very long-term. It was generally agreed that lack of resources, including time, was a major impediment to creativity.

In industry: A controversial assertion was made—that we have lost the culture of innovation. The group generally agreed that the lack of a culture of innovation was conditioned on such factors as the industry, so that the aerospace industry, according to one participant, had lost a culture of innovation, while the computing industry, according to other participants, had not. This led to speculation that the life-stage of an industry or company influences its balance of creativity and rigor. This assertion came with the caveat that this is probably not a conflict per se but may be a function of an early entrepreneurial phase versus a later steady-state phase of a project or organization. Finally, high levels of comfort, or perhaps fears of greater discomfort (or perceived relative comfort) mediated a balance of creativity and rigor. Do lack of resources (discomfort) fuel rigor and/or fuel creativity? Does comfort diminish one or both?

In general, it seemed that lack of resources, insufficient rewards for taking risks, and agent comfort were impediments to creativity and possibly rigor as well. Again, it is possible that both fall and rise together in many circumstances.

Encouraging and facilitating creativity

Participants suggested several strategies for encouraging creativity and rigor.

Extend time horizons and resources: An industry participant indicated that grants were given for long-range efforts, and that these long-range efforts had to be innovative to justify the funds. An educator described a strategy for advising graduate students that gave students short-term goals to work on and a problem that required long-term thought, not to be confused with long-term number crunching. Extended time horizons are one way of extending resources.

Practiced risk taking: One participant described the practice in his lab of holding weekly brainstorming meetings—“half-baked lunches”—where expression of half-baked ideas was encouraged and critiqued. The participant reported growing more comfortable with the process in time; to paraphrase the participant, “It was neat to see senior people having their ideas picked apart.” Presumably there are many variations on this theme, but an essential idea is that brainstorming is practiced to a point of comfort, at least within an accepting community.

Another participant told of having graduate students prepare as their final class project a National Science Foundation (NSF)-style proposal that is reviewed by the instructor according to NSF criteria of intellectual merit and broader impacts, rather than turning in a finished project. Another participant described an “idea tracker” that allowed employees to post ideas within their own or other departments that went before department committees for possible action. The participant reported comfort with having their “wacko ideas” discussed. One participant summarized well—focusing on mechanisms that increase comfort with creative exploration and risk-taking are more important than focusing on creative end products per se. There are no shortcuts.

Reducing compartmentalization: Various participants gave examples of reducing compartmentalization in various forms—between disciplines, people, and technical and societal issues—to encourage creativity.

NSF's broader impacts criteria was given as one example of reducing compartmentalization between research and its societal, environmental, and other-science impacts. Broader impact considerations can attract traditionally under-represented groups (e.g., women) to engineering, which can increase diversity. This in turn nourishes engineering with new perspectives that further support discipline-based integration as well as integration between engineering and its broader impacts. It was noted that the symposium talks on computer and data security and privacy wove together technical discussion and societal implications in a very fluid manner.

Participants mentioned examples of reducing various forms of compartmentalization, such as interdisciplinary, cross-functional teams for class projects; contextualizing computing with business or law to make it more appealing to students; and special courses on technology and innovation or computation and society. Expanding the diversity of first-class stakeholders leads to more comprehensive utility functions for assessing system-development success, which can promote both increased rigor and increased creativity.

In summary, the group concluded that:

- Creativity and rigor can be synergistic.
- Lack of resources, including time, is a major impediment to creativity.
- Tailor reward systems to a desired weighted balance of creativity and rigor and apply rewards consistently.
- Reduce compartmentalization to encourage creativity or at least prevent stagnation.

Group 2

It was noted that entering the terms "rigor and creativity" in Google generates 587,000 hits. Also, both terms commonly occur in promotion and tenure documents. Questions posed included:

- What are examples of conflict and synergy between rigor and creativity?
- How do these issues play into specific areas of engineering, computer science, and design?
- How do these issues play into careers? Do publication-driven positions reward rigor? Do innovation-driven positions reward creativity?

Why is this topic of interest to you?

- From the teaching perspective, it helps with understanding students who come from different backgrounds.
- The issue isn't about balance. There are scientific methods that a true scientist should adhere to (rigor). Creativity cannot be taught. Most people are not creative; the rare individual possesses it.
- Rigor needs to be taught but the environment should support creativity.
- There could be creativity flowing constantly; however, looking at problems in depth with rigor requires an extended period of time.

Can/How do rigor and creativity co-exist?

- Creativity may require a higher-level approach. The interaction between rigor and creativity is complex.
- There are many ideas of varying quality, therefore, there's a wide spectrum of creativity. However, some ideas are inherently limited due to the approach that's being taken, which constrains the research to incremental advances rather than being disruptive in nature.
- The way to excite people about a scientific idea is through creativity. Creativity can be seen as at odds with rigor. For instance, interdisciplinary research is generally perceived as not very rigorous.
- Rigor is required to understand and manage creativity.
- Creativity can be aligned with rigor. Creativity can flow out of rigor, but cannot be necessarily taught. There is a synergy between creativity and rigor if arranged properly.
- Creativity without rigor may not lead to usable ideas and products. However, you cannot let rigor get in the way of creativity.

Are there differences in reward and utility functions in academia with respect to creativity?

- Earlier career academics get their recognition through rigor as demonstrated by number of publications, etc.
- In Google, some pages contrast creativity with rigor. The typical marketing approach is to claim something is both rigorous and creative.
- In defining rigor and creativity, engineering and basic science approaches are different because engineers commonly work on a product.
- What does "balancing" mean? Both creativity and rigor are necessary for success.
- Creativity is still first. No matter how rigorously you approach a problem you will not necessarily have a good idea/solution.

- Rigor is not necessarily valued in industry unless it helps with a problem or it leads to some improvement.

How does innovation intersect with creativity and rigor?

- Innovation can be based on either creativity or rigor or both.
- Innovation can arise from rigor by defining a problem thoroughly and understanding the fundamental issues.
- There is a process that cycles between creativity and rigor, and this is probably necessary to create realistic products/developments.
- Education may favor rigor over creativity.
- Intuition is necessary for creativity, but learning the fundamentals is important to becoming creative.
- Rigor allows you to extend your creativity and to be creative in other fields.
- You can be creative without being rigorous.
- The Ph.D. process inherently requires both rigor and creativity.
- If creativity and rigor were separated there could be more creative ideas.
- Creative work may be high risk, but the potential payoff may also be very high.
- There is a theory of inventive processes developed in Russia. They looked at the patent literature and asked what is required to come up with patents. One can develop a checklist that helps people be more creative but if it becomes a prescription for creativity, people will react negatively to it.
- At the initial stage, ideas can be checked with back-of-the-envelope calculations to eliminate fundamentally flawed ideas.

Individual personalities and strengths

- Different cultures place different values on creativity and rigor. In the U.S., there seems to be more of a balance between creativity and rigor.
- Give people the environment to be creative. Pushing them one way or the other may not be very effective.
- Those focused on rigor should have the patience to listen to creative people.
- High-level vision requires balance in creativity and rigor.
- The educational system is also influenced by rigor and creativity. In the United States, the education system is less rigorous but more creative.
- Developing teams with both creative and rigorous people is important.
- Personality types tests may be used (Belban test).
- Interpersonal skills are important within a team.
- People come from many different backgrounds, and it is important to understand this for a well-functioning team.

Summary observations

- In terms of relative risk between rigor and creativity, creativity is more risky.
- In creating a portfolio that balances creative (high-risk) and rigorous (low-risk) ideas, there is a higher probability of success even for a riskier idea if the idea is presented in terms of a hypothesis that is based on previous knowledge. There is ample creativity in ideas within the paradigm even though we tend to acknowledge paradigm-shifting creativity more. The latter only comes once in a while, such as Einstein's vs. Newton's theories about physics.
- Downtime may be necessary to come up with a creative idea.
- Determining what the questions are rather than what the solutions are, is very critical to being creative. This is especially important across different fields.

Group 3

Is there a tension (need for balance) between rigor and creativity? In industry where the product design process has a tight schedule and product testing is required, rigor is necessary. In the academic environment/engineering education, the body of knowledge and tools is growing fast. What constitutes a well-rounded engineering education? With tools like Matlab, does the student still need to learn inversion of a 3-by-3 matrix by hand? Freshman design classes that let students go through a creative process where they can fail enhance creativity and promote rigor. Later, more tools will be taught. In senior design class the students can balance creativity and rigor (risk) more conscientiously.

The balance between rigor and creativity is industry specific. In typical civil engineering projects such as bridge building, a premium is put on eliminating risk, which demands rigor. In software and high-tech industries, a premium is often put on new feature sets, which demands more creativity. In the natural progression of any industry from high-margin unsaturated capacity to a low-margin mature market with saturated production capacity, there is a continuous change in the emphasis on risk vs. creativity. Practices in the automotive industry may have been highly creative in the 1910s, but may be less so at present. (Reuse is strongly encouraged to save cost and time in product validation, which often takes 5-10 years.) In space and other risk-averse industries, there is often a 15-year lag between commercial technologies that are used now and state-of-the-art.

Market pull defines what kind of engineer is in demand. Changes in market conditions and industry maturity make the requirements a moving target. In engineering education, students will be trained to have many skill sets but soft skills such as the ability to communicate well are also important. One current problem in education is the hesitation to give negative feedback, i.e., As are given too readily and participation counts more than results. This may have a negative impact on rigor.

Corporate research (AT&T Bell Labs, Xerox PARC) has been very successful in producing new knowledge and high returns on investment (ROI), but these returns are often very long term such that the historical ROI for a company has not been as impressive. Most large corporations now seek short-term performance, quarter to quarter or at most a few years, and buy successful technologies instead of developing them. Is there a future for corporate research, especially fundamental research? Should there be separation in roles so that creativity is the domain of universities and venture capital companies, and less creativity and more rigor is the purview of large corporations? This could have fundamental importance to the engineering students we train. In terms of the U.S. role in the world, can we maintain a creative edge over the rest of the world while still imparting sufficient skills to our students?

The value of multidisciplinary research is that sometimes creativity comes by just bringing the right people together, for example business development people with engineers, where there is no such thing as a stupid question. Solving the energy problem is an example of this because it demands effort from many disciplines.

TOPIC 5: HOW CAN WE BETTER PREPARE PH.D. ENGINEERS FOR THE COMPETITIVE GLOBAL MARKETPLACE?

Seventy percent or more of Ph.D. engineers in the United States were born and educated (B.S. level) outside the United States. As a result, they are already exposed to foreign culture and are more globally savvy than domestic students. Overseas experience is important and typically missing at the B.S. level; therefore, international experience should be mandatory for B.S. students and incorporated into the engineering curriculum.

Domestic students at the B.S. level should be encouraged to do a semester of international work or a project with an international component for which they receive course credit. B.S. and graduate students in engineering should obtain experience in business management to be better prepared for global markets. Many companies such as Dow have education programs to expose their new Ph.D. engineers to business management. At the University of Illinois some undergraduate students are involved with organization such as Engineers Without Borders that conduct international engineering projects in developing countries. At Carnegie Mellon University an engineering course called Tech Bridge is taught simultaneously to students in the United States, Greece, and Portugal, and students are encouraged to cross-exchange for a couple of weeks among these countries. One way to give students international experience is through post-doctoral fellowships in foreign countries. Federal grants should require domestic graduate students to spend some time conducting research overseas.

Most domestic students have good hands-on experience but are not as driven as international students. There is a need to have students involved in engineering and science through outreach programs at high school and middle school. Faculty members also should have international experience. Some universities organize immersion trips for early- or mid-career faculty members to understand global issues related to bioenergy, bioproducts, agriculture, and climate change. Frequently, domestic undergraduate students in engineering do graduate work in business management, medicine, or law, and exposing students earlier to contextual issues in engineering may help retain them in the engineering profession.

Innovative skills are required to be globally competitive. If curiosity can be developed in students, innovation usually follows. Developing and teaching courses in entrepreneurship and innovation to engineering students is needed because interdisciplinary programs enhance innovation. The current process of education does not allow cross-pollination of students from different disciplines. Graduate students from different engineering and science disciplines should be encouraged to work together on interdisciplinary problems. In engineering schools throughout the United States, there is discussion about reducing the number of credit hours required for graduation, yet we need to add more courses to give students international exposure and teach innovation. There is a mismatch between these two strategies.

Companies prefer to hire students trained in core disciplines and do not want students trained in less-focused interdisciplinary themes. Companies foster interdisciplinary skills in their employees by forming teams of experts from different fields. They emphasize communication skills so their employees can function well in these teams. Companies are expanding globally in terms of both markets and talent. What role can government play in preparing engineers for the global market? What policies can be put in place for federal funding to encourage global competitiveness of engineering students? Development of other skills (e.g., communication and confidence) is important for engineers to be globally competitive.

TOPIC 6: RESPONSIBILITIES OF ENGINEERS AS MEMBERS OF SOCIETY—WHAT ARE THEY? HOW DO WE FULFILL THEM?

Responsibilities

- To innovate—create/improve/develop engineering artifacts.
- Competitiveness—ensure the country can compete and advance in the global economy and maintain and improve the standard of living.
- Responsible creation—safe, sustainable engineering, conservation, minimal impact.
- Inform the public debate, provide engineering facts and point of view.

- Provide society with engineering expertise to organize complex, large projects.
- Educate engineering professionals.
- Educate the general public on the virtues of engineering.
- Provide policy guidance and advice.
- Adhere to a code of conduct (ethics, standards, licensing, etc.).

Observations

- Engineering enrollment is down at colleges.
- Engineering public relations should be improved. People routinely rely on devices created by engineers but don't understand the connection between engineering and the devices they use.
- Massachusetts Institute of Technology enrolls students who can also communicate and interact well socially.
- Supply chain management (SCM) is largely viewed as management innovation, but engineers solve SCM problems.
- The National Academy of Sciences does a decadal survey of physics, and perhaps this should be done for engineering.
- The Engineering Marvels series is a good example of public relations.
- Intel Inside is a good example of advertising intrinsic capabilities.
- Engineers tend to communicate poorly to the general populace because their terminology is very domain specific.

Recommendations

1. NAE should select an Engineering Grand Challenge, such as:

- Self-sustaining energy across a spectrum of production, transport, and conservation. This includes *energy production,* such as artificial photosynthesis, biomass fuel, solar photovoltaic, geothermal, tidal, nuclear (fission, fusion, decay); *energy transport,* such as room temperature superconductors, energy dense materials (gallium/aluminum, synthetic fuels), fuel cells, new batteries (better lithium, organic, etc.), microwave, laser; and *energy conservation,* such as passive solar architecture, improved electric motors, heating, cooling, lighting, and computing, networks.
- Ubiquitous information access: networked information, significant content, and advanced organization of information to all locales.
- Space exploration, including solar system bodies (analysis, protection, utilization, colonization, asteroids, comets, moons, planets); robotic exploration; and search for life (terrestrial planet finder, extra-solar missions and surveys).

2. NAE should improve public relations by

- Promoting engineering in commercials and through other materials similar to Intel Inside, where the origin of significant advancements is made clear through advertising, and individual engineers or teams and their background are emphasized. Products or areas were this could be utilized are: the integrated circuit, ARPANet, computer graphics, moon missions, personal computers, nuclear fission power, powered flight, and medical scanning.

- Helping shape public policy on engineering education and other initiatives through interactions with presidential candidates and Congress and facilitating improvements to national engineering capabilities through education incentives, policy, and regulations. For example, help establish engineering education curricula that include engineering public relations beyond the usual engineering communications classes where engineers learn how to write a technical paper.

Contributors

Carina Maria Alles leads Engineering Evaluations and Sustainability at DuPont Engineering Research and Technology in Wilmington, Delaware. Her team of corporate technical consultants evaluates novel products and processes with regard to technical feasibility, economic merit, environmental footprint, and societal benefit in support of DuPont's strategic technology, research, and business decisions. Previously, she was a research fellow at the Universität Karlsruhe in Germany and a visiting scientist at the University of Houston and the University of Massachusetts, Amherst. Dr. Alles holds a doctorate in chemical engineering from the Universität Karlsruhe, with graduate-level training in process engineering, environmental engineering, and biotechnology in Germany, France, Switzerland, and the United States.

Jess C. Brown is manager of the Carollo Engineers Research Group, where he leads Carollo Engineers' biological drinking water treatment program. He specializes in drinking water processes and applied research and has extensive experience with analytical chemistry, water-quality testing methods, and process selection. He has performed bench-, pilot-, and full-scale studies covering conventional treatment to advanced processes. Dr. Brown is a member of the American Water Works Association (AWWA) for which he chairs the Inorganic Contaminants Research Committee. He is also a member of the Florida section of the AWWA Biological Contaminants Committee, the International Water Association, and the Water Environment Federation. He received a B.A. in environmental science and public policy from Harvard University, and a B.S. in civil engineering and an M.S. and a Ph.D. in environmental engineering from the University of Illinois at Urbana-Champaign.

Amy E. Childress is an associate professor in the Department of Civil and Environmental Engineering and director of the Environmental Engineering Program at the University of Nevada, Reno. In 2004, she was a visiting associate professor at the University of New South Wales in Sydney, Australia. Dr. Childress teaches an undergraduate introductory course in environmental engineering and graduate courses in physical and chemical processes in water and wastewater treatment and colloidal and interfacial processes. Her research focuses on membrane contactor processes, pressure-driven membrane processes, and membrane bioreactor technology. Dr. Childress is president-elect of the Association of Environmental Engineering and Science Professors and serves on the advisory board of *Desalination*. She is also a member of the American Society of Civil Engineers, the American Chemical Society, the American Water Works Association, and the North American Membrane Society. She was recipient of a National Science Foundation CAREER Award in 2001. She received M.S. and Ph.D. degrees in environmental engineering from the University of California, Los Angeles, in 1993 and 1997, respectively.

Bruce S. Dien is a biochemical engineer with the U.S. Department of Agriculture's National Center for Agricultural Utilization Research in Peoria, Illinois. In this position, he conducts process development for conversion of corn fibrous streams from the corn milling industry into ethanol. He has served on several multi-instructional research projects in this regard with other federal laboratories, as well as with industrial and academic partners. Dr. Dien is also part of an interdisciplinary team that developed microorganisms for converting sugars mixtures to ethanol or lactic acid at yields greater than 90 percent of theoretical. He received a B.S. in food process engineering and in biochemistry from Purdue University (1988) and a Ph.D. in chemical engineering from the University of Minnesota (1994).

Edward W. Felten is a professor of computer science and public affairs at Princeton University and is the founding director of Princeton's Center for Information Technology Policy. His research interests include computer security and privacy, especially relating to media and consumer products, and technology law and policy. He has published about 80 papers in the research literature and two books. His research on topics such as web security, copyright and copy protection, and electronic voting has been covered extensively in the popular press. Dr. Felten's weblog at <http://www.freedom-to-tinker.com> is widely read for its commentary on technology, law, and policy. He was the lead computer science expert witness for the U.S. Department of Justice in the Microsoft antitrust case, and he has testified in other important lawsuits. He has testified before the Senate Commerce Committee on digital television technology and regulation and before the House Administration Committee on electronic voting. In 2004, *Scientific American* magazine named him to its list of 50 worldwide science and technology leaders. He received a B.S. in physics from the California Institute of Technology (1985), and an M.S. (1991) and a Ph.D. (1993) in computer science and engineering from the University of Washington.

Kevin A. Gluck is a senior research psychologist at the Air Force Research Laboratory's Mesa (Arizona) Research Site, where he is the team lead for the Performance and Learning Models (PALM) research team. The team's mission is to improve scientific understanding of human perceptual, cognitive, and motor processes by conducting empirical research and creating computational and mathematical models of results and phenomena of interest. Current research areas include spatial competence, the effects of fatigue on cognition, language-enabled synthetic entities, the micro-dynamics of learning and forgetting, and the development of large-scale computational infrastructures that enable faster scientific progress in cognitive science. Applications of interest include synthetic teammates, pedagogical agents, and analysis tools for performance optimization. Dr. Gluck is also the conference officer for the Cognitive Science Society. He received a B.A. in psychology from Trinity University (1993) and an M.S. (1997) and a Ph.D. (1999) in cognitive psychology from Carnegie Mellon University.

Laurent Itti is an associate professor of computer science, psychology, and neuroscience in the Computer Science Department at the University of Southern California. Dr. Itti's research interests are in biologically inspired computational vision, particularly in the domains of visual attention, scene understanding, control of eye movements, and surprise. This basic research has technological applications to areas including video compression, target detection, homeland security, and robotics. He has co-authored over 75 publications in peer-reviewed journals, books, and conferences; holds two patents; and developed an open-source neuromorphic vision software toolkit used by more than 1,750 research groups and individuals worldwide. He also led the development of the first encyclopedia of attention research. Dr. Itti received an M.S. in image processing from the Ecole Nationale Supérieure des Télécommunications in Paris, France (1994) and a Ph.D. in computation and neural systems from the California Institute of Technology (2000).

Matthew J. Lang is an assistant professor of mechanical engineering and biological engineering at the Massachusetts Institute of Technology. His research interests are in molecular biophysics and functional measurement of molecular and cellular machinery. He was the Jane Coffin Childs Postdoctoral Fellow in the Departments of Biological Sciences and Applied Physics at Stanford University, a postdoctoral fellow in the Department of Molecular Biology and the Princeton Materials Institute at Princeton University, and a postdoctoral fellow in the Department of Chemistry and the Physical Biosciences Division of Lawrence Berkeley National Laboratory at the University of California, Berkeley. Dr. Lang received a B.S. from the University of Rochester and a Ph.D. from the University of Chicago.

Karl G. Linden is a professor and Liebman Faculty Fellow in the Department of Civil, Environmental, and Architectural Engineering at the University of Colorado at Boulder. He teaches classes on fundamentals of environmental engineering, physical and chemical treatment processes, and fundamentals and applications of ultraviolet (UV) processes in environmental systems. Dr. Linden's research has focused on the efficacy of UV irradiation for inactivation of microbial pathogens, the use of UV and oxidation technologies for the degradation of environmental pollutants in water, and innovative technologies for water reuse. In 2004, he received the Klein/Stansell Family Distinguished Research Award and was a Switzer Environmental Foundation Leadership Fellow from 2001-2003. Dr. Linden's current research is funded by the American Water Works Association Research Foundation, the Water Reuse Foundation, the U.S. Environmental Protection Agency, the U.S. Department of Agriculture, the National Institute of Environmental Health Sciences, and various utility and industrial sponsors. He is the author of over 70 peer-reviewed publications, serves as an associate editor of the *ASCE Journal of Environmental Engineering*, and is a founding board member and an international vice president of the International Ultraviolet Association. Dr. Linden received a B.S. from Cornell University and an M.S. and a Ph.D. from the University of California, Davis.

Sanjay V. Malhotra is head of the Laboratory of Synthetic Chemistry at the National Cancer Institute in Frederick, Maryland. He is also a visiting scientist at the National Institutes of Standards and Technology and at Brookhaven National Lab. Previously, he was an assistant professor in chemistry and environmental science at the New Jersey Institute of Technology. Since 2000, he has served on the editorial board of *Catalysis in Organic Reactions*. In 2006, he received the American Chemical Society-New Jersey Chapter Pro Bono Award. Dr. Malhotra received M.S. (1993) and Ph.D. (1995) degrees in chemistry from Seton Hall University in South Orange, New Jersey.

Greg Morrisett is associate dean of Computer Science and Engineering and Allen B. Cutting Professor of Computer Science at Harvard University. Previously, he spent 7 years on the faculty of the Computer Science Department at Cornell University. In 2002-2003, he took a sabbatical at Microsoft's Cambridge Research Laboratory. He is serving on the editorial board for *Information Processing Letters,* on DARPA's Information Science and Technology Board, and on Fortify Software's technical advisory board. He also serves on numerous Association for Computing Machinery steering and program committees. Dr. Morrisett received a B.S. in mathematics and computer science from the University of Richmond (1989) and a Ph.D. in computer science from Carnegie Mellon University (1995).

Rama Ranganathan is an associate professor of pharmacology in the Graduate School of Biomedical Sciences and director of the Green Center for Systems Biology at the University of Texas Southwestern Medical Center in Dallas. He is also affiliated with the Howard Hughes Medical Institute. The goal of his lab is to understand the process of information flow in cell signaling systems at the atomic level. Dr. Ranganathan did post-doctoral research in neurobiology at Harvard Medical School and in structural biology at the Salk Institute. He received a B.S. in bioengineering from the University of California, Berkeley (1985) and a Ph.D. in biology and an M.D. from the University of California, San Diego (1994).

Diana K. Smetters is a senior member of the research staff at the Palo Alto Research Center (formerly Xerox PARC). Her research focuses on usability of security, and in particular how to design new cryptographic and security technologies that make it easier to build secure systems that are easy to use. Other areas of interest include approaches to key distribution and key management, as well as security for mobile and wireless devices. Before joining PARC in 1999, she worked at CertCo, Inc., designing and building commercial systems to support high-security public key infrastructures. Dr. Smetters received a B.A. in cognitive science from Ohio State University and a Ph.D. from the Massachusetts Institute of Technology in computational and experimental neuroscience. She did post-doctoral work at the Salk Institute and Columbia University.

Vanessa L. Speight is an associate with Malcolm Pirnie, Inc. in Cincinnati, Ohio, where she has performed research on water-quality sampling plans, disinfection byproduct occurrence and prediction, and disinfectant decay on projects funded by the American Water Works Association Research Foundation (AwwaRF), the U.S. Environmental Protection Agency (EPA), the Centers for Disease Control and Prevention, and individual water utilities. Her expertise is regularly sought as a member of project advisory committees for the American Water Works Association (AWWA), AwwaRF, and the WateReuse Foundation; on technical review panels for EPA; and as a peer reviewer for professional journals. She is a member of AWWA's Technical Advisory Workgroup on revisions to the Total Coliform Rule, chair of the AWWA Distribution Research Committee, and a member of planning committees for several leading industry conferences. Dr. Speight received a B.S. in civil engineering from McGill University (1993) and an M.S. and a Ph.D. in environmental engineering from the University of North Carolina at Chapel Hill in 1995 and 2003, respectively.

Michael van Lent is the associate director for games research and a research assistant professor at the Institute for Creative Technologies (ICT) at the University of Southern California. Since joining ICT in 2001, he has been the lead researcher for a number of game-based training tools including Full Spectrum Warrior and Full Spectrum Command. His research interests include adaptive intelligence for games, social and cultural simulation, and the use of game technology for learning and training. Dr. van Lent has also been active in building a link between artificial intelligence (AI) researchers and game developers as the first editor-in-chief of the *Journal of Game Development*, as an organizer of the AI and Interactive Digital Entertainment Conference, and as a frequent presenter at the Game Developers' Conference. He received a B.A. in computer science from Williams College, an M.S. in computer science from the University of Tennessee-Knoxville, and a Ph.D. in computer science from the University of Michigan, Ann Arbor.

Rebecca N. Wright is an associate professor of computer science at Rutgers University. She is also deputy director of the DIMACS Center for Discrete Mathematics and Theoretical Computer Science. Previously, she was a professor of computer science at Stevens Institute of Technology and a researcher in the Secure Systems Research Department at AT&T Labs and AT&T Bell Labs. Her research is in the area of information security, including cryptography, privacy, foundations of computer security, and fault-tolerant distributed computing. Dr. Wright serves as an editor of the *Journal of Computer Security* and the *International Journal of Information and Computer Security*, and she was a member of the board of directors of the International Association for Cryptologic Research from 2001 to 2005. She is currently serving on the National Academies' Committee on State Voter Registration Databases. She received a B.A. from Columbia University (1988) and a Ph.D. in computer science from Yale University (1994).

DINNER SPEAKER

Henrique (Rico) Malvar is a Distinguished Engineer at Microsoft and the managing director of Microsoft Research in Redmond, Washington. He received a Ph.D. from the Massachusetts Institute of Technology in electrical engineering and computer science. Before joining Microsoft in 1997, he was vice president of research and advanced development at PictureTel Corporation, and prior to that he was with the faculty of the University of Brasilia for 14 years. As the managing director at Microsoft Research, he oversees the activities of over 25 research groups in many areas related to computer science. His technical interests include multimedia signal compression and enhancement, fast algorithms, multi-rate filter banks, and wavelet transforms. He has over 140 publications and over 70 patents in those areas. Dr. Malvar is a fellow of the IEEE, a member of the editorial board of the journal *Applied and Computational Harmonic Analysis*, and a former associate editor of the journal *IEEE Transactions on Signal Processing*. He received the Young Scientist Award from the Marconi International Fellowship in 1981, the Best Paper Award in Image Processing from the IEEE Signal Processing Society in 1992, the 2002 Technical Achievement Award from the IEEE Signal Processing Society, and the Wavelet Pioneer Award from the SPIE in 2004.

Program

NATIONAL ACADEMY OF ENGINEERING

2007 U.S. Frontiers of Engineering Symposium
September 24-26, 2007

Chair: Julia M. Phillips, Sandia National Laboratory

ENGINEERING TRUSTWORTHY COMPUTER SYSTEMS

Organizers: Annie I. Antón and John Dunagan

Privacy in a Networked World
Rebecca N. Wright

Unifying Disparate Tools in Software Security
Greg Morrisett

Usable Security: Oxymoron or Challenge?
Diana K. Smetters

Freedom to Tinker: The Struggle to Access Devices You Own
Edward W. Felten

***

CONTROL OF PROTEIN CONFORMATIONS

Organizer: Donald J. Leo

The Evolutionary Design of Proteins
Rama Ranganathan

Lighting Up the Mechanome
Matthew J. Lang

***

BIOTECHNOLOGY FOR FUELS AND CHEMICALS

Organizers: Richard T. Elander and Vijay Singh

Corn-Based Materials
Sanjay V. Malhotra

Process Review of Lignocellulose Biochemical Conversion to Fuel Ethanol
Bruce S. Dien

Sustainable Biorefineries
Carina Maria Alles

***

MODELING AND SIMULATING HUMAN BEHAVIOR

Organizers: Christian Lebiere and Robert Wray

Computational Cognitive Neuroscience and Its Applications
Laurent Itti

Barriers, Bridges, and Progress in Cognitive Modeling for Military Applications
Kevin A. Gluck

Modeling of Culturally Affected Human Behavior
Michael van Lent

***

SAFE WATER TECHNOLOGIES

Organizers: Carol R. Rego and Paul K. Westerhoff

Ultraviolet Irradiation: An Age-Old Emerging Technology for Water Treatment
Karl G. Linden

Membrane Processes to Address the Global Challenge of Desalination
Amy E. Childress

Biological Treatments of Drinking Water
Jess C. Brown

Distribution Systems: The Next Frontier
Vanessa L. Speight

***

DINNER SPEECH

Microsoft's Future: An Overview of Microsoft Research
Henrique (Rico) Malvar

Participants

NATIONAL ACADEMY OF ENGINEERING

2007 U.S. Frontiers of Engineering Symposium
September 24-26, 2007

Dongchan (Shaun) Ahn
Research Engineering Specialist
Business & Technology Incubator
Dow Corning Corporation

Vitaly Aizenberg
Advanced Scientific Associate
ExxonMobil Biomedical Sciences
ExxonMobil Corporation

Carina Maria Alles
Engineering Consultant
DuPont

Matthew Andrews
Member of Technical Staff
Mathematical and Algorithmic Sciences Center
Alcatel-Lucent, Bell Labs

Ana I. Antón
Associate Professor
Department of Computer Science
North Carolina State University

David A. Bader
Executive Director of High-Performance Computing
Associate Professor
College of Computing
Georgia Institute of Technology

Randy A. Bartels
Associate Professor
Department of Electrical and Computer Engineering
Colorado State University

Jennifer T. Bernhard
Associate Professor
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

Joshua Binder
Advanced Structural Analysis Manager
Phantom Works: Materials and Structures Technology
Boeing Company

Jess Brown
Project Manager for Biological Treatment
Carollo Engineers, P.C.

Markus J. Buehler
Assistant Professor
Department of Civil and Environmental Engineering
Massachusetts Institute of Technology

Barrett S. Caldwell
Associate Professor
School of Industrial Engineering
Purdue University

Wai (Victor) Kin Chan
Assistant Professor
Department of Decision Sciences and Engineering Systems
Rensselaer Polytechnic Institute

Jia Chen
Research Staff Member
IBM T. J. Watson Research Center
IBM Corporation

Amy Childress
Associate Professor
Department of Civil Engineering
University of Nevado, Reno

Eric Pei-Yu Chiou
Assistant Professor
Department of Mechanical and Aerospace Engineering
University of California, Los Angeles

Alfred J. Crosby
Assistant Professor
Polymer Science and Engineering
University of Massachusetts, Amherst

Rula A. Deeb
Senior Associate
Malcolm Pirnie, Inc.

Shrikant Dhodapkar
Technical Leader
Process Fundamentals Group, Performance Plastics and Chemicals
Dow Chemical Company

Bruce Dien
Chemical Engineer
National Center for Agricultural Utilization Research
U.S. Department of Agriculture

Peter A. Dinda
Associate Professor
Department of Electrical Engineering and Computer Science
Northwestern University

John Dunagan
Researcher
Distributed Systems and Security Research Group
Microsoft Research

Richard T. Elander
Advanced Pretreatment Team Leader
BioProcess Engineering Group
National Bioenergy Center
National Renewable Energy Laboratory

Michael J. Fasolka
Director, Combinatorial Methods Center
Polymers Division
National Institute of Standards and Technology

Edward W. Felten
Professor of Computer Science and Public Affairs
Director, Center for Information Technology Policy
Princeton University

Andrew Fernandez
Member of Technical Staff
Agilent Labs
Agilent Technologies

Timothy Fisher
Associate Professor
School of Mechanical Engineering
Purdue University

Raja N. Ghanem
Senior Scientist
Cardiac Rhythm Disease Management
Medtronic Inc.

Reza Ghodssi
Associate Professor
Department of Electrical and Computer Engineering
Institute for Systems Research
University of Maryland

Rajat S. Ghosh
Senior Project Leader
Sustainable Production Technology
Alcoa Inc.

Anouck Girard
Assistant Professor
Department of Aerospace Engineering
University of Michigan, Ann Arbor

Kevin Gluck
Senior Research Psychologist
Air Force Research Laboratory

Samuel Graham
Assistant Professor
Department of Mechanical Engineering
Georgia Institute of Technology

Tyrone W. Grandison
Research Manager
Data Privacy and Mining Technologies
IBM Almaden Research Center

Johney Green
Group Leader
Fuels, Engines, and Emissions Research Center
Oak Ridge National Laboratory

Michelle L. Gregory
Research Scientist
Computational and Information Science Directorate
Pacific Northwest National Laboratory

William J. Grieco
Director, Process Development
Alkermes

Gregory A. Hebner
Manager
Lasers, Optics, Remote Sensting, Plasma Physics, and Computer Systems Department
Sandia National Laboratories

John Hecht
Technology Leader, Drying and Mass Transfer
Corporate Engineering
Procter & Gamble

Hugh W. Hillhouse
Associate Professor
Department of Chemical Engineering
Purdue University

Laurent Itti
Associate Professor
Department of Computer Science
University of Southern California

Jeffrey S. Kanel
Senior Research Associate
Eastman Research Division
Eastman Chemical Company

Jeffrey M. Karp
Instructor of Medicine
Department of Medicine at the Brigham and Women's Hospital
Harvard-MIT Division of Health Sciences and Technology
Harvard University

Theodore J. Kim
Distinguished Member of the Technical Staff
Navigation, Pointing, and Control Department
Sandia National Laboratories

Fred A. Kish
Vice-President, Research, Development, and Manufacturing
Photonic Integrated Circuit Department
Infinera Corporation

Efrosini Kokkoli
Assistant Professor
Department of Chemical Engineering and Materials Science
University of Minnesota

Ilya Kolmanovsky
Technical Leader
Powertrain Control R&D
Ford Research and Advanced Engineering

Michael R. Krames
Director
Advanced Laboratories
Philips Lumileds Lighting Co.

Raj Krishnaswamy
Senior Scientist
Polymer Product Development
Metabolix

Rajesh Kumar
Intel Fellow and Director, Circuit & Low Power Technologies
Digital Enterprise Group
Intel

Sanjay Lall
Assistant Professor
Department of Aeronautics and Astronautics
Stanford University

Matthew J. Lang
Assistant Professor
Department of Mechanical Engineering
Massachusetts Institute of Technology

Christian Lebiere
Research Scientist, Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University

Eugene J. LeBoeuf
Associate Chair and Associate Professor
Department of Civil and Environmental Engineering
Vanderbilt University

K. Rustan M. Leino
Principal Researcher
Programing Languages and Methods
Microsoft Research

Donald J. Leo
Associate Dean of Research and Graduate Studies
College of Engineering
Virginia Tech

Philip Levis
Assistant Professor
Department of Computer Science and Electrical Engineering
Stanford University

Mark E. Lewis
Associate Professor
School of Operations Research and Information Engineering
Cornell University

Ju Li
Associate Professor
Department of Materials Science and Engineering
University of Pennsylvania

Karl G. Linden
Professor and Liebman Faculty Fellow
Department of Civil, Environmental, and Architectural Engineering
University of Colorado at Boulder

Jennifer Lukes
William K. Gemmill Assistant Professor
Department of Mechanical Engineering and Applied Mechanics
University of Pennsylvania

Yiorgos Makris
Associate Professor
Department of Electrical Engineering
Yale University

Sanjay Malhotra
Head, Laboratory of Synthetic Chemistry
Developmental Therapeutic Program Support
National Cancer Institute

Ajay P. Malshe
21st Century Endowed Chair Professor
Department of Mechanical Engineering
University of Arkansas

Michele Marcolongo
Associate Professor/Associate Vice Provost for Research
Department of Materials Science and Engineering
Drexel University

Wade Martinson
Principal Engineer
Process Solutions Technology Development Center
Cargill Inc.

Katherine McMahon
Assistant Professor
Department of Civil and Environmental Engineering
University of Wisconsin-Madison

Swarup Medasani
Senior Research Engineer
Information Systems and Sciences Laboratory
HRL Laboratories, LLC

Greg Morrisett
Allen B. Cutting Professor of Computer Science
Division of Engineering and Applied Science
Harvard University

Wilbur L. Myrick
Assistant Vice President, Senior Analyst
Intelligence, Security, and Technology Group
Science Applications International Corporation (SAIC)

Priya Narasimhan
Associate Professor
Department of Electrical and Computer Engineering
Carnegie Mellon University

Ravi Narasimhan
Assistant Professor
Department of Electrical Engineering
University of California, Santa Cruz

Roseanna M. Neupauer
Assistant Professor
Department of Civil, Environmental, and Architectural Engineering
University of Colorado

Sissy Nikolaou
Associate
Mueser Rutledge Consulting Engineers

Burak Ozdoganlar
Assistant Professor
Department of Mechanical Engineering
Carnegie Mellon University

Jianbiao Pan
Assistant Professor
Department of Industrial and Manufacturing Engineering
California Polytech State University

Alexander Parkhomovsky
Senior Engineering Manager
Motor Design Division
Seagate Technology, LLC

Julia M. Phillips
Director
Physical and Chemical Sciences Center
Sandia National Laboratories

Arshan Poursohi
Staff Researcher
SunLabs
Sun Microsystems

Lisa Purvis
Research Lab Manager
Xerox Innovation Group
Xerox Corporation

Marcus Quigley
Associate
Water Resources Engineering
Geosyntec Consultants

R. Michael Raab
President and CEO
Corporate Development and Research
Agrivida, Inc.

Rama Ranganathan
Associate Professor
Graduate School of Biomedical Sciences
University of Texas Southwestern Medical School

Carol R. Rego
Vice President
CDM

Lee Rosen
Manager, Combustion Applications R&D
Research and Development
Praxair, Inc.

Rachel A. Segalman
Charles Wilke Assistant Professor
Department of Chemical Engineering
University of California, Berkeley

Deborah Shands
Engineering Specialist
Information Assurance Technology Department
Aerospace Corporation

Samuel Sia
Assistant Professor
Department of Biomedical Engineering
Columbia University

Vijay Singh
Associate Professor
Department of Agricultural and Biological Engineering
University of Illinois at Urbana-Champaign

John E. Smee
Senior Staff Engineer, Manager
Corporate Research and Development
QUALCOMM, Inc.

Diana K. Smetters
Senior Member of the Research Staff
Computer Science Laboratory
PARC

Vanessa L. Speight
Associate
Malcolm Pirnie, Inc.

Randy Stiles
Principal Investigator
ATC Modeling, Simulation & Information Sciences
Lockheed Martin Space Systems Company

Michael S. Strano
Associate Professor
Department of Chemical Engineering
Massachusetts Institute of Technology

Joseph Szczerba
Staff Researcher
Vehicle Development Research Lab
General Motors Corporation

David N. Thompson
R&D Engineer III
Department of Biological Sciences
Idaho National Laboratory

Jean W. Tom
Director
Development Engineering/Process RD
Bristol-Myers Squibb

Michael van Lent
Research Assistant Professor and Project Leader
Institute for Creative Technologies
University of Southern California

Sundeep N. Vani
Technical Director
Bioproducts Division
Archer Daniels Midland

James Scott Vartuli
Chemical Engineer
Optical Materials Lab
GE Global Research

S. Travis Waller
Associate Professor
Fellow of Clyde E. Lee Endowed Professorship
Department of Civil, Architectural, and Environmental Engineering
University of Texas, Austin

Annemarie Ott Weist
Global Lead, Gas Recycle and Recovery
Advanced Technology
Air Products and Chemicals, Inc

Mei Wen
Research Scientist
Functional Additives
Arkema Inc.

Paul K. Westerhoff
Professor
Department of Civil and Environmental Engineering
Arizona State University

Joan Wills
Technical Advisor
Controls and Diagnostics, Advanced Engineering
Cummins, Inc.

Robert Wray
Chief Scientist
Soar Technology

Rebecca N. Wright
Associate Professor
Department of Computer Science
Deputy Director, DIMACS Center
Rutgers University

Lifang Yuan
Senior Systems Engineer/Analyst
Systems Engineering Department
Defense Systems Division
EDO Corporation

Randy Zachery
Director
Mathematical and Information Science Directorate
U.S. Army Research Office

Xin Zhang
Associate Professor
Department of Manufacturing Engineering
Boston University

Victor Zhirnov
Research Scientist
Semiconductor Research Corporation

Guests

Douglas H. Fisher
Program Director, Robust Intelligence Cluster
CISE/IIS
National Science Foundation

Robert Herklotz
Program Manager, Software and Systems
Air Force Office of Scientific Research

Igor Vodyanoy
Program Officer
Life Sciences Research Division
Office of Naval Research

Dinner Speaker

Henrique (Rico) Malvar
Managing Director
Microsoft Research

National Academy of Engineering

Charles M. Vest
President

Lance A. Davis
Executive Officer

Janet Hunziker
Senior Program Officer

Gin Bacon
Senior Program Assistant

Microsoft

Richard Rashid
Senior Vice President
Research

Jane Prey
Program Manager
External Research and Programs

Beth Winkler
External Research and Programs