Grand Challenges for Multimodal Bio-Medical Systems

Jie Chen[1], Edward Dougherty[2], Semahat S. Demir[3], Charles Friedman[4], Chung Sheng Li[5], and Stephen Wong[6]

1. Introduction

The Multimodal Bio-Medical Systems Workshop (MBM) was held on October 22, 2004 in the Lister Hill Auditorium at the National Library of Medicine, Bethesda, Maryland. The workshop was sponsored by the IEEE (Institute of Electrical and Electronic Engineers), National Library of Medicine (NLM), and National Science Foundation (NSF). This workshop was initiated by life sciences and physical sciences experts to identify and highlight new computational paradigms and technology infrastructure in biomedicine owing to recent advance in biosciences and instrumentation. It also allowed for unique interactions among experts of life sciences and physical sciences working on various aspects of multimodal biomedical systems research, whose goals are to collect, process, analyze, integrate, model, synthesize, visualize, and simulate vast amounts of heterogeneous multi-modal, multi-scale data for emerging real world applications in life science as shown in Figure 1.

We sought recommendations on large-scale challenges and their short-term and long-term resolution strategies from these leading scientists and engineers (refer to Figure 2).

The workshop targeted both the micro/nano level and the macro level of biomedical systems. At the micro/nano level, unlike alphanumeric genetics and genomics data, multimodality biological data are increasingly generated by high throughput or high content data acquisition devices, including a broad spectrum of microscopes (e.g., florescence, light, optical, confocal, two-photon laser scanning, time-lapse, laser capture dissection, and electronic), biosensors, and microarrays. This enables the biomedical community, for the first time, a large scale approach to identify and understand the functions and interactions of macromolecules in cells and to decode the intertwined signal pathways and biological mechanisms of disease formation and organ development. Such analysis, extending to work in tissues or clinical samples, offers the potential to speed the identification of toxic compounds during therapeutic drug development and the targeting of drug effects to specific subtypes of cells. The pragmatic goal is to translate scientific discovery into clinical practice, including diagnosis, prognosis, therapeutic prevention, repair, drug development, and disease management, in an effective manner.

At the macro level, on the other hand, environmentally-related activities include global climate changes, deforestation, natural disasters, forest fires, and air pollution. Monitoring disease outbreaks for bio-surveillance public health and their environmental epidemiology has been accomplished for a number of vector-born diseases, such as Hantavirus Pulmonary Syndrome (HPS), malaria, and Dengue fever. To date, health activity monitoring (HAM) concepts have also been applied in the early detection of disease outbreaks by recording subtle human behavior changes. These systems can provide advance warnings before significant casualties occur. The alerts generated from HAM systems can be triggered through the fusion of both traditional and nontraditional multimodal, multiscale heterogeneous data sources. Traditional data includes data generated from clinical sources such as inpatient and outpatient data. Nontraditional data sources include data collected from remote sensing (including satellite images), video/audio surveillance, cellular and microbial analysis, and other biological data, from which we can extrapolate human behavior.

This invitation-only workshop brought together a selected group of about forty worldclass experts in the areas of biology, engineering, and public and environmental health for a full-day interaction and discussion in the Lister Hill Auditorium. The morning session consisted of a series of overview talks. The afternoon sessions focused on talks to set the stage for the discussion.

- Three overview talks: These overview talks covered the interactions between biology and engineering, multiscale modeling in biosystems, and integrative informatics issues in public and environmental health.

[1]Brown University,
[2]Texas A&M University,
[3]NSF and University of Memphis & University of Tennessee,
[4]National Library of Medicine, NIH and University of Pittsburgh,
[5]IBM Research, and
[6]Brigham & Women's Hospital & HCNR, Harvard Medical School

 1540-7977/05/$20.00©2005 IEEE

- Seven focused talks: These talks presented specific, emerging issues of multimodality and multi-scalability in systems biology, drug discovery, public and environmental health, standards and interoperability in public health and life science, as well as new high throughput approaches in proteomics and cytological profiling.

We list the workshop speakers and chairs of breakout sessions in the left hand column of the previous page.

This workshop also included two three-hour breakout sessions at the micro/nano and macro levels. These breakout sessions began with position statements given by individual session chairs, followed by discussion with sponsoring federal agency representatives. The breakout sessions focused on identifying grand scale challenges and drafting white papers based on consensus recommendations established on these challenges. Each challenge was further addressed in the phase of immediate term (in 18–24 months), near term (in 2–5 years) and long term (in 5–10 years). The primary objective of this workshop was to establish consensus to address fundamental research, experimental, and deployment challenges faced by multiscale, multimodal biomedical systems through leveraging common methodologies. More information on this MBM workshop and related activities can be found at (http://www.ieee-lssatc.org).

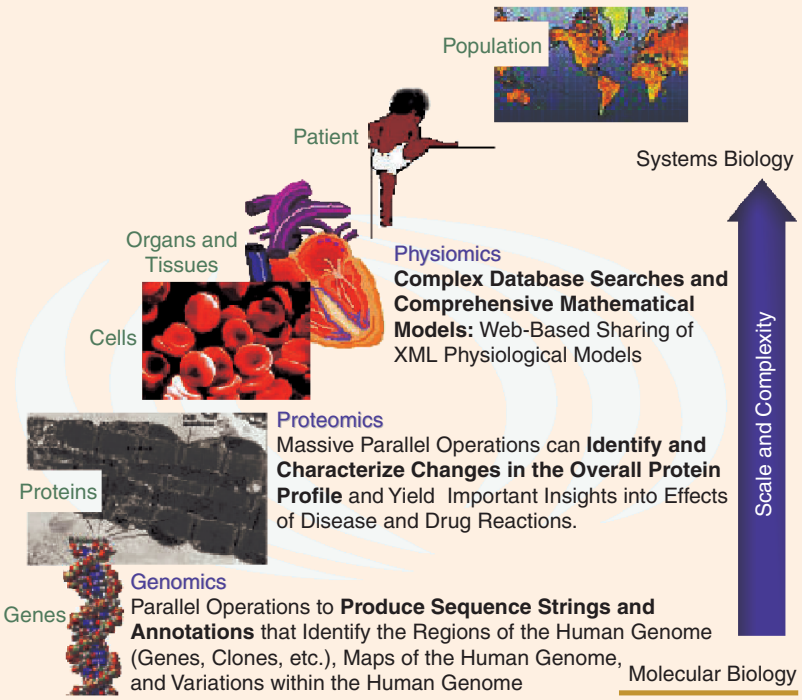


Figure 1. Multimodal and multiscale bio-medical systems.

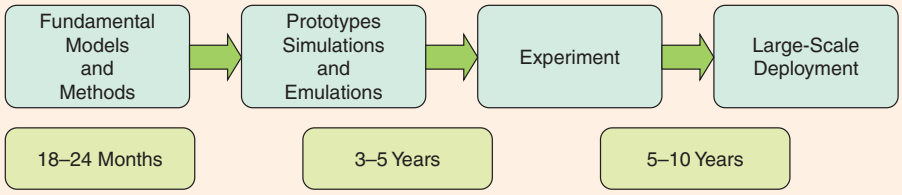


Figure 2. Timeline for addressing challenges.

2. Grand Challenges of MBM Systems

During the workshop, the attendees reached considerable consensus to address the following items as the selected grand challenges of multimodal biomedical systems:

2.1: To Allow Early Detection of Where and When an Infectious Disease Outbreak Occurs, Whether it is Naturally Occurring or Man-Made, in Real Time; This grand challenge should address those commonly occurring infectious diseases such as influenza, and less frequently occurring diseases, such as West Nile encephalitis, Denge Fever, malaria, as well as those made by bioterrorists.

To address this grand challenge, this community needs to develop ontology, semantics, and self-describing data services that cover clinical genomics, proteomics, healthcare/clinical data and public health data. Developing self-describing data and services will enable and greatly simplify the exposure and consumption of content and services, thereby facilitating the development of

large-scale early warning systems for various infectious diseases. Interoperable databases are critical to the success of multiscale and multimodal research. These databases should include (i) standard semantics and ontology so that researchers can agree on the meaning of data; and (ii) standard schema and data models so that investigators can easily share schema and concepts. Note that these two goals can only be achieved in the context of formalized biological knowledge. The data should be scalable or be able to be scaled to larger problems in the future. This community recommends developing infrastructure to enable the seamless sharing of healthcare data in real time, which is another important facet in facilitating the development of large-scale early warning systems (refer to Figure 3). Both of the previously mentioned items are for addressing one urgent research challenge - the lack of sharable data and/or data services.

Many papers published in multimodal or systems biology claim advantages to their proposed approaches. These claims need to be checked and compared, but this is impossible with the current lack of benchmarks. Therefore, it is necessary to include ground truth measurements with which to gauge various methods. The group strongly recommends that benchmarks and ground truth measures be established to evaluate methodologies based on both synthetic data from a wide spectrum of models and on real biological data wherever it is available in sufficient amount to develop statistical test beds. It is also important to create case-focused benchmark data to address the challenge of integrating data from different sources into a coherent model.

To accomplish this grand challenge, we require the following capabilities: First, changing the ownership model of patient data either by care providers or by patients themselves (it works like bank accounts). The separation of patient data from health care providers will greatly impact and dramatically simplify the process of public health data collection and sharing.

Second, the capability of mapping, integrating, and fusing data across macro and micro levels, and even beyond health care data by including clinical genomics data.

Third, the development of various analysis methodologies, including data mining, machine learning, and pattern recognition to identify correlation among various data modalities.

Fourth, the establishment of socially, culturally, and politically acceptable methodologies for data sharing. This should include developing methods for data de-identification and privacy protection. Some members thought that this item by itself could be another grand challenge.

2.2: To Develop Multidimensional Drug Profiling DataBases to Facilitate Drug Discovery and to Identify Biomarkers for Diagnosis and Monitoring the Progress of Individual Disease Treatments.

This grand challenge is similar to the previous one, except the focus is on microscopic cell population as opposed to macroscopic human population. The challenge in the drug discovery area is to generate a comprehensive, multidimensional and multimodal compound profiling database for open access to research communities. Such a database will speed up drug discovery and biological research.

In drug discovery, profiling technologies are used to measure both drug action on a desired target in cell-based assays and drug action on other targets. This profiling should be performed as a function of drug concentration because drug effects are highly dose dependent. For example, the degree to which a primary target is perturbed may affect various downstream pathways differently. Drugs can also bind to multiple targets with different affinities. To date, drug effects have been broadly profiled by transcript analysis,

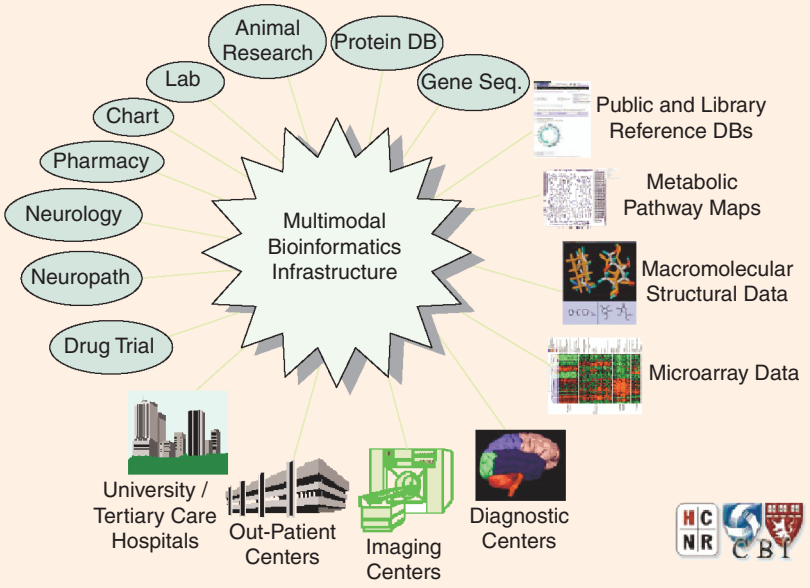


Figure 3. Multimodal Bioinformatics Infrastructure.

proteomics, and the measurement of cell line toxicity dependence. In these studies, multidimensional profiling methods were applied at a singe drug concentration.

Recent advances in automated florescence microscopy have the potential to complement transcriptional and proteomic profiling approaches by allowing fast and economic means to collect data describing protein behaviors and biological pathways within individual cells. Cytometric dose-response profiling is a fast and cheap method for quantitatively surveying broad ranges of individual cell responses. It is conceivable to use the combinations of targeted phenotypic imaging screens to generate the profiles of drug activity. The development of a publicly available database containing large sets of such unbiased measurements might serve as a high dimensional, but simple and inexpensive means to allow extensive cytometric dose-response profile for many drugs.

One can assemble in this database a test set of most available compounds (up to a few hundred thousands) with drugs of known action mechanism, as well as unknown mechanism. The known drug set would be selected to cover common toxicity mechanisms or therapeutic action in cancers and other diseases. It would also be chosen to include several groups with a common target (macromolecular or pathway) but unrelated structures. One can choose tens or even hundreds of distinct probes that cover a comprehensive range of cell biology, thereby collecting images of up to a few thousand cells from each kind. One can then examine the population response of each descriptor to increase the concentration of a given drug, and allow the comparison between dose-response profiles that are independent of the starting dose.

The multidimensional drug profiling database can support methods to assign mechanisms for blind and uncharacterized drugs and to suggest interrelationships between signal pathways. Complex dose-response curves and large cell-to-cell variability frequently observed reinforce the utility of unbiased multidimensional characterization of drug effects over wide dose ranges. The database should include both fixed and live cell imaging as well as be linked and integrated with other profiling databases, e.g., proteomics and transcriptional.

Such drug profiling databases can also be extended to incorporate the response characterization of subpopulations defined by variables, such as cell cycle state, cell density, or neighboring environment. This analysis, extending to tissues or clinical samples, offers great potential to speed the identification of toxic compounds during therapeutic drug development, and the targeting of drug effects to specific cell subtypes.

To address this grand challenge, we need new methodologies and new algorithms for automated cellular image analysis and multidimensional cytological data modeling.

2.3: To Connect Activities and Events Derived from Cellular Processes to High Level Cognitions.

This grand challenge is similar to developing the functional maps of gene sequence and biomedical imaging modalities at the micro scale, such as confocal and light microscopy, and the macroscale, such as PET (Positron Emission Tomography) and fMRI (functional Magnetic Resonance Imaging).

To address this challenge, we need to develop a comprehensive biological atlas of the function and morphology of neurons and glial cells, and the complete characterization of lower level neural physiological and behavior changes. The goal is to define higher level biomarkers for neurological disorders, e.g., dementia, Parkinson's disease, and epilepsy.

2.4: To Support Personalized Medical Care and Clinical Decision for Patients

This grand challenge will require the development of computational models that allow the "*what-if*" analysis of intervention strategies that are optimized for patient biology, personal preference and demographics.

Decision support systems will need capabilities to incorporate, integrate, or fuse data at multiple scales, including genome, proteome, and other clinical phenotypes. They will require data from different sources, including cross generation and genes from other life forms in the related "ecosystem." In particular, they should be able to integrate patient history, individual preference, risk assessment, and public and environmental data in order to customize an optimal solution for personalized patient care. It is expected that many of the new systems will be developed from the interplay of clinical genomics, imaging, and medical informatics.

3. Enabling Technologies

To address the grand challenges listed above, the workshop attendees also considered that it is necessary to adequately address the following technological issues: the development of common methods and informatics infrastructure such as formalism, analytic algorithms, shared databases, interoperability standards, and predictive models that enable us to gain a better understanding of disease mechanisms. The technology will in turn benefit translational research to improve diagnosis, treatment, and disease monitoring. There are also important potential benefits to patients including better life quality, more personalized and lower medical treatment cost, and more therapeutic options for a wide range of diseases. To help realize these benefits, in addition to the

aforementioned technological issues, the group has developed the following additional recommendations:

3.1: Formalization of Biological Knowledge into Predictive Models for Systems Biology and System-Based Analysis

We should devote our focus on developing and validating systems biology models. These models can be inferred from data in conjunction with *a priori* biological knowledge. These models, once developed and fully validated, can then form the basis of future biological knowledge. In order to achieve this goal, we need to develop a methodology to formalize biology knowledge in life sciences in a similar way as in medical informatics, in which the Unified Medical Language System (UMLS) is developed for clinical systems such as electronic medical records and hospital information systems. The methodology will facilitate the formulation of universal definitions of biology knowledge into a formal symbolic structure. Such structure will provide a framework to enable the use of formal relations to constitute and validate (through appropriate experiment) objective scientific knowledge.

Since acquiring a comprehensive understanding of genomic and proteomic networks requires analyzing data from different levels, it is important to develop models to describe system structures and to predict evolving network behavior from multiple data sources. These models would provide structural and functional representations of the biology knowledge and facilitate the integration of multiscale and multimodal data sources.

We also consider that it is essential to analyze algorithms in a systems context. For instance, many data normalization procedures are being proposed and evaluated without reference to the systems behavior where the data are to serve as input. Complex systems are inherently multimodal, and the real issue is how individual procedures function relative to the goals of the whole system, not in how they function in isolation.

The life sciences rely on predictive models. Serious thought needs to be devoted to their epistemology. Models should include existing biological information, suitable mathematical formulas, and necessary data. Predictive mathematical models are necessary to move biology in the direction of a predictive science. They are also necessary to the application of engineering methods to translate biological knowledge into therapies with a mathematical and computational basis.

The need for new formalisms for representing knowledge requires the development of machine learning techniques that can automatically discover and build systems from data, taking advantage of existing information, both structural and parametric. Analyzing and synthesizing these models requires that they be of minimal complexity for the goals at hand.

3.2: Interdisciplinary Training

It is crucial to educate scientists, mathematicians, and engineers for the challenges and opportunities posed by life sciences. In the long run, biologists may need to be educated in mathematics just as much as physicists, but in different types of mathematics. In the intermediate term, statisticians, applied mathematicians, and engineers can receive retraining in the fundamental principles of biology and specific domain knowledge where necessary. This can be accomplished by developing programs that retrain engineers in active multidisciplinary environments and provide directed educational courses, facilitating ongoing collaboration between engineers and biologists, and developing interactive cross-disciplinary faculty workshops so that researchers with different backgrounds can exchange ideas.

Incidentally, a number of attendees have expressed the need to change the current academic reward system which encourages individual academic excellence and performance into a system that rewards sharing, collaboration, teamwork, training and education.

3.3: Develop Open Source, Multiscale Multimodality Informatics Toolkits

Open-source based tools have the potential to facilitate the development of reusable software and the contribution of universally understood modules to software libraries. The software tools should be able to run on portable platforms independent of operating systems. The availability of these tools would allow research laboratories to concentrate on solving hard, biological problems instead of diverting energy and time into building proprietary tools. The recent NIH sponsored program in National Centers for Biomedical Computing is considered a significant step forward in this direction.

Ideally, such toolkits should include:

- Modeling formalism connecting measurements at different spatio-temporal scales from different modalities
- knowledge representation applicable to healthcare and life science taxonomy, ontology, etc.
- knowledge extraction, data mining, etc.
- formulism and methodology for data, information, knowledge and decision integration
- algorithms for optimal image analysis of different modalities of biological and medical images.

The toolkits should contain complete

- Phenome characterization of disease associated cell systems (e.g., cancer, cardiovascular, and dementia), for scientific hypothesis development;

- searchable cytological profiling of all molecular/drug compound libraries to speed up the drug discovery process, especially for orphan drugs.

In addition, mechanisms for evaluating and validating these open source toolkits are needed. These toolkits, once available, will dramatically accelerate the development of multimodal biomedical systems capable of addressing the grand challenges listed earlier.

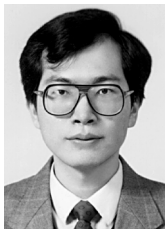
Acknowledgements

This publication is based upon independent research/development (IRD) work supported by the National Science Foundation (NSF) while Dr. S.S. Demir served at NSF.

Disclaimer

Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author and do not reflect the views of the National Science Foundation.

Charles P. Friedman is currently a Senior Scholar and Program Officer in the Extramural Programs Division of the National Library of Medicine (NLM). He is on an extended leave from the University of Pittsburgh. He is overseeing NLM's extramurally-supported informatics training activities, serving as NLM's representative to the program of National Centers for Biomedical Computing, and administering NLM's research grant program in bioinformatics. He is also teaching and consulting on a range of research and evaluation projects. At Pittsburgh, Friedman is appointed as Professor of Medicine and held the titles of Associate Vice Chancellor for Biomedical Informatics and Director of the Center for Biomedical Informatics.



Chung-Sheng Li [S'87,M'92,SM'94,F'03] received the BSEE degree from National Taiwan University, Taiwan, R.O.C., in 1984, and the M.S. and Ph.D. degrees in electrical engineering and computer science from the University of California, Berkeley, in 1989 and 1991, respectively. He has been with the Computer Science Division at the IBM T.J. Watson Research Center as a research staff member since Sept. 1991, and has been the Associate Director of Computer Science since Oct. 2004.

His research interests include broadband and wireless applications, which include digital library, information and media marketplace, content-based retrieval of images and image sequence, knowledge discovery and data mining, content adaptation, and pervasive commerce; broadband network and switching, which includes all-optical networks, storage area networks, and fiber channel; and broadband technologies, which include optical chip interconnects, opto-electronics, and high-speed analog/digital VLSI circuit design. He has initiated and coinitiated several research programs in IBM on fast tunable receiver for all-optical networks, content-based retrieval in the compressed domain for large image/video databases, federated digital libraries, and biosurveillance.

He received an Outstanding Innovation Award from IBM in 2000 for his leadership and major contribution to the IBM/ NASA digital library project, and a Research Division award from IBM in 1995 for his major contribution to the tunable receiver design for WDMA, and numerous invention and patent application awards. He is currently an associate editor for the IEEE Transaction on Multimedia and the Journal of Computer Vision and Image Understanding, and the technical editor for the IEEE Communication Magazine. He has authored or coauthored more than 120 journal and conference papers and received the best paper award from IEEE Transactions on Multimedia in 2003. He is a Fellow of the IEEE Circuit and System Society, the IEEE Laser Electro-Optic Society, the IEEE Communication Society, and the IEEE Computer Society.

Edward R. Dougherty is a professor in the Department of Electrical Engineering at Texas A&M University in College Station. He holds a Ph.D. in mathematics from Rutgers University and an M.S. in Computer Science from Stevens Institute of Technology. He is author of twelve books, editor of four others, and author of more than one hundred and sixty journal papers. He is an SPIE fellow, is a recipient of the SPIE President's Award, and has served as editor of the Journal of Electronic Imaging for six years. Prof. Dougherty has contributed extensively to the statistical design of nonlinear operators for image processing and the consequent application of pattern recognition theory to nonlinear image processing. His current research is focused in genomic signal processing, with the central goal being to model genomic regulatory mechanisms. He is head of the Genomic Signal Processing Laboratory at Texas A&M University and adjunct professor in the Department of Pathology of the University of Texas M.D. Anderson Cancer Center.

Jie Chen received his Ph.D. degree in Electrical and Computer Engineering from the University of Maryland, College Park. He is currently an Assistant Professor at Brown University's Division of Engineering. Dr. Chen's research interests include nanoscale devices and architecture design, genomic signal processing, and multimedia

communications. He is a Distinguished Lecturer of IEEE Circuits and Systems Society (2004–2005). He has been invited as speakers in different conferences and workshops, and as the guest editors of two special issues, "Multimedia over IP" for IEEE Transaction on Multimedia, and "Multimedia over wireless networks" for EURASIP Journal on Applied Signal Processing. Dr. Chen has published 46 scientific papers in refereed journals, conference proceedings, and the book, "Design of Digital Video Coding Systems: A Complete Compressed Domain Approach" (New York: Marcel Dekker 2001); and co-edited another book "Genomic Signal Processing and Statistics" (EURASIP Book Series, 2004). He has invented or co-invented several U.S. patents. He is an associate editor for IEEE Signal Processing Magazine, and has been associated editors for IEEE Trans. on Multimedia and EURASIP Journal on Applied Signal Processing. He is a senior member of IEEE signal processing society.

Semahat S. Demir received B.S. in Electronics Engineering, (Istanbul Technical University, 1988), M.S. in Biomedical Engineering (Bosphorous University, 1996), and a second M.S. and a Ph.D. both in Electrical and Computer Engineering (Rice University, 1992 and 1995). She did postdoctoral training in Biomedical Engineering (Johns Hopkins University, 1995–96). In industry, Semahat worked as a technical manager and medical laser engineer for both Messerschmidt Bolkow Blohm and Rodenstock (Germany and Turkey, 1988–89) and as a research and development engineer for Siemens Company (Germany and Turkey, 1988). She was an assistant professor (1996–200) and has been an associate professor of Biomedical Engineering at the Joint Program of University of Memphis and University of Tennessee Health Science Center since 2000. Her research is in computational modeling of bioelectricity in cardiac electrophysiology and neuroscience and in the development of simulation resources. She served as an Expert Scientist and a Consultant (for the Bioinformatics Research Initiative funded by the World Bank) for The Scientific and Technical Research Council of Turkey under the Prime Ministry of Turkey (1999–2000). Dr. Demir has been Program Director, Biomedical Engineering and Research to Aid Persons with Disabilities at National Science Foundation since June 1, 2004.

Recently Dr. Demir received **Special Recognition Award** at the Women of Color in Health, Science and Technology Award Conference (2002), and **Distinguished New Engineer** Award at the Society of Women Engineers National Conference (2003). Dr. Semahat Demir was selected as 2004 Featured Engineer by Memphis Institute of Electrical Electronics Engineers (IEEE) and an **Achievement Awardee** by Memphis Joint Engineers Council (2004).

Dr. Demir is a member of Eta Kappa Nu Electrical Engineering Honor Society, Sigma Xi Scientific Research Society, American Society for Engineering Education (ASEE), Biomedical Engineering Society (BMES), Biophysical Society, Institute of Electrical and Electronics Engineers (IEEE), Engineering in Medicine and Biology Society (EMBS), Society of Women Engineers (SWE), Women in Engineering Programs and Advocates Network (WEPAN). Dr.Demir's numerous professional society leadership activities include: Members-at-Large Representative of IEEE EMBS AdCom (governing board) 2001–2003; Chair, Exhibits/Job Committees, Joint Meetings of IEEE/EMBS and BMES, 1999 & 2001; Chair and Vice-chair, Memphis IEEE/EMBS Chapter; Founder-Advisor, EMBS Student Club (Brainwave); Deputy Conference Chair, 22nd Conference of IEEE/EMBS, 2000; Job Fair Coordinator, World Congress 2000; and Chair, IEEE/EMBS Industrial Relations Committee, 2000–2001; Chair, IEEE/EMBS Distinguished Lecturers Committee; Associate Editor for Industry Affairs of EMB Magazine since 2001; Vice-chair of Professional Development, Biomedical Engineering Division of ASEE (2003–2005); Founder-Advisor, Female Engineers Mentor for Success (FEMS), Peer-mentoring group for female biomedical engineering students (University of Memphis and University of Tennessee, 2003–present); and Chair of Women in Academia Committee of SWE 1999–2004. She has serves as theme, track and session chair of a number of conferences at EMBS, BMES, SWE, ASEE and other societies. She is a senior member of IEEE.

Stephen Wong, Ph.D. (CS), PE (EE) is the founding Director of the Center for Bioinformatics, Harvard Center of Neurodegeneration and Repair (HCNR) and an Associate Professor of Radiology, Harvard Medical School. Stephen is a hybrid scientist who has successfully straddled the fields of computational science and biomedical research. He has 20 years of R&D experience building large scale systems for leading institutions in academia and industry. Stephen received his executive education from MIT Sloan School, Stanford University and Columbia University. His current research focuses on non-invasive techniques for biomarker development for disease diagnosis and therapy, high content analysis of life science imaging, and computational neuroscience. Stephen is the founding co-chair and current chair of IEEE CAS Life Science Systems and Applications Technical Committee.