# Chapter 3
# BIOINFORMATICS CHALLENGES AND OPPORTUNITIES

Peter Tarczy-Hornoch[1] and Mark Minie[2]

*[1]Department of Medical Education and Biomedical Informatics; [2]Health Sciences Libraries, University of Washington, Seattle, WA 98195*

## Chapter Overview

As biomedical research and healthcare continue to progress in the genomic/post genomic era a number of important challenges and opportunities exist in the broad area of biomedical informatics. In the context of this chapter we define bioinformatics as the field that focuses on information, data, and knowledge in the context of biological and biomedical research. The key challenges to bioinformatics essentially all relate to the current flood of raw data, aggregate information, and evolving knowledge arising from the study of the genome and its manifestation. In this chapter we first briefly review the source of this data. We then provide some informatics frameworks for organizing and thinking about challenges and opportunities in bioinformatics. We use then use one informatics framework to illustrate specific challenges from the informatics perspective. As a contrast we provide also an alternate perspective of the challenges and opportunities from the biological point of view. Both perspectives are then illustrated with case studies related to identifying and addressing challenges for bioinformatics in the real world.

## Keywords

bioinformatics; computational biology; biomedical informatics; Human Genome Project; applied informatics; foundations of informatics; information access; sociotechnical dimensions; evaluation

# 1. INTRODUCTION

As biomedical research and healthcare continue to progress in the genomic/post genomic era, a number of important challenges and opportunities exist in the broad area of biomedical informatics. Biomedical informatics can be defined "as the scientific field that deals with biomedical information, data, and knowledge – their storage, retrieval, and optimal use for problem-solving and decision making" (Shortliffe et al.., 2001). To understand the challenges and opportunities for informatics within the field of bioinformatics (defined most broadly as informatics in the domains of biology and biomedical research) it helps to understand the broader context in which they exist.

In the broader context, the key challenges to bioinformatics essentially all relate to the current flood of raw data, aggregate information, and evolving knowledge arising from the study of the genome and its manifestation. The genome can be thought of as the machine code or raw instructions for creation and operation of biological organisms (its manifestation). The information encoded in DNA results in the creation of proteins which serve as the key building blocks for biological function (a protein on the surface of one cell (neuron) in the brain can recognize a chemical signal sent by a neighboring neuron). Proteins physically aggregate to create more complex units of biological function termed protein complexes (the protein that recognizes the signal from a neuron might be part of a protein complex that translates that signal into an action such as turning on another protein that was in "standby mode"). Proteins and protein complexes interact with one another in networks or pathways to carry out higher level biological processes (such as the neuronal signaling pathway). These pathways include regulatory mechanisms whereby the function of the pathway overall is controlled by relevant input parameters (such as frequency and intensity of input from the part of the nervous system related to sensing pain). This regulation is complex and can include feedback and interaction among the proteins and protein complexes of the pathway, as well as regulation and interaction of other pathways. Interestingly, mechanisms include also the regulation of the conversion (translation) of the raw information encoded in the DNA into the intermediate messages (mRNA) and regulation of the conversion of the mRNA into proteins, as well as modification of the proteins themselves. The pathways in turn are assembled into more complex systems of multiple interacting pathways (pathways involved in evasive response to painful stimuli). In multi-cellular animals these complex systems in turn interact to control the function of their basic building blocks, namely the cells (for example, a brain cell or neuron). The cells in turn interact with one another and form higher order structures termed organs (the brain, for example). These organs interact with one another to form systems (such as the nervous system, which includes the brain as well as the input from sensory organs and the output to muscles and other organs). These systems interact to carry out higher order functions such as seeking out food sources (thus for example the nervous system guides the organism to seek food, the digestive system breaks down food, the metabolic system helps control the conversion of food to sugars, and the circulatory system helps deliver this energy to cells). Expanding beyond this level one can think of organisms interacting to form ecosystems in turn resulting in the Earth's biosphere. This hierarchical progression is illustrated in Figure 3-1. This cursory overview of the modern view of biological systems begins to shed light on the challenges faced by the fields of modern biology and biomedical research and the roles that bioinformatics might play.

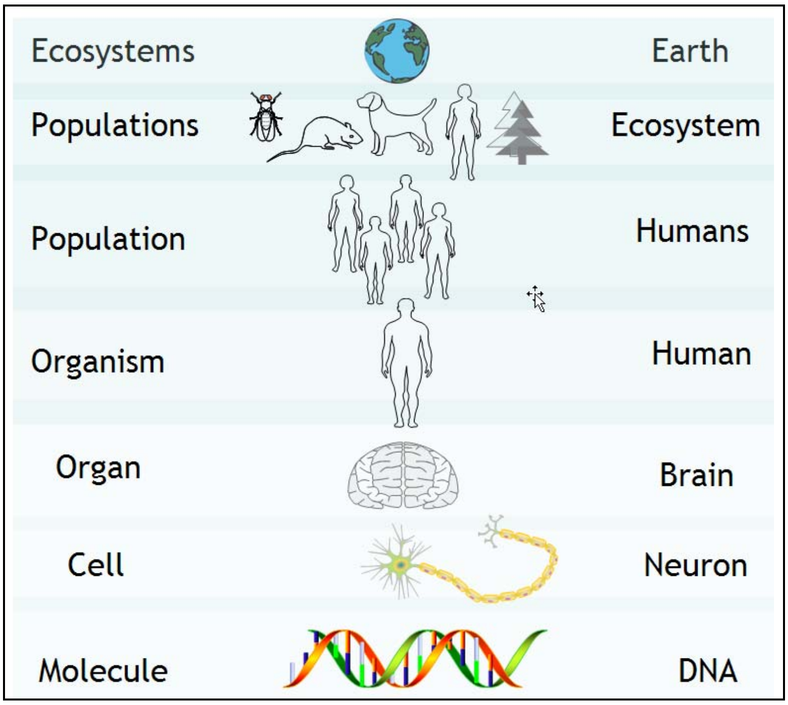


*Figure 3-1.* Hierarchy of biological systems.

In the broader context, to understand the opportunities for both biomedical research and bioinformatics, it helps to understand the genesis of this flood of information and more importantly the vision of how this information might be used. The roots of both the large quantity of information and the guiding vision can be traced to the start of the modern era of biomedical research, which is felt to be the discovery by Watson and Crick in 1953 of DNA as the information storage mechanism for cells. Research into the genome continued at a relatively linear pace until the establishment in 1989 of the National Center for Human Genome Research (NCHGR) to carry out the role of the National Institutes of Health (NIH) in the International Human Genome Project (HGP: see Online Resources). The HGP served to accelerate the pace of data generation from a linear to an exponential growth pattern as shown in Figure 3-2.

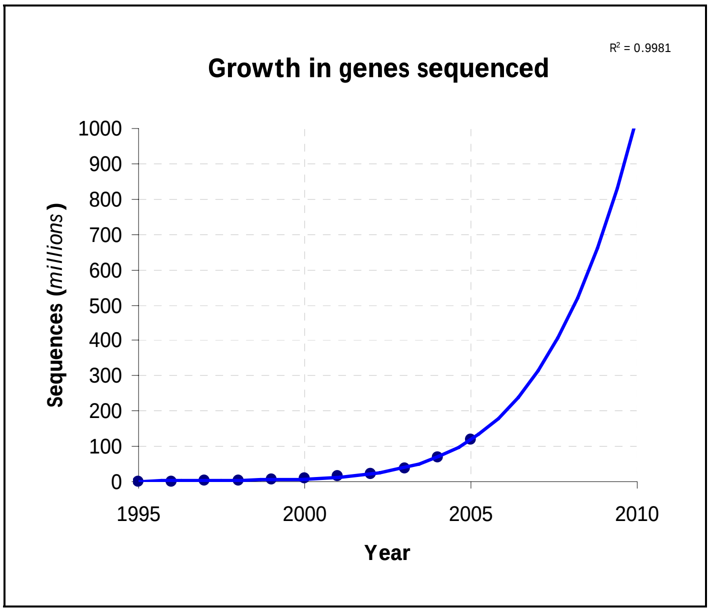


*Figure 3-2.* Growth in genes sequenced.

The seed of the vision for the HGP and the investment that has been made can be found in the mission of the National Institutes of Health (NIH) which is “science in pursuit of fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to extend healthy life and reduce the burdens of illness and disability.” The relationship of this mission to the grand vision of the HGP was published in 1990 as part of the first five year plan for the HGP: “The information generated by the human genome project is expected to be the source book for biomedical science in the 21st century and will be of immense benefit to the field of medicine. It will help us to understand and eventually treat many of the more than 4000 genetic diseases that afflict mankind, as well as the many multifactorial diseases in which genetic predisposition plays an important role." (See Online Resources). The flood of data, information, and knowledge we face today in biology and biomedical research can be traced directly to the coordinated international investment of large amounts of funding to sequence the human genome as a first step in arriving at a deeper understanding of the basis of human health and disease (Collins and McKusick, 2001). Research into the genomics and basic biology of diverse other organisms was galvanized by this effort as well and has been proceeding in parallel over the last decade and a half. With the completion of the sequencing of the DNA of humans and other organisms we have however only begun to explore the hierarchy discussed above and shown in Figure 3-1.

A guiding vision for the next phases of the HGP was articulated in a paper published in *Nature* on the 50th anniversary of Watson and Crick's discovery (Collins et al., 2003). This paper outlines fifteen grand challenges clustered into three broad areas: Genomics to Biology (improving our understanding of complex biological systems), Genomics to Health (developing and applying our understanding of the genomic basis for health and disease), and the sometimes underappreciated Genomics to Society (broadly, the ethical, legal, and social implications of our understanding).

These challenges, of course, present opportunities as well. As an example of a grand challenge presenting opportunities for biologists and informatics researchers in the Genomics to Biology area, consider, "Grand Challenge I-2: Elucidate the organization of genetic networks and protein pathways and establish how they contribute to cellular and organismal phenotypes." An example from the Genomics to Health area is, "Grand Challenge II-3: Develop genome-based approaches to prediction of disease susceptibility and drug response, early detection of illness, and molecular taxonomy of disease states." In response to the challenges posed by a post-genome sequencing era of biomedical research the NIH has identified the intersection of the computing and biological and biomedical fields as a key opportunity for future research based on the challenges and potentials outlined above. A critical articulation of this was provided by the report that led to the creation of the National Institutes of Health Biomedical Information Science and Technology Initiative (BISTI). (See Online Resources for the URL.) This introduction provides a high level overview of the opportunities and challenges for the field of bioinformatics. In the following sections we outline from an informatics perspective some more specific challenges and illustrate this with case studies/examples.

## 2. OVERVIEW OF THE FIELD

### 2.1 Definition of Bioinformatics

The definition of bioinformatics used in this chapter is the broadest possible definition of the field, namely *all informatics research and application in support of the biological research endeavor*. In the context of the definition of biomedical informatics given in the introduction "as the scientific field that deals with biomedical information, data, and knowledge – their storage, retrieval, and optimal use for problem-solving and decision making" (Shortliffe et al., 2001), we define bioinformatics as the subset of the field that focuses on information, data, and knowledge in the context of biological and biomedical research. By our definition the culture and environment (context) in which bioinformatics is studied and applied are that of the researcher in the laboratory seeking new knowledge. This includes a broad range of research ranging from a) basic molecular and cellular level research seeking to understand the way cancer results in unregulated growth of cells to, b) whole animal applied research looking at ways to block the spread of cancers, to c) clinical research involving patients looking at genetic factors influencing susceptibility to cancer. It is distinct from clinical informatics which focuses on the culture and environment of clinical care involving patients and healthcare providers in settings ranging from one's own home, to outpatient (clinic) and inpatient (hospital) care. This definition is similar to the one used by the BISTI website: "Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data" (see Online Resources).

There are a number of other definitions of the term "bioinformatics" and in reading the literature it is important to be sure one is clear on the meaning being used. For some, the term is fairly narrow and refers primarily to developing and validating and applying algorithms for processing and analyzing sequences of DNA (the phrase "computational biology" is also being used for this area). Others expand the definition of bioinformatics to include any algorithmic or statistical approach to the analysis of biological data. Some make a distinction between mathematical modeling in biology and bioinformatics, whereas others view the former a subset of the later. For some, bioinformatics refers to the basic research in the area, whereas the applied side of deploying systems is termed biocomputational infrastructure. For others, bioinformatics refers to the set of computational tools used by biologists to carry out their research. A very interesting alternate broad definition is, "The study of how information is represented and transmitted in biological systems, starting at the molecular level" (Bergeron, 2002). For the remainder of this chapter, we will use this last broader, more inclusive definition of the term.

## 2.2 Opportunities and Challenges – Informatics Perspective

### 2.2.1 Frameworks for Describing Informatics Research

The field of biomedical informatics is relatively young and there are a number of ways to organize important research questions and areas (and in turn to discuss challenges and opportunities).

The American Medical Informatics Association developed the following framework, shown in Tables 3-1 and 3-2, categorizing research papers in the discipline submitted for review at the 2003 annual meeting (Scientific Program Committee Chair: Mark A. Musen, Foundations Track Chair: Charles P. Friedman, Applications Track Chair: Jonathan M. Teich, see http://www.amia.org/meetings/archive/f03/call.html#categorizing).

The Foundations Track, shown in Table 3-1, focuses on theories, models, and methods relevant to biomedical informatics broadly (applicable to clinical informatics, bioinformatics, and public health informatics). Bold faced categories are foundational approaches often referred to in publications in the bioinformatics arena. Each of these represents ongoing areas of inquiry and thus potential challenges and opportunities for bioinformatics, both in terms of research and in terms of application. As will be discussed later in this chapter, some foundational areas are not currently active areas of research in bioinformatics and may represent important opportunities for future research (in particular many of the areas in C).

*Table 3-1.* Categories of Informatics Research*

| I. Foundations of Informatics Building Models and Methods for Biomedical Information Systems |
|---|
| A. Modeling Data, Ontologies, and Knowledge |
| **1. Controlled terminologies and vocabularies, ontologies, and knowledge bases** |
| **2. Data models and knowledge representations** |
| **3. Knowledge acquisition and knowledge management** |
| B. Methods for Information and Knowledge Processing |
| **1. Information retrieval** |
| **2. Natural-language processing, information extraction, and text generation** |
| **3. Methods of simulation of complex systems** |
| 4. Computational organization theory and computational economics |
| 5. Uncertain reasoning and decision theory |
| **6. Statistical data analysis** |
| **7. Automated learning, discovery, and data mining methods** |

continued

| I. Foundations of Informatics Building Models and Methods for Biomedical Information Systems |
|---|
| B. Methods for Information and Knowledge Processing *(continued)* |
| 8. Software agents, distributed systems |
| 9. Cryptography, database security, and anonymization |
| **10. Image representation, processing, and analysis** |
| **11. Advanced algorithms, languages, and computational methods** |
| C. Human Information Processing and Organizational Behavior |
| 1. Cognitive models of reasoning and problem solving |
| **2. Visualization of data and knowledge** |
| 3. Models for social and organizational behavior and change |
| 4. Legal issues, policy issues, history, ethics |

**Used with permission from the American Medical Informatics Association*

The Applications Track, shown in Table 3-2, focuses on real world systems: their design, implementation, deployment, and evaluation. Bold faced categories are applications often referred to in publications in the bioinformatics arena. Each category represents ongoing areas of inquiry and thus potential challenges and opportunities for bioinformatics, both in terms of research and in terms of application. As will be discussed later in this chapter, some application areas, similar to the theoretical track, are not currently active and may represent important opportunities for future research: for example, B, or the intersection of bioinformatics with C1.

*Table 3-2.* Categories of Informatics Research*

| II. Applied Informatics - Real World Solutions for Real World Problems |
|---|
| A. Advanced Technology and Application Infrastructure |
| 1. Data standards and enterprise data exchange |
| 2. System security and assurance of privacy |
| 3. Human factors, usability, and human-computer interaction |
| 4. Wireless applications and handheld devices |
| **5. High-performance and large-scale computing** |
| 6. Applications of new devices and emerging hardware technologies |
| B. Evaluation, Outcomes, and Management Issues |
| 1. Organizational issues and enterprise integration |
| 2. System implementation and management issues |
| 3. Health services research: health care outcomes and quality |
| C. Information, Systems and Knowledge Resources for Defined Application Areas |
| 1. Care of the patient |
| a. Electronic medical records |
| b. Computer-based order entry |
| c. Clinical decision support, reference information, decision rules, and guidelines |
| d. Workflow and process improvement systems |
| e. Nursing care systems |

*continued*

II. Applied Informatics - Real World Solutions for Real World Problems

C. Information, Systems and Knowledge Resources for Defined Application Areas
1. Care of the patient *(continued)*
f. Ambulatory care and emergency medicine
g. Telemedicine and clinical communication
h. Patient self-care, and patient-provider interaction
i. Disease management
2. Care of populations
a. Disease surveillance
b. Regional databases and registries
c. Bioterrorism surveillance and emergency response
d. Data warehouses and enterprise databases
3. Enhancements for education and science
a. Consumer health information
b. Education, research, and administrative support systems
c. Library applications
**4. Bioinformatics and Computational Biology**
**a. Genomics**
**b. Proteomics**
**c. Studies linking the genotype and phenotype**
**d. Determination of biomolecular structure**
**e. Biological structure and morphology**
**f. Neuroinformatics**
**g. Simulation of biological systems**

**Used with permission from the American Medical Informatics Association*

The University of Washington Biomedical and Health Informatics Graduate Program has taken a less granular approach to categorizing the broad field of biomedical informatics with three application domains and four foundational areas. The three application domains are: a) Biomedical Research, b) Clinical Care, and c) Public Health. The four foundational areas are: a) Biomedical Data and Knowledge, b) Biomedical Information Access and Retrieval, c) Biomedical Decision Making, and d) Socio-Technical Dimensions of Biomedical Systems. In addition to the application domains and the foundational areas the University of Washington requires grounding in methodologies including programming, statistics, research design and evaluation. The need for evaluation methodologies is especially important as is discussed below. The next sections will use these foundational areas to illustrate challenges and opportunities in the bioinformatics domain.

### 2.2.2 Opportunities and Challenges – Biomedical Data and Knowledge

The volume and diversity of biomedical data is growing rapidly, presenting a number of challenges and opportunities ranging from data capture, data management, data analysis, and data mining. The analysis of this data is generating new knowledge that needs to be captured. As the volume of this knowledge grows, so does the need to develop formal ways of representing this knowledge. Knowledge bases and formal approaches including ontologies are potential solutions. This particular area of biomedical data and knowledge will be explored in more depth than the other areas given the emphasis of this book.

Analysis of gene expression (microarray) experiments illustrates diverse aspects of the problem with modern biological data. In a gene expression experiment the biologist measures the level of expression of all genes in a particular tissue under a given condition, and then frequently compares expression levels to those in the same tissue under a different condition (a process known as differential gene expression). Thus, for example, one might measure the level of gene expression (the degree to which certain genes are turned on or off) by comparing cancer cells that have received a cancer drug to ones that have not.

The first challenge is management of the experimental data since a single gene expression measurement results in thousands of data points. In turn typically one repeats each experimental condition and control condition multiple times. Frequently, the measurements are repeated at multiple time points (for example, before treatment with a drug, one hour after, four hours after, eight hours after, twenty-four hours after). A number of open source and commercial packages help researchers collect and manage gene expression data.

The next challenge is data analysis and data mining. There are a number of commercial expression array analysis packages but they often do not implement the latest algorithms and methods for data analysis. Important open source collaborations aim to develop tools to assist researchers in developing and using new tools for array analysis. This collaboration is the BioConductor project (http://www.bioconductor.org) and is built on top of the R programming environment (Ihaka and Gentleman, 1996).

Finally, there is the need to mine large data sets of gene expression data. A number of studies have been published using a variety of data mining techniques from computer science and this is still a rapidly evolving area. An example of this class of problems is trying to predict the outcome of cancer patients based on analyses of the gene expression in their cancerous tissue (e.g. gene expression in a piece of breast cancer removed by the surgeon). A classic study used DNA microanalysis and a supervised classifier to predict outcome of breast cancer far better than any other classifiers (van 't Veer et al., 2002).

The data capture and data management problem is compounded by the fact that modern biological experiments frequently involve diverse types of data ranging from analysis of mutations (changes in the DNA sequence) to gene expression to protein expression to biochemical measurements to measurements of other properties of organisms (frequently termed phenotype). In order to make sense out of these diverse experimental results and to incorporate data, information, and knowledge from public domain databases (such as databases of protein function) data integration is needed. A number of data integration systems for biomedical data have been developed. These data integration approaches are reviewed in a number of articles (Sujansky, 2001). The BioMediator system (formerly GeneSeek) (Donelson et al., 2004; Mork et al., 2001; Mork et al., 2002; Shaker et al., 2002*;* and Shaker et al., 2004) is one such system for data integration. It is designed to allow biologists to develop their own views of the way in which diverse private (experimental data) and public databases and knowledge bases relate to one another and to map this view (the mediated schema) onto the specific sources they are interested in querying. The interfaces, or wrappers, to these diverse sources are written in a general purpose fashion to permit the same wrappers to be reused by diverse biologists. The custom views (mediated schemata) are captured in a frames based knowledge base (implemented in Protégé) (Stanford, 2002). The system architecture permits in a single environment both the integration of data from diverse sources and the analysis of this data (Mei et al., 2003). The system works well but an important set of challenges surrounds the need to develop tools that permit the biologists to manipulate the mediated schema in a more intuitive fashion. Another challenge is to incorporate such systems into the workflow of the typical biological lab.

Ultimately all this data generates new knowledge which needs to be captured and shared. The volume of this knowledge is growing only linearly as shown in Figure 3-3 in contrast to the growth of the data.

An important challenge to knowledge creation is developing ways to increase the rate of knowledge generation to keep up with the rapid growth of data. Even with the linear growth of knowledge the volume of it is such that it is becoming difficult for one person to keep up with it all systematically. In order to access and use this knowledge it is becoming more and more important that the knowledge be captured in computable form using formalisms from the computer science community such as ontologies. These topics are discussed in more detail in other chapters.

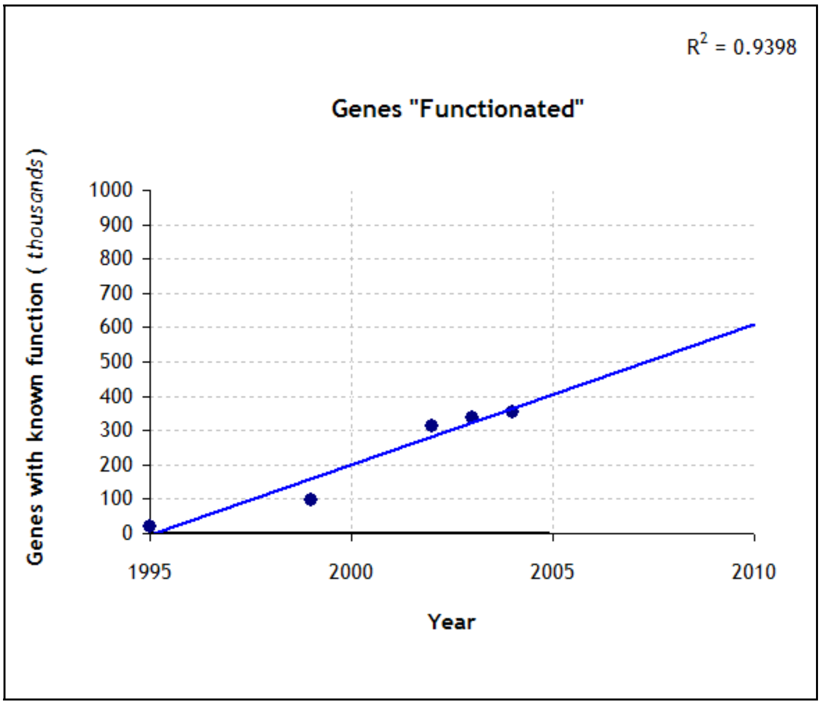


*Figure 3-3.* Growth rate of Genes with known function

The power and the challenges of these approaches can be illustrated by three important bioinformatics related knowledge bases. The first is the Foundational Model of Anatomy (FMA) (Rosse and Mejino, 2003) which is a centrally curated knowledge base capturing anatomic knowledge about the body from levels of granularity ranging from the whole body down to cells, sub cellular compartments and molecules (proteins). The FMA is becoming widely adopted as a reference standard for describing a variety of biologic processes in terms of where they occur and what they impact (serving as the anatomic component of the Unified Medical Language System (Tuttle, 1994; Bodenreider et al., 2002, among others). Some important challenges remain, though, in that: a) the FMA describes only human anatomy yet much work is being done on other species; b) the centralized curation process ensures internal consistency and quality control yet does not scale well to match the expansion of the FMA; c) the FMA describes normal physical structures but needs to be extended to describe abnormal (or disease related) structures; and d) the FMA needs to be extended to describe processes and functions of the physical structures.

The Gene Ontology (GO) Consortium (Gene Ontology Consortium, 2000) takes a different approach to describing the current state of knowledge about proteins and their functions. Given the evolving nature of the field a centralized top down approach such as that taken by the FMA was not possible. The GO is thus created and curated in a distributed fashion by a consortium of experts in molecular biology. The strength of this approach is that it scales well and adapts well to the rapidly changing state of our knowledge. Challenges to the GO approach include a) difficulty in maintaining internal consistency of the knowledge base; b) capturing in computable form from biologists subtle aspects of function; and c) maintaining referential integrity as the knowledge base evolves.

The third example of a bioinformatics knowledge base is the PharmGKB project (Klein et al., 2001, and PharmGKB, n.d.*)* which is a sophisticated pharmacogenomics knowledge base. The strength of this knowledge base is that it was centrally designed with distributed input to capture in a computable form a large amount of knowledge relevant to the field of pharmacogenomics - the interaction between an individual’s genes, the medicines taken and the variability in response to these medicines. The challenges with this approach, however, are: a) it is dependent up on human curation (this is a shared challenge with FMA and GO as well); and b) extending the knowledge base to other areas of biology will be a challenge since unlike GO and FMA, the scope of PharmGKB was designed to be deep and narrow (pharmacogenomics) rather than broad and comprehensive (anatomy or molecular function).

### 2.2.3 Opportunities and Challenges – Biomedical Information Access and Retrieval

As the volume of data and knowledge grows it is becoming critical to biologists that they be able to access and retrieve the relevant pieces when they need it. The older paradigm of keeping up with the contents of the handful of top journals relevant to one’s biological research area no longer works.

There are three key factors contributing to this. The first factor is that the sheer volume of new information is such that systematically keeping up is no longer a viable option. The second and related factor is that with the growth in new information has come a growth in the number of places in which information is published. Related to the dispersion of information across diverse sources is the fact that interdisciplinary and interprofessional research is becoming the norm, thus important research findings are published in a wider range of journals. The third factor is that information is becoming more and more available in electronic form and no longer just in condensed form in journals, resulting in a proliferation of biological databases, knowledge bases, and tools.

The University of Washington BioResearcher Toolkit (see Online Resources) illustrates the opportunity and the challenge this presents for biologists and for bioinformatics researchers and developers. Simply trying to find the right resource for a particular task from among hundreds is a challenge to say nothing of finding the right information within that resource. Given the volume of data and the fact that it exists as a combination of data in databases and free text, an important part of information access and retrieval has been both data integration and data mining, as discussed above. Intelligent parsing of queries, frequently involving natural language processing of both queries and sources, is becoming a key component of information access and retrieval. The challenges, opportunities, and state of the art of information retrieval (and data mining) in bioinformatics is covered in more depth in other chapters.

### 2.2.4 Opportunities and Challenges – Biomedical Decision Making

Thus far the field of bioinformatics has done little explicit research into the area of decision making. Within clinical or medical informatics there is a rich history of research into systems designed to help care providers and patients (healthcare consumers) make optimal decisions surrounding diagnosis (what disease or illness is it that a patient has) and management (which of the options for treatment are best factoring in details of the circumstances and the values of the patient). Approaches and methods used have included Bayesian belief networks, decision analytic models, and rule-based expert systems, among others. An important area on the clinical side for decision support systems has been genetic testing which has obvious ties (though one step removed) to bioinformatics research. Though this area of decision making is outside the primary scope of this book, it is worth noting that there appears to be a great potential opportunity to explore the development of tools for biologists to explicitly assist them in their decision making processes. The challenge is the paucity of literature and study in this potential arena. The first steps likely would be needs assessments and development and validation of models of decision making for biologists to see if in fact there is a niche for decision making tools in biomedical research.

### 2.2.5 Opportunities and Challenges – Evaluation and Socio Technical Dimensions of Biomedical Systems

The bioinformatics literature has a large number of papers published on theoretical frameworks for bioinformatics systems and a large number of papers on specific bioinformatics applications. There is, however, a relative lack of formal evaluations of bioinformatics systems and models. There is also a relatively sparse literature that formally and systematically examines the needs of biologists for specific tools (for example, Yarfitz and Ketchell, 2000). In part this is due to the relatively young nature of the field. A related factor is that to date much tool development has been driven by experienced biologists solving recurring problems they face through computational tools and sharing these tools with others. Though evaluation per se is outside the scope of this book it is important to learn from the experience of the clinical (medical) informatics community. Careful assessment and evaluation of the needs of users of the system is an important factor in guiding future development both on the theoretical (foundational) front as well as the applied front. Equally importantly formal evaluations and comparisons of alternate solutions (both applied and theoretical) are needed in order to guide development as well. An excellent resource on evaluation of systems in the clinical (medical) informatics arena is *Evaluation Methods in Medical Informatics* (Friedman and Wyatt, 1997); to date there is no similar book for bioinformatics evaluation.

The socio-technical environment in which informatics research and application development occur is becoming increasingly important on the clinical (medical) informatics front. It appears likely this will be true on the bioinformatics front as well. There are a number of ways of looking at this contextualization of informatics. The AMIA community has coalesced interests and activities in this area around the "People and Organizational Issues" working group. Their mission as quoted from their website is “a) To apply the knowledge of human behaviors toward the use of information technology within a health care environment; b) To effectively describe the benefits and impacts of information technology before paradigm shifts fully occur; c) To incorporate organizational change management and human concerns into information technology projects; and d) To distinguish between the human and technology issues when system successes or failures occur.” As the field of bioinformatics grows and matures many of these challenges and opportunities will arise and need to be addressed. Already there are anecdotal reports of the purchase and deployment of complex expensive bioinformatics software packages that are unused despite apparent demand - a finding not unlike what has been seen with the development and deployment of unsuccessful clinical information systems.

Another perspective is provided by the description of the core graduate program courses at the University of Washington, "Sociotechnical Issues in Biomedical Informatics"; quoting from the course description: "Essentially all informatics work - whether purely theoretical or purely applied - is conceived, designed, built, tested, and implemented in organizations. Organizations are comprised of individuals and individuals are human beings, complete with philosophies, ideas, biases, hopes and fears. To build effective and valued informatics systems, the informaticist must understand how and why people behave as individuals, in groups, in organizations, and in society, and then build tools and systems that consider these human factors. The premise of this course is that the thoughtful consideration and application of the management sciences offers the opportunity to mitigate these risks." As bioinformatics projects are smaller in scope, these issues have not risen to the forefront, but as larger scale bioinformatics endeavors are undertaken it is almost certain they will.

## 2.3 Opportunities and Challenges – Biological Perspective

The exponential growth in basic biological data and the incorporation of that raw information into highly integrated databases on the Internet, along with the relatively linear but nonetheless rapid changes in our understanding of biological systems present several opportunities and challenges. These challenges faced by biologists and biomedical researchers present a complementary view to the perspective of the bioinformatics researcher. As noted in the section on Socio-technical Dimensions, understanding and addressing the challenges of the biologists in the trenches are critical to successful deployment of bioinformatics applications. We now discuss some of the challenges and opportunities viewed from the biological perspective.

### 2.3.1 Data Storage, Standardization, Interoperability and Retrieval

The huge growth in biological information being acquired at every level of the biological organization, from simple DNA sequences on up to the global ecosystem, has created serious challenges in data storage, retrieval and display. These challenges are being met by new developments in nanotechnology, search algorithms, and virtual/augmented reality tools as well as more conventional approaches.

### 2.3.2 Data Publication and Knowledge Sharing

NIH now requires all data generated by research it funds to be published in easily accessible and sharable electronic format, creating overwhelming challenges for current approaches such as journals and websites. New technologies such as wikis (see http://wiki.org/ and http://en.wikipedia.org/) and bibliomics tools (such as Telemakus: http://www.telemakus.net/ and PubGene: http://www.pubgene.org/) will need to be applied to these challenges in publication. The very meaning of "publication" has already started to evolve, and libraries in particular are becoming directly involved in providing for the distribution and archiving of raw data from scientific experiments (see DSpace: http://www.dspace.org/). Additionally, increased use of "telepresence" tools such as the Access Grid (http://www.accessgrid.org/) and online collaboration/knowledge sharing tools such as AskMe (http://www.askmecorp.com/) provide new and novel infrastructure in support of the basic biology research effort.

### 2.3.3 Analysis/annotation Tool Development and Distribution/access

The intense development of Open Source bioinformatics tools within different departments/groups at Universities and other institutions has created a need to develop the means of making these "home brew" tools available to the general bioresearch community. At present there is no integrated package analogous to Microsoft Office or an electronic medical record for biomedical researchers. The BioResearcher Toolkit (http://healthlinks.washington.edu/bioresearcher) provides a mechanism for the dissemination and sharing of such tools via its "UW HSL Bioinformatics Tools" section. There, tools developed by national biomedical researchers as well as local biomedical researcher (such as the web based protein structure prediction tool developed by Dr. Robert Baker of the UW Biochemistry Department, Robetta (http://robetta.bakerlab.org/), are made available to users. Other networked software tools, such as Vector NTI and PubGene are also available through the BioResearcher Toolkit site.

### 2.3.4 Hardware Development and Availability

Many bioinformatics applications require tremendous computational power. This challenge is being met by the availability of clusters constructed from readily available desktop computers (http://www.bio-itworld.com/news/083004_report5927.html) as well as specially constructed supercomputing devices such as IBM's BlueGene (http://www.research.ibm.com/bluegene/). Furthermore, the evolution of a new class of "BioIT" specialists such as "The BioTeam" (http://www.bioteam.net/) has increased the availability and utility of hardware needed to meet developments in bioinformatics. Though this may not per se be a challenge for bioinformatics researchers, it does present a challenge to biomedical researchers seeking to use powerful tools; thus, it is a challenge for the discipline of bioinformatics.

### 2.3.5 Training and Education

The constantly changing nature of bioinformatics tools and the rapid growth in biological information has created a need for the development of better and more effective training and education programs in bioinformation data retrieval and analysis. The EDUCOLLAB Group at the National Center for Biotechnology Information (NCBI) has developed a series of introductory and advanced training programs for bioinformatics tool use, and the University of Washington Health Sciences Library has developed a 3-Day intensive training program to train students, faculty and staff in the use of NCBI online resources, commercial software and new developments in biology such as RNAi. These training sessions have been successfully given using telepresence tools such as the Access Grid. Additionally, commercial training companies such as OpenHelix (http://www.openhelix.com/) are now developing to meet the challenge and opportunity presented by the need for such training and education. There has also been a growing realization that a new type of profession, that of "bioinformationist", may be necessary to contend with the vast amount of data and analysis requirements resulting from what is essentially the digital imaging of Earth's biosphere (Lyon et al., 2004; and Florance et al., 2002 ].

### 2.3.6 Networking and Communications Tools

The highly dispersed nature of the modern biological research enterprise has from its inception required a very high degree of networking and communications among individual researchers and organizations—the Human Genome Project itself would not have been possible without the use of the Internet to promote and facilitate the distributed approach to sequencing and annotating the human genome. This had led to more extensive use of telecommunications tools such as WebEx and also to the development of so-called "virtual" organizations such as VirtualGenomics.org (http://www.virtualgenomics.org/). NIH Director Elias A. Zerhouni has specifically described the need for the development of research teams spread out over large distances and many disciplines as a critical part of the NIH Roadmap, and the particular challenge provides the opportunity to develop new organizational structures and networking and communications tools. The Cornell University Life Sciences Initiative VIVO website (http://vivo.library.cornell.edu) provides a prototype for such a tool in a University context, while the Community of Science (COS-http://www.cos.com/) is a commercial enterprise tool for promoting collaborative research.

### 2.3.7 Publication/comprehension of Biological Information

Novel means of publication of data—wikis with their potential for rapid and constant peer review, data posting on websites such as the Gene Expression Omnibus (GEO: http://www.ncbi.nlm.nih.gov/geo/), modeling efforts such as the e-cell Project (http://www.e-cell.org/) and virtual disease models such as the Entelos Diabetes virtual patients (http://www.entelos.com/) and computer generated animations (http://www.wehi.edu.au/education/wehi-tv/dna/index.html) to help understand biological systems—are becoming essential to making efficient use of digital biological information for both clinicians and basic biology researchers. Additionally, new paradigms such as Systems Biology are providing new and important intellectual frameworks for comprehending biological information.

### 2.3.8 Physical Infrastructure and Culture

Conferencing facilities at university libraries for virtual meetings, computer laboratories for training, and architectural designs to promote contact among researchers can further promote collaboration and sharing of data, knowledge and expertise. Bio-X (http://biox.stanford.edu/) at Stanford University is an example of one such effort.

### 2.3.9 Research Center Coordination

Many of the resources for biological research are extremely expensive and mechanisms for sharing such resources must be developed. One example of the use of high speed Internet systems to allow the sharing and operation for advanced tools remotely is the Telescience Portal at the University of California, San Diego (https://telescience.ucsd.edu/), which provides for a collaborative environment for telemicroscopy and remote science. As high-speed connectivity and real-time videoconferencing tools become the norm, “Portals” allowing the use of complex and expensive scientific instruments such as high voltage electron microscopes remotely will allow researchers all over the world to perform experiments remotely and to form collaborative research teams driven by research needs rather than location.

### 2.3.10 Public Outreach

As the stem cell research issue and sometimes emotional debates concerning biodefense, genetics, nanotechnology, and robotics (GNR) developments show (Joy, 2000) it is critical to educate the public as to the science behind such fields as bioinformatics. Public understanding of the Human Genome Project, for example, will greatly enhance decision making as to how the results of that project will be used in the delivery of genomics based health care and technologies. High School Education projects such as the Seattle Biomedical Research Institute’s BioQuest (http://www.sbri.org/sci-ed/index.asp) as well as direct connection with public media such as the Sci-Fi Channel (which has recently elected to produce science fiction classics such as the “Andromeda Strain” and Greg Bear’s “Darwin’s Radio and Darwin’s Children”) and other organizations with influence in the public understanding of science and its roles and effects on society are critically important.

## 3. CASE STUDY

### 3.1 Informatics Perspective – The BIOINFOMED Study and Genomic Medicine

The BIOINFOMED study funded by the European Commission (Martin-Sanchez et al., 2004) is an excellent case study at multiple levels. First it is a study focusing on formally developing a list of challenges and opportunities within bioinformatics and thus provides yet another perspective on opportunities and challenges. Secondly, it explicitly identifies these challenges in a particular sociotechnical context providing a first hand example of the issues identified under evaluation and sociotechnical dimensions. Finally, it articulates the fact that in order to achieve the promise of the Human Genome Project it is critical that work be done at the intersection of bioinformatics and clinical (medical) informatics.

The broad context of the BIOINFOMED study is that of the promise of the Human Genome Project as articulated in the beginning of this chapter. The specific focus is captured by the title of the resulting paper, "Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care." The methods used were a prospective study of the relationships and potential synergies between bioinformatics and medical informatics. The starting point for the study was a written survey developed by the lead institute (Institute of Health Carlos III in Spain) that addresses a number of questions related to research directions and the future of both bioinformatics and medical informatics with an emphasis on opportunities to exchange knowledge across the two subdisciplines. A group of thirty professionals with expertise in medical informatics, bioinformatics, genomics, public health, clinical medicine and bioengineering met twice to analyze and synthesize the results of the survey.

The sociotechnical perspective was the articulation of the various stakeholders' interests and the resultant opportunities and challenges. For the focus of their paper (informatics in support of genomic medicine) they identified the following stakeholders: a) scientists/researchers; b) those executing clinical trials; c) health care professionals; d) health care consumers; e) systems providing healthcare; f) policy decision makers; g) industry; and h) society at large. For each stakeholder they identified different challenges and opportunities for biomedical informatics overall. From an evaluative point of view the study identified a number of gaps and synergies between the fields of bioinformatics and medical informatics.

The result of the study was a list of research priorities proposed by the BIOINFOMED study. Each item on the list included a description of the barrier(s) (e.g. the challenges), a proposed solution (e.g. the opportunities), a priority rating and a risk rating. The prioritization was High vs. Medium. The risk was defined as the probability that focusing on the research priority would fail to deliver results and given a rating of High, Medium, or Low risk. The items were grouped into four areas. The first area was enabling technologies. An example of one item is, "Barrier: Need to expand current interoperability standards for new genetic data infrastructure, Proposed Solution: Data Communication Standards, Priority: High, Risk: Medium." The second area was medical informatics in support of functional genomics. An example of one item is, "Barrier: Patient care data have not been systematically used in genomic research, Proposed Solution: phenotype databases suitable for genomic research, Priority: High, Risk: Low." The third area was bioinformatics in support of individualized healthcare. An example of one item is, "Barrier: Unavailability of models for including genetic data into electronic health records, Proposed Solution - Genetics data model for the EHR, Priority: Medium, Risk: Medium." The fourth area was the unified field of biomedical informatics in support of genomic medicine. An example of one item is, "Barrier: Linking environmental and lifestyle information to genetic and clinical data, Proposed Solution: Population based repositories, Priority: High, Risk: Low."

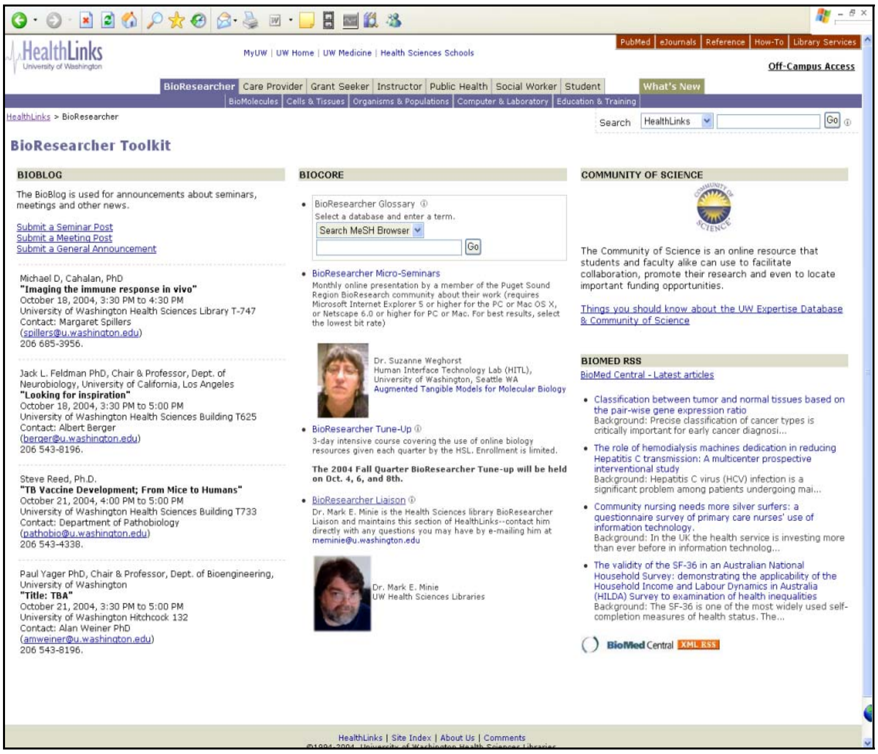
## 3.2 Biological Perspective – The BioResearch Liaison Program at the University of Washington

The University of Washington Health Sciences Library BioResearcher Liaison (http://healthlinks.washington.edu/hsl/liaisons/minie/) provides direct access to bioinformation consulting tools and training, and is a model program for contending with the issues discussed in Section 2.3. The BioResearcher Liaison program evolved out of an earlier effort called the BioCommons, and has been fully integrated into the Library's "informationists" infrastructure. The most visible part of this program is the BioResearcher Toolkit (http://healthlinks.washington.edu/bioresearcher) as shown in Figure 3-4, which provides a "portal" to biological information links, laboratory services, bioinformatics tools and consulting through the Library's Liaisons program (http://healthlinks.washington.edu/hsl/liaisons/). The contrast between the BIOINFOMED study and the BioResearcher toolkit is that the former lays out a research agenda for the future at the intersection of bioinformatics and medical informatics whereas the later is designed to address problems here and now. It is informative to compare and contrast the two case studies looking at the difference between grand challenges and on the ground realities.

The BioResearcher Toolkit is the second most visited part of the HealthLink's website (http://healthlinks.washington.edu/) after the more clinically oriented Care Provider Toolkit (see website at: http://healthlinks.washington.edu/care_provider/) with over 3,000 unique hits per month.

Since the consolidation of the BioCommons into the Library in 2002, the networked software and webware offerings have been the most used part of the BioResearcher Toolkit part of the website, with over 800 registered users of the various software packages available from the site and over 1,200 downloads over the past two years. These users are from that total pool of faculty, staff and students at the University of Washington, and come from large variety of departments as shown in Figure 3-5.

In addition to the BioResearcher Toolkit, the BioResearcher Liaison also provides a 3-day course given every quarter, the BioResearcher Tune-Up. The BioResearcher Tune-Up is a 3-Day intensive class with three modules—Module I: NCBI Online, Module II: Bioinformatics Software Workshop and Module III: Advanced Topics. Module I is a highly interactive tutorial which is taught in a computer lab using a web based PowerPoint template that allows students to directly follow the trainer through a tutorial on how to use NCBI databases using a single biologically relevant example: Huntington's disease.



HealthLinks
University of Washington
MyUW | UW Home | UW Medicine | Health Sciences Schools
PubMed | eJournals | Reference | How-To | Library Services
Off-Campus Access
BioResearcher | Care Provider | Grant Seeker | Instructor | Public Health | Social Worker | Student
What's New
BioMolecules | Cells & Tissues | Organisms & Populations | Computer & Laboratory | Education & Training
HealthLinks > BioResearcher
Search
HealthLinks
Go

## BioResearcher Toolkit

**BIOBLOG**

The BioBlog is used for announcements about seminars, meetings and other news.

Submit a Seminar Post
Submit a Meeting Post
Submit a General Announcement

Michael D, Cahalan, PhD
**"Imaging the immune response in vivo"**
October 18, 2004, 3:30 PM to 4:30 PM
University of Washington Health Sciences Library T-747
Contact: Margaret Spillers
(spillers@u.washington.edu)
206 685-3956.

Jack L. Feldman PhD, Chair & Professor, Dept. of Neurobiology, University of California, Los Angeles
**"Looking for inspiration"**
October 18, 2004, 3:30 PM to 5:00 PM
University of Washington Health Sciences Building T625
Contact: Albert Berger
(berger@u.washington.edu)
206 543-8196.

Steve Reed, Ph.D.
**"TB Vaccine Development; From Mice to Humans"**
October 21, 2004, 4:00 PM to 5:00 PM
University of Washington Health Sciences Building T733
Contact: Department of Pathobiology
(pathobio@u.washington.edu)
206 543-4338.

Paul Yager PhD, Chair & Professor, Dept. of Bioengineering, University of Washington
**"Title: TBA"**
October 21, 2004, 3:30 PM to 5:00 PM
University of Washington Hitchcock 132
Contact: Alan Weiner PhD
(amweiner@u.washington.edu)
206 543-8196.

**BIOCORE**

- BioResearcher Glossary
  Select a database and enter a term.
  Search MeSH Browser
  Go
- BioResearcher Micro-Seminars
  Monthly online presentation by a member of the Puget Sound Region BioResearch community about their work (requires Microsoft Internet Explorer 5 or higher for the PC or Mac OS X, or Netscape 6.0 or higher for PC or Mac. For best results, select the lowest bit rate)

  Dr. Suzanne Weghorst
  Human Interface Technology Lab (HITL),
  University of Washington, Seattle WA
  Augmented Tangible Models for Molecular Biology
- BioResearcher Tune-Up
  3-day intensive course covering the use of online biology resources given each quarter by the HSL. Enrollment is limited.

  **The 2004 Fall Quarter BioResearcher Tune-up will be held on Oct. 4, 6, and 8th.**
- BioResearcher Liaison
  Dr. Mark E. Minie is the Health Sciences library BioResearcher Liaison and maintains this section of HealthLinks--contact him directly with any questions you may have by e-mailing him at meminie@u.washington.edu

  Dr. Mark E. Minie
  UW Health Sciences Libraries

**COMMUNITY OF SCIENCE**

The Community of Science is an online resource that students and faculty alike can use to facilitate collaboration, promote their research and even to locate important funding opportunities.

Things you should know about the UW Expertise Database & Community of Science

**BIOMED RSS**

BioMed Central - Latest articles

- Classification between tumor and normal tissues based on the pair-wise gene expression ratio
  Background: Precise classification of cancer types is critically important for early cancer diagnosi...
- The role of hemodialysis machines dedication in reducing Hepatitis C transmission: A multicenter prospective interventional study
  Background: Hepatitis C virus (HCV) infection is a significant problem among patients undergoing mai...
- Community nursing needs more silver surfers: a questionnaire survey of primary care nurses' use of information technology.
  Background: In the UK the health service is investing more than ever before in information technolog...
- The validity of the SF-36 in an Australian National Household Survey: demonstrating the applicability of the Household Income and Labour Dynamics in Australia (HILDA) Survey to examination of health inequalities
  Background: The SF-36 is one of the most widely used self-completion measures of health status. The...

HealthLinks | Site Index | About Us | Comments
©1994-2004, University of Washington Health Sciences Libraries

*Figure 3-4.* The BioResearcher Toolkit.

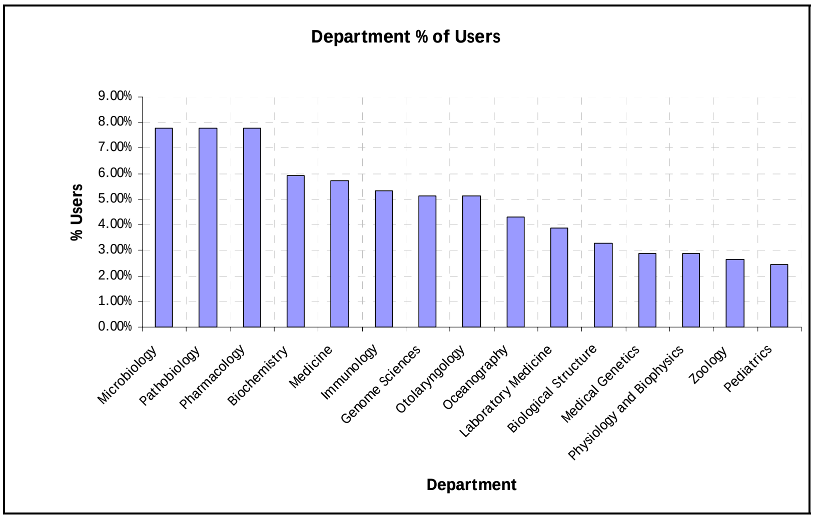


*Figure 3-5.* Users of the BioResearcher Toolkit

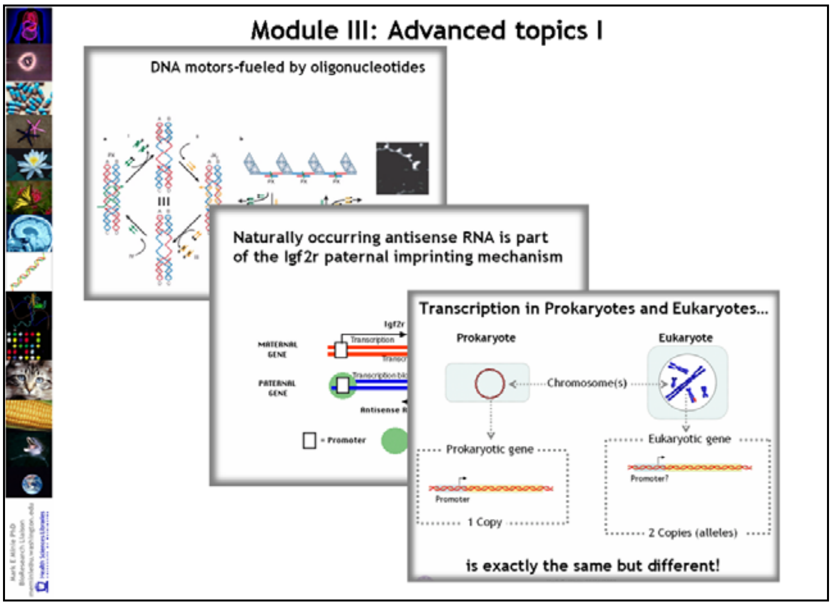
The Huntington Disease theme allows exploration of every database on Entrez, Blast and LocusLink (as well as Entrez Gene) and thus allows for a simple way to follow a story that touches on all aspects of the disease from the molecular level on up. Additionally, the module allows for the demonstration of discovery through digital data mining—a heretofore little known relationship between Huntington's disease and Type I diabetes is revealed while exploring the expression resources on GEO and SAGE Genie. The course is highly popular, with more applicants than there is room each quarter, and each attendee evaluates each module using an online form identical to that used by NCBI to evaluate similar modules taught by the EDUCOLLAB Group. Typically, this Module is rated as "Very Good" to "Excellent."

Module II is usually run by a guest vendor, who typically presents a tutorial workshop on a bioinformatics tool available from the BioResearcher Toolkit Computing and Laboratory section. For example, a Workshop/Tutorial on GeneSifter (www.genesifter.net), a web-based microarray analysis tool for analyzing gene expression data has been given as part of the BioResearcher Tune-Up with students participating both in the computer lab onsite and offsite via WebEx. The use of WebEx in particular is highly interactive, and has allowed more students, faculty and staff to attend than would be otherwise possible. Note also that one of the services provided by GeneSifter is the archiving of raw data, and the means to release that data in a highly interoperable format to the general scientific public in compliance with NIH's new rules on this issue.

Module III, the "Advanced Topics" part of the Tune-Up, covers a wide variety of relevant research oriented topics, ranging from seminars on DNA based nanotechnology to eukaryotic gene regulatory mechanisms as shown in Figure 3-6.

The BioResearch Liaison program also provides for one-on-one consulting with basic biology researchers at the University of Washington. For example, a client recently requested assistance in identifying a simple bioinformatics tool that would process molecular sequence data into graphical maps of alternative splice products for the gene studied. This led to a recommendation for the freely available NIH/NCBI tool "SPIDEY." A web-based open source program that was readily adapted to the clients needs. All consult encounters are followed-up with online evaluation forms to track the BioResearch Liaison's effectiveness, and the results are usually "Very Good" to "Excellent."

Recently, the Health Sciences Library BioResearch Liaison provided a version of the BioResearcher Tune-Up as an online training session to Alaska, Hawaii, Montana, Nevada, Utah, and Wyoming using the Access Grid videoconferencing technology shown in Figure 3-7. This provided a training session focused on the use of the sequence alignment tool BLAST (http://www.ncbi.nlm.nih.gov/BLAST/), and was a success from both the technical and teaching perspectives—the conferencing technology worked without glitches and online evaluations of the course by attending students gave it a "Very Good" to "Excellent" rating.



Module III: Advanced topics I

DNA motors-fueled by oligonucleotides

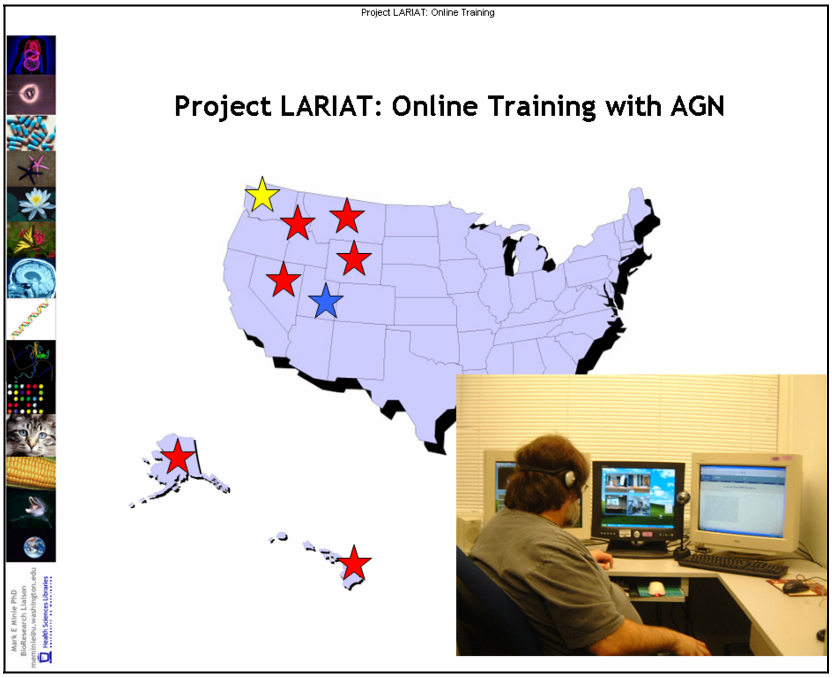
Naturally occurring antisense RNA is part of the Igf2r paternal imprinting mechanism

Transcription in Prokaryotes and Eukaryotes…

is exactly the same but different!

*Figure 3-6.* Advanced Topics section of the Tune-Up

Finally, the BioResearch Liaison program has had a number of notable successes in Public Outreach, with an important one involving providing a presentation to a large audience of science fiction writers and their publishers on how to access genomics information online as part of the "Science Friday" part of the recent 2004 Nebula Awards Conference in Seattle. One end result: an offer to publish a scientifically factual review of molecular biology and genomics in a prominent science fiction magazine widely read by the general public.



Project LARIAT: Online Training

Project LARIAT: Online Training with AGN

Mark E Minie PhD
BioResearch Liaison
meminie@u.washington.edu
Health Sciences Libraries

*Figure 3-7.* An Online BioResearcher Tune-Up

## 4. CONCLUSIONS AND DISCUSSION

The field of bioinformatics (defined both as foundational research and applied development of systems in support of biomedical research) presents a number of exciting challenges and opportunities for biologists, computer scientists, information scientists and bioinformaticists. These challenges sit at the intersection of biology and information. Ideally, larger scale work in this broad area involves a partnership between those with expertise in relevant foundational domains (e.g. computer scientists) and application domains (e.g. biologists) as well as bioinformaticists to serve as a bridge. The potential benefits of addressing some of these challenges are great both in terms of improving our understanding in general of how biological systems work and in terms of applying a better understanding of how the human biological system works in order to help improve health and treat disease.

Though many definitions of bioinformatics exist we have chosen to focus on the more inclusive definition to provide a richer picture of the opportunities and challenges. Indeed, it is possible that from the new perspectives of this more broadly defined bioinformatics the very informational nature of living systems may lead to a paradigm shift in biology. Our illustrations using specific examples nevertheless represent only a subset of the potential opportunities, and the inclusion of the broader framework for categorizing research papers will perhaps stimulate a reader of this book to look at the domain in a new way leading to unanticipated benefits to the field more broadly.

An important and often neglected area in biomedical informatics broadly (and bioinformatics by extension) is the human dimension captured in the socio-technical aspects of biomedical systems. In this context it is important to note two observations from the field of clinical (medical) informatics: a) the majority of applications developed in the lab have failed to be successfully deployed in the real world, and b) the majority of time, these failures relate to human factors rather than technical factors. It is also noteworthy that in addition to training scientists in the field in the use of online bioinformatics resources such as NCBI's Entrez, the very accessibility of these tools on the Internet allow for the possibility that the general public may directly use and possibly even participate in the further development of a "digital biology." Unlike other major developments in science in the 20th Century, the inherently "webified" nature of genomics information makes it relatively accessible to all - amateur scientists really can "try this at home."

We have presented two closely related but contrasting perspectives (the biological and the informatics perspectives) on the opportunities and challenges for bioinformatics. We have done so to a) illustrate some subtle but important distinctions, and b) demonstrate the value of having diverse perspectives as one explores the field of bioinformatics.

Finally, we have illustrated again the potential benefits of further work in this field through two case examples which also illustrate how researchers are going about trying to realize this potential. Here again the differences in the informatics and biological perspectives are worth noting. Particularly intriguing are both the emerging realization from both perspectives that biological systems are inherently digital, and the emerging parallel "Digital Biosphere" deriving from bioinformatics research activities. A true theoretical biology is at last emerging, where it may eventually be possible to understand complex biological systems by modeling them *in silico*. Significant progress in this direction has already taken place, with the publication of a detailed computer model of the regulatory network responsible for the control of flagellar biosynthesis in E.coli based on quantitative gene expression data (Herrgard and Palsson, 2004). This model is now being tested against the well defined genetic system of E. coli, and has already provided a system for developing new insights into this biological process. Particularly intriguing and revealing is the ready exchange of information between the *in vivo* and *in silico* systems.

## 5. ACKNOWLEDGEMENTS

Peter Tarczy-Hornoch: NIH NHGRI and NLM. “Interdisciplinary Center for Structural Informatics,” #P20 LM007714; “BioMediator: Biologic Data Integration and Analysis System,” # R01 HG02288; “Biomedical and Health Informatics Training Program,” #T15 LM07442.

## REFERENCES

Bergeron, B. (2002). *Bioinformatics Computing,* Upper Saddle River NJ: Prentice Hall.

Bodenreider, O., Mitchell, J., and McCray, A. (2002), “Evaluation of the UMLS as a Terminology and Knowledge Source for Biomedical Informatics,” in *Proceedings of the AMIA Symposium,* p. 504-8.

Collins, F.S. and McKusick, V.A. (2001). “Implications of the Human Genome Project for Medical Science,” *JAMA*, 285(5), 540-544.

Collins, F.S., et al. (2003). “A Vision for the Future of Genomics Research,” *Nature*, 422(6934), 835-847.

Donelson, L., Tarczy-Hornoch, P., and Mork, P. (2004). “The BioMediator System as a Data Integration Tool to Answer Diverse Biologic Queries,” in *Proceedings of MedInfo*, IMIA, an Francisco, CA.

Florance, V., Guise, N., and Ketchell, D. (2002). “Information in Context: Integrating Information Specialists into Practice Settings,” *Journal of the Medical Library Association,* 90(1), 49-58.

Friedman, C. and Wyatt, J. (1997). *Evaluation Methods in Medical Informatics,* New York: Springer.

Gene Ontology Consortium, (2000). “Gene Ontology: Tool for the Unification of Biology,” *Nature Genet.* 25, 25-29.

Herrgard, M. and Palsson, B. (2004). “Flagellar Biosynthesis in Silico: Building Quantitative Mdels of Regulatory Networks,” *Cell*, 117(6), 689-90.

Ihaka, R. and Gentleman, R. (1996). R. “A Language for Data Analysis and Graphics,” *Journal of Computational and Graphical Statistics,* 5, 299-314.

Joy, B. (2000). “Why the Future Doesn't Need Us,” *Wired,* 8(4).

Klein, T.E., et al. (2001). “Integrating Genotype and Phenotype Information: An Overview of the PharmGKB Project,” *The Pharmacogenomics Journal*, 1, 167-170.

Lyon, J., et al. (2004). “A Model for Training the New Bioinformationist,” *Journal of the Medical Library Association*, 92(2), 188-195.

Martin-Sanchez, F., et al. (2004). “Synergy between Medical Informatics and Bioinformatics: Facilitating Genomic Medicine for Future Health Care,” *Journal of Biomedical Informatics,* 37(1), 30-42.

Mei, H., et al. (2003). “Expression Array Annotation Using the BioMediator Biological Data Integration Systems and the BioConductor Analytic Platform,” in *Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium.* Washington, DC: American Medical Informatics Association.

Mork, P., A. Halevy, and Tarczy-Hornoch,P. (2001). “A Model for Data Integration Systems of Biomedical Data Applied to Online Genetic Databases,” in *Journal of the American Medical Informatics Association, Fall Symposium Supplement,* p. 473-477.

Mork, P., et al., (2002). “PQL: A Declarative Query Language over Dynamic Biological Schemata,” in *Proceedings of the American Medical Informatics Association Fall Symposium*, p. 533-537.

PharmGKB, *http://www.pharmgkb.org*.

Rosse, C. and Mejino, J. (2003). “A Reference Ontology for Bioinformatics: The Foundational Model of Anatomy,” *Journal of Biomedical Informatics*, 36, 478-500.

Shaker, R., et al. (2004). “The BioMediator System as a Tool for Integrating Biologic Databases on the Web,” in *Proceedings of the Workshop on Information Integration on the Web, held in conjunction with VLDB 2004.*

Shaker, R., et al. (2002). “A Rule Driven Bi-directional Translation System Remapping Queries and Result Sets between a Mediated Schema and Heterogeneous Data Sources,” in *Journal of the American Medical Informatics Association, Fall Symposium.*

Shortliffe, E., et al. (2001). *Medical Informatics: Computer Applications in Health Care and Biomedicine.* Second ed. New York, Springer-Verlag.

Stanford (2002). Protegé Home Page, *http://protege.stanford.edu/*

Sujansky, W. (2001), “Heterogeneous Database Integration in Biomedicine,” *Journal of Biomedical Informatics,* 34(4), 285-98.

Tuttle, M., Nelson, J. (1994). “The Role of the UMLS in 'Storing' and 'Sharing' across Systems,” *Internet Journal of BioMedical Computing,* 34(1-4), 207-37.

van 't Veer, L., et al. (2002). “Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer,” *Nature*, 415(6871), p. 530-6.

Yarfitz, S. and D. Ketchell, “A Library-based Bioinformatics Services Program,” *Bulletin of the Medical Library Association*, 88(1), 36-48.

## SUGGESTED READINGS

Dan E. Krane. Michael L. Raymer. *Fundamental Concepts of Bioinformatics*. Publisher: Benjamin Cummings, 2003.

Provides a good overview of Molecular Biology and Biological Chemistry for the non-biologist then addresses important problems in bioinformatics for both the biologist and informaticist (sequence alignment, substitution, phylogenetics, gene identification, structure prediction, and proteomics).

Andreas D. Baxevanis. B.F. Ouellette, eds. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins.* 2nd Edition. Publisher: Wiley-Interscience. 2001.

Provides an overview of internet accessible tools for the biologist with an emphasis on NCBI resources aimed at a mixed audience of biologists and developers. Each section is a blend of the underlying biology, the computing principles involved, and some practical hands on advice and tips.

Stanley I. Leovsky, ed. *Bioinformatics Databases and Systems.* Publisher: Kluwer Academic Publishers, 1999.

Provides an excellent overview of biological databases and computing systems aimed more at the developer than the biologist but useful to both. A series of deployed bioinformatics databases are described in some detail by the developers of the databases (e.g. NCBI, KEGG, FlyBase). Then a series of deployed tools are described by their developers in a 2nd section (e.g. BioKleisli, SRS, ACDEB).

Z. Lacroix, and T. Critchlow, eds. *Bioinformatics: Managing Scientific Data.* Publisher: Morgan Kaufmann, 2003.

Provides an excellent overview from more of a computer science standpoint of data management issues in biological research with an emphasis on data integration using selected examples from both academics and industry.

## ONLINE RESOURCES

NIH Biomedical Information Science and Technology Initiative (BISTI)
http://www.bisti.nih.gov/

Report of the Working Group on Biomedical Computing, Advisory Committee to the Director, National Institutes of Health, Co-Chairs: David Botstein and Larry Smarr "The Biomedical Information Science and Technology Initiative"
http://www.nih.gov/about/director/060399.htm

NIH All About the Human Genome
http://www.genome.gov/10001772/

The University of Washington BioResearcher Toolkit
http://healthlinks.washington.edu/bioresearcher/

The NCBI website; the entry point for such search and analysis tools as Entrez and BLAST etc. Also a source for online tutorials on the use of PubMed, CN3D (free structure viewing tool), BLAST etc.
http://www.ncbi.nlm.nih.gov/

University of Pittsburg biolibrary website
http://www.hsls.pitt.edu/guides/genetics/

The Cornell University Life Sciences Library website, VIVO
http://vivo.library.cornell.edu/

## QUESTIONS FOR DISCUSSION

1. Define interesting challenges in knowledge management and data mining in biomedical informatics based on the primary bioinformatics literature.

2. Define interesting challenges in knowledge management and data mining in biomedical informatics based on the primary biomedical research literature or based on interviews with biomedical researchers.

3. Define key unmet needs related to bioinformatics tools based on the primary bioinformatics literature and the primary biomedical research literature. Compare and contrast the unmet needs from these two perspectives.

4. What are the implications of the different perspectives from biology, biomedical research, and bioinformatics research?

5. What are the possibilities for theoretical biology based on knowledge mining of biological databases?

6. Discuss the implications for medical systems of virtual patients and disease modeling.

7. Given the informational nature of biological systems—what are the implications for our definition and understanding of life?

8. Point-of-Care Diagnostics and biomedical informatics—what might be the implications for future medical care and costs?