Bioinformatics and the Future of Medical Research and Clinical Practice

DynPort Vaccine Company LLC, A CSC Company

DVC LLC
A CSC COMPANY

64 Thomas Johnson Drive
Frederick, Maryland 21702
Phone: 301-607-5000
www.csc.com/dvc

retweet this
Click here to post this on Twitter

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

The enormous advances in biological technology over the past four decades have led to a profound change in how information is processed; conceptual and technical developments in experimental and molecular biology disciplines such as genomics, transcriptomics, proteomics, metabolomics, immunomics, and countless other "omics" have resulted in a veritable sea of data with the potential to radically alter biomedicine. Yet, with this wealth of data comes a challenge, namely how to transform the data into information, the information into knowledge, and the knowledge into useful action.

Nearly coincident with the advances in biological science, and in fact rapidly outpacing such advances, has been the advent of the modern computer and the associated advances in information storage, retrieval, and processing made practical with microelectronics and informatics. The power of modern information technology is ideal for capturing and storing the huge volume of biological data being generated; however, the respective languages and concepts of biology and computer sciences have, until recently, been disparate enough to prevent the logical next step of combining the two disciplines into a more powerful tool. The discipline of bioinformatics has emerged to capture the information stored in living systems and help turn it into actionable technology. In this paper we will explore the precepts of this discipline, the tools, and the potential for the future inherent in this powerful meta-technology.

1.1 Definition of Terms

"When I use a word, it means just what I choose it to mean -- neither more nor less."
– *Through the Looking Glass*

A truly meaningful discussion of bioinformatics must be grounded in a concise and consistent definition of terms. Unfortunately, there are multiple terms used interchangeably with "bioinformatics" that tend to confuse the discussion. For the purpose of this paper, we will define that term thusly: The creation and implementation of algorithms, computational and biostatistical techniques, and theory to solve formal and practical problems posed by or inspired from the management and analysis of biological data. A more thorough, and more eloquent, definition has also been proposed: "Bioinformatics is an interdisciplinary field that blends computer science and biostatistics with biological and biomedical sciences such as biochemistry, cell biology, developmental biology, genetics, genomics, and physiology. An important goal of bioinformatics is to facilitate the management, analysis, and interpretation of data from biological experiments and observational studies. Thus, much of bioinformatics can be categorized as database development and implementation, data analysis and data mining, and biological interpretation and inference" (Moore, 2007).

Along with this working definition of bioinformatics, it is instructive to also define several related terms that will be used in this review:

Computational Biology: The investigation of a specific biological problem using computers, carried out with experimental and simulated data, with the primary goal of discovery and the advancement of biological knowledge.

Convergence: Conceptual and practical linkages in a number of high-concept technologies that have the potential to be both self-reinforcing and transformational to human knowledge and experience. One key nexus is the so-called nanotechnology/biotechnology/information technology/cognitive science (NBIC) convergence.

Medical/Health Informatics: The design and implementation of systems and algorithms for improving the communications and application of medical treatment and health care. This would include but not be limited to electronic medical records, decision support systems and algorithms and medical databases. In the context of the present discussion, it differs from bioinformatics in that it concerns the management of existing clinical information, rather than the management and interpretation of raw biological data.

Synthetic Biology: A combination of biology/molecular biology and engineering concepts in which complex biological systems – particularly genetic sequence data - are subjected to various engineering techniques such as fabrication and modeling. Practitioners of synthetic biology seek to modify living systems at a fundamental (molecular) level. The tools offered by bioinformatics are key components of this discipline.

Systems Biology: An antithetical approach to biology as compared with the traditional reductionist model. Rather than "disassemble" complex biological systems and try to understand them in isolation, a systems biology approach views the complex system and seeks to understand it using novel models and techniques, including some of the tools provided by bioinformatics.

Translational Medicine: An evolution of evidence-based medicine that incorporates not only biological and clinical scientific observations, but epidemiology, social sciences, and various other disciplines (in this respect similar to systems biology, although with some obvious differences). Complex mathematical tools may be part of this discipline.

Figure 1 illustrates the relationship of several of these concepts.

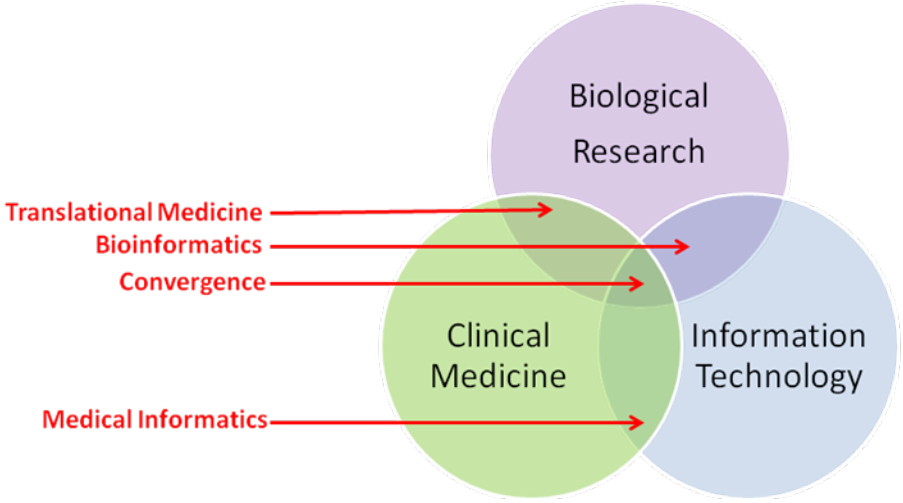

Figure 1 Relationship of Terms

1.2 The Language of Life

A truly adequate exploration of bioinformatics begins with the language of life – the genetic code of humans and other living organisms and how this set of instructions can be read and understood. The basic operating code for all life on Earth is encoded by the complex molecule deoxyribonucleic acid, DNA. Information in DNA takes the simple form of the letters A, G, C and T, standing for four chemical entities ("bases") within the DNA molecule, and the DNA molecule is double-stranded so each of these bases is paired with complementary base (A pairs with T, and G pairs with C). Various combinations of these four bases represent individual amino acids, and combinations of various amino acids form peptides and proteins which are the essential building blocks for all physiological processes. Upon the sequencing of the first human genome (the total sequence of all bases in human DNA), it was discovered that there are approximately 20,000 to 25,000 genes, comprising a surprisingly small percentage of the total genome. The remaining genome contains various regulatory sequences and other forms of DNA, some of which have no known function. The human genome contains approximately three billion base pairs, representing an enormous data set. Naturally, all organisms have unique genomes, many of which are much larger than the human genome, and must be sequenced individually. The various processes for decoding and understanding DNA sequences, regardless of species, are collectively known as ***genomics***.

For a cell to make use of this information, it must first be copied (transcribed) and the message must be translated into proteins. To accomplish this, a second molecule, ribonucleic acid (RNA) "unzips" the double-stranded DNA molecule and then creates a negative copy of the stored information. The RNA molecule then works with a cellular structure known as the ribosome;

using the RNA template, the ribosome stitches together various amino acids into proteins. The study of this process is termed ***transcriptomics***.

To begin to understand life processes on a fundamental level, it is next necessary to determine the identity and function of the multitude of proteins. Proteins are complex molecules which are the functional and structural building blocks of all living processes. Their chemical makeup determines their three-dimensional geometry (their so-called tertiary structure) which is crucial for how they interact with each other and their environment. The study of proteins at this level is termed ***proteomics***, and this is perhaps one of the fastest-growing omics-type analyses.

Finally in our brief tour are the metabolic processes resulting from this complex interaction of the proteins. Living organisms must processes basic materials into fuel and building and repair materials for maintenance of life, the process of **metabolism** in which oxygen, water and various nutrients are changed from one form into another and waste products are formed. At a higher level, any materials impinging on a life form (such as toxins, environmental agents or pharmaceuticals) are also metabolized by organisms, with either deleterious or beneficial results. Analysis of these processes is known as **global metabolic profiling**, with the primary concerns being **metabolomics** (the evaluation of by-products resulting from metabolic processes [that is, metabolites]) and the more detailed **metabonomics,** which is the assessment of metabolomics in a systems biology framework.

While these are four forms of "omic" analysis, the suffix has been appended to scores of other biological disciplines, perhaps reflecting the trend toward increasingly sophisticated technologies available to study such processes (for example, lipidomics, bibliomics and the tongue-twisting chromononics). Each of these evolving disciplines is beginning to generate huge swaths of data, requiring a holistic approach for transforming the data into useable information.

1.3 Systems Biology, Synthetic Biology and the Future

Perhaps the two disciplines that will, in the future at least, derive the greatest potential impact from the use of bioinformatics to draw disparate data together are systems biology and synthetic biology. Systems biology is the concept of understanding biology as a system of systems which do not exist in isolation, but rather are in turn part of much metasystems. For example, to understand biology it is also necessary to understand chemistry, physics, mathematics, and so on, and to understand how biological processes function in relation to these other disciplines. This way of seeing biology is holistic, which stands in contrast to the more reductionist approach that has characterized biology in the past which studied life by highly controlled and narrowly focused experiments. Instead, a bioinformatics approach is useful in systems biology since the various processes described above can not only be examined in minute detail, while also facilitating the comparison of data across multiple systems. Using systems biology, researchers seek to determine the complex interaction between genes, proteins and metabolic products to understand how cellular functions are affected by disease, drug toxicity or drug efficacy. One example of how this can be accomplished is by the previously described global metabolic profiling which can be applied in both preclinical and clinical development stages of drug development, as well as to study disease mechanisms. The metabolic profile encodes the

phenotype, which is composed of the genotype and the result of environmental factors (Schnackenberg, 2007).

Using the environmental approach, the human organism can be evaluated as a system of systems. In fact, the modeling of mammalian systems (such as humans) represents an enormous challenge given not only that the external environment must be taken into consideration, but the internal environment as well. Humans carry such a diverse community of commensal and parasitic microorganisms that they can be considered to be superorganisms (Nicholson et al., 2004). The powerful tools of genomics and bioinformatics are vital to understanding this dynamic system as well as the concepts of network theory (Almaas, 2007).

Synthetic biology, as defined earlier, is a nexus of biology and engineering in which biological processes are parsed to their most elemental form, and these structures and functions are then engineered using molecular biology, nanotechnology, and microfabrication, resulting in standardized "parts" which can then be reassembled into purpose-driven structures. One of the most intriguing efforts in synthetic biology is work by Craig Venter and his colleagues to create what is in essence a man-made organism. In this work, Venter's team has synthesized large sequences of a bacterial genome *de novo* using standard gene synthesis tools, and some novel approaches to stitch together these large sequences into a full genome (Gibson et al., 2008). Eventually, this synthetic genome will be inserted into a small bacterium (*Mycobacterium genitalium*) that has had its own genome stripped out. Ideally, the bacterial cell will "reboot" from instructions encoded in the synthetic genome. While not technically an artificial life-form, it will be very close to one. This will then serve as a precedent for the creation of genetically tailored organisms with functionalities not previously imagined.

While this technology is expected to open a whole new era in industrial biology with great benefits to mankind, the very power in this approach suggests a dark side as well. As detailed in the report entitled *Synthetic Biology: Social and Ethical Challenges* (www.bbsrc.ac.uk/organisation/policies/reviews/scientific_areas/0806_synthetic_biology.pdf), there are many challenges raised by synthetic biology including the danger of creating biological weapons, uncontrolled release of engineered life-forms with unexpected consequences, and the ethics regarding the creation of life from non-life. Clearly, synthetic biology has the potential to become a true Pandora's box.

2.0 THE TOOLS OF BIOINFORMATICS

For the sake of this discussion, we will limit our discussion of tools to two broad categories. First, the standard and emerging tools used to generate genomic/transcriptomic/proteomic data (**data generation tools**), and second the heuristic and systems biology tools necessary to store, retrieve and process these data into actionable information (**data analysis tools**). These concepts will appear again later in this paper in terms of future needs.

2.1 Data Generation Tools

Fantastic increases in both the scope and the speed of technological innovations over the past several decades have resulted in the ability to parse the language of life with ever-increasing speed and fidelity. The human genome consists of slightly more than three billion base-pairs comprising about 25,000 protein-coding genes, along with miscellaneous genetic sequences such as regulatory sequences, non-coding (so-called "junk" DNA, although this is almost certainly an oxymoron), repeat elements, pseudogenes, and so forth. Surprisingly (and to some, disappointingly), the human genome is not even the largest known genome. Thus, one can quickly see that efforts to sequence the genome of multiple species must deal with huge amounts of raw data. In keeping with the pattern established above, we will next briefly discuss the tools currently in use to extract such data at the genomic, transcriptomic, proteomic and metabolomic levels.

2.1.1 Gene sequencing (genomics)

The original methods for gene sequencing (Maxam-Gilbert and chain-termination) are very laborious and not conducive to large-scale analysis of entire genomes. With the advent of automated sequencers, the process became somewhat easier, but still fairly time-consuming. In addition, the techniques are subject to a certain degree of error. Newer methods of sequencing, termed "next-generation", are now making this process much more streamlined and are allowed the relatively rapid elucidation of entire genomes. Some of these newer techniques include *in vitro* clonal amplification, emulsion PCR (part of the "454" technology), parallelized sequencing, ligation sequencing and microfluidic sequencing. The specific mechanisms whereby these various technologies work are beyond the scope of the present paper. However, the result is that all of these technologies produce massive amounts of data. More importantly, there are not technologies that can sequentially read the entire three-billion base-pair genome (of humans) with complete fidelity. Rather, all of these techniques rely on sequencing large fragments of the genome, and the resulting maps must be assembled into an intact sequence. Tools to such data assembly are in use, but more research is needed.

2.1.2 Microarrays (transcriptomics)

Currently, the bulk of bioinformatic-relevant information derives from DNA microarrays. In microarray technology, thousands of DNA elements are bound to a solid substrate (generally glass or silicon). These elements represent genes of interest or expressed sequence tags. Next, the total RNA in a sample of interest is extracted, and a complementary DNA (cDNA) construct is created. This cDNA is then plated on the microarray and any areas of complementarity are recognized by special tags, often fluorescent dyes. The patterns of hybridization are then analyzed by special software. In this way, the gene expression in any particular sample can be determined as long as the cognate genes are encoded by the microarray.

Microarray technology is fairly mature and, as a consequence, many sources of data analysis have arisen. Data formats for microarrays include: Gene Expression Markup Language (GEML), Microarray and Gene Expression Markup Language (MAGE-ML), Minimal Information about Microarray Experiment (MAIME), and Microarray Markup Language (MAML). Table 1 is a

highly abbreviated listing of software packages available for design of microarrays and visualization and interpretation of data from such arrays.

Table 1 Highly Abbreviated Listing of Software Packages Available for Design of Microarrays and Visualization and Interpretation of Data from Such Arrays

Note: Underlined text in this table is hyperlinked to the corresponding Web site.

Software Packages Available for Design of Microarrays and Visualization and Interpretation of Data from Such Arrays (Highly Abbreviated)		
ACID	Cyber-T	GEPAS
Acuity Enterprise Microarray Informatics	DAVID	GO-TermFinder
AMIADA	Dragon hopkins	HeatMap Builder
ArrayAssist (StratGene)	DNA-Chip Analyzer (dChip)	J-Express
Array Designer	Engene	MAExplorer
ArrayMiner	Expression Profiler (EBI)	Partek Discover
ArrayTools (BRB)	GEDA	Rosetta
Array Viewer	GeneCluster 2	S+ Arrayanalyzer (Insightful)
BAGEL	GeneMaths XT	SNOMAD
BASE	GenePattern	SpotFire DecisionSite (Functional Genomics)
BioConductor	GeneSifter	TIGR
CAGED - Cluster Analysis of Gene Expression Dynamics	GeneSight	TreeArrange
CGH-Miner	genetide	Vector Xpression™
Cleaver	GeneX	
Cluster	Gene Xplorer	

Sources:
http://lyle.smu.edu/~mfonten/research/MAsoftware.html
http://genome-www5.stanford.edu/resources/restech.shtml
http://www.genetools.us/genomics/Microarray%20software%20catalog.htm

While these programs are all very similar in their overall features, the sheer volume of available options suggests that some consolidation and standardization may be helpful going forward.

2.1.3 Protein analysis (proteomics)

Much of the current work in proteomics uses standard protein chemistry techniques such as 2D gel electrophoresis and 2D high-performance liquid chromatography to separate the protein mixtures (samples of interest) into discrete proteins, followed by identification by mass spectrometry (Bernas et al., 2006). Further discrimination is provided by immunoproteomics, in which electrophoresed gels are probed with labeled antibodies. A powerful new technology that should accelerate proteomics is the antibody microarray, in which the DNA elements of the standard microarray are replaced by antibodies specific for various proteins. This technology will accommodate the use of pattern recognition software similar to DNA microarrays (Wingren and Borrebaeck, 2009).

2.1.4 Metabolic analysis (metabolomics)

Currently, the three main types of metabolomic evaluation are targeted analysis, metabolic fingerprinting and metabolic profiling (Shulaev, 2006). Targeted analysis is the most fully developed of the three. In targeted analysis, a finite number of known metabolites are evaluated. Although this method provides good sensitivity and specificity, it is limited in terms of evaluating large numbers of potential metabolites, and fails to detect unknown metabolites. Metabolic profiling does not seek to identify specific metabolites; rather, this analysis forms a "snapshot" of the entire metabolic output, and then pattern-recognition analysis is performed to compare the results with a comparator data set. This type of analysis is well-suited for biomarker discovery and diagnosis in which actual mechanisms may be unknown. Metabolic profiling is similar to targeted analysis except that a more global (and less selective) range of metabolites are examined by means of nuclear magnetic resonance, mass spectrometry, and various types of spectrometry (Shulaev, 2006). As with other types of high-throughput analysis (such as genomics and transcriptomics), the potential data stream from metabolomics is huge, especially when one factors in the multitude of potential variables affecting metabolism (genetic predisposition, species differences, concomitant environmental exposure and potential toxicity, and so forth).

2.2 Data Analysis Tools

2.2.1 Data mining

To this point, we have essentially discussed techniques and technologies for prospective, *de novo* data generation and analysis. This presumes that such data are generated with a definite end in mind…that is, answers are formulated in response to specific questions. However, the volume of data being generated is now doubling every few years, and new paradigms are constantly being developed. As a result, "old" data should be re-evaluated in light of newer data, and previously unseen relationships between various types of data may now become more apparent. What is needed is a way to index, collate and analyze the mountain of data residing in the scientific literature, as well as other sources. This process of extracting new patterns hidden in data is known as data mining.

The theories and practices associated with data mining are collectively known as knowledge discovery in databases (Fayyad et al., 1996). A variety of languages and software programs have been developed for data mining including Predictive Model Markup Language (PMML), Cross-Industry Standard Process for Data Mining (CRISP-DM), KnowledgeSEEKER, GhostMiner, KEEL, Clementine, R, Viscovery and many other open-source and commercial programs. These programs all have in common characteristics that allow them to extract data from existing databases. Before such analysis can happen, however, multiple databases must be subjected to data cleaning (mapping data to consistent conventions, a non-trivial exercise; accurately representing missing data points; and accounting for "noise" in the system), and methods must be provided to create a logical access to the various forms of data, including off-line data and metadata. The process of cleaning and ensuring access is referred to as *data warehousing* (Fayyad et al, 1996). This latter point is particularly important in the context of bioinformatics due to the variety of data, and particularly as regards the sheer complexity of the data. Genomic sequences are massive data sets, and the chance for inadvertent error is always present.

Once data warehousing has been accomplished, data mining generally proceeds along six tasks:

- Classification: Arrangement of the data into predefined groups. Common algorithms include nearest neighbor, naive Bayes classifier and neural network;
- Regression: Seeking to identify a function which models the data with the least error;
- Clustering: Similar to classification but the groups are not predefined, so the algorithm attempts to group similar items together;
- Summarization: methods for finding a compact description of a data subset;
- Dependency modeling: methods for finding significant dependencies between variables in a model;
- Change and deviation detection: discovering the most significant changes in a data set from a pre-established norm.

An important factor to consider with data mining is that hypotheses can be set, and in some cases addressed/supported, without the need for wet laboratory work. That is, *in silico* models can be developed and tested, perhaps with supportive data being derived from traditional laboratory investigations (Loging et al., 2007). Data mining using bioinformatics has already shown much promise, and successes enjoyed to date should pave the way for additional successes in the future (Baudis, 2006; Haoudi and Bensmail, 2006; Phan et al., 2006).

2.2.2 Systems for sharing data and results

One easily overlooked need as bioinformatics-type work expands is how data networks will evolve (or be developed) to allow individuals and groups to organize into data communities (my term). The term for such networks is content management systems. At present, there are many open-source tools available to accomplish the needs of such communities, including collaborative Web sites, conference Web sites, databases of content and laboratory intranets (such as wikis) (Mooney and Baenziger, 2008). The sheer diversity of such systems makes it difficult to determine which tools to use, and in the future some sort of standardization might be of great benefit.

3.0 THE PROMISE OF BIOINFORMATICS IN MEDICINE AND PUBLIC HEALTH

3.1 Building a Better Vaccine

The immune system is arguably one of the most complex of all vertebrate physiological systems. This complexity makes it a logical candidate for bioinformatics applications. One especially important application of immunological research is the development of vaccines, particularly vaccines for infectious agents (as opposed to the newer and less-proven therapeutic vaccines). Vaccines have had immeasurable impact on human health, preventing death and morbidity in many millions of individuals. Until recently, vaccines have been developed essentially the same way that Jenner developed the first smallpox vaccine, namely by trial and error. In this approach, whole organisms against which vaccination is desired (generally killed, fractionated or otherwise inactivated to prevent infection), are injected into animals and the strength, quality, and duration of protective immunity are evaluated. Obviously, if fractions of the organisms are used, many hundreds of individual fractions might have to be tested; this is a time-, labor- and resource-intensive process. However, at least one advance in how vaccines are designed has arisen from the use of bioinformatics, namely reverse vaccinology (Davies and Flower, 2007).

In reverse vaccinology, the genome of a bacterium or other infectious organism is decoded, and the various genes are determined. Using sequence analysis, the genes most likely to make good vaccine candidate, such as those encoding proteins that are expressed on the organism's surface or those known to be associated with disease, are selected and proteins are expressed from those genes using recombinant technology. This way, a much more limited panel of candidates can be tested, with an expected higher rate of success than the chance associated with older technology (Capecchi et al., 2004; Scarselli et al., 2005). The tools of bioinformatics (sequence analysis, etc.) are of great value in the reverse vaccinology approach. However, as powerful as this approach is compared to older technologies, there has to date been a continuation of the previous approach of developing individual vaccines for each pathogenic organism. A more powerful approach would be to develop vaccines for entire classes of pathogens, or perhaps even disease indications irrespective of the causative agent. For this, new approaches will be required.

One such approach is the comparison of genomes across multiple strains of pathogens within selected species, or among various species within genera. The recent rapid advances in sequencing technology have resulted in the complete sequencing of several hundred bacteria, and over a thousand more are currently underway. Using a technique known as comparative genomic hybridization, the genomes of bacteria can be compared for sequence homology; from this, researchers can determine which genes are shared among various species (the so-called "core genome") (Mora et al., 2006). Such pan-genomic information will eventually be crucial in developing universal vaccines for certain classes of infectious organisms (Kaushik and Sehgal, 2008).

Obviously, the power of bioinformatics comes into play here due to the huge amount of data that must be analyzed. For example, rarely do isolated components of an organism produce effective immunity when injected as a vaccine; rather, specific combinations appear to produce the best result. This combination of potential vaccine components stimulates what has been termed the

"immunome" of the host (De Groot and Martin, 2003). An increasing number of bioinformatics programs are available to parse these data (see for example Davies and Flower, 2007 and De Groot et al., 2008).

An especially interesting possible application of bioinformatics to vaccine development would involve not only early discovery aspects, as have been described so far in this review, but also the integration of data streams from multiple simultaneous systems to accelerate and advance vaccine clinical trials. This concept is presented in Figure 2 below.

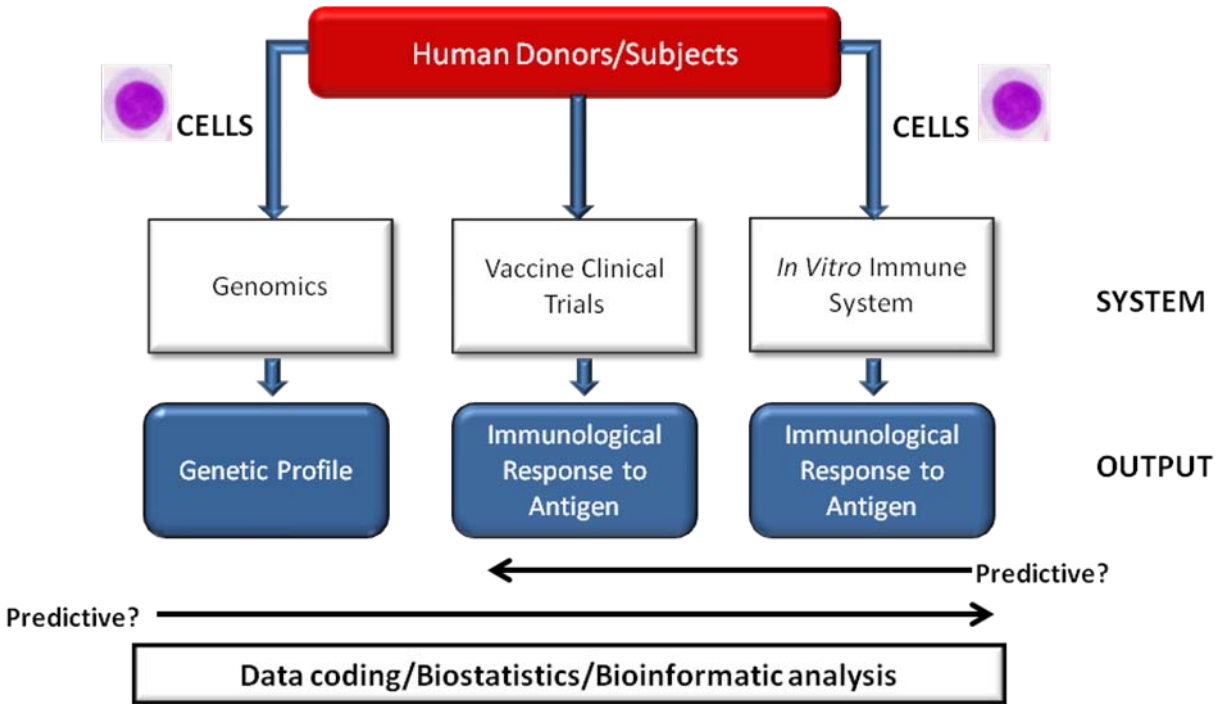

Figure 2 Next-generation Clinical Trial Design

In the initial stages of such an approach, volunteers would undergo a standard clinical trial in which the *in vivo* response (assessed by humoral immunity, cell-mediated immunity, or both) to a vaccine antigen is quantified. Concomitantly, the genetic phenotype of the volunteers is determined by donated cells; this addresses the important emerging concept of "vaccinomics" in which the genetic propensity of an individual dictates response to specific vaccination parameters (Poland et al., 2007). Combining this information with actual clinical results will form an important data set. Finally, the immune response would also be measured using a novel *in vitro* immune system surrogate such as the systems under development at VaxDesign (www.vaxdesign.com). Results from these disparate sources would be analyzed using specialized biostatistics and bioinformatics. The ultimate aim of such work would be 1) to determine whether genetic phenotype would predispose toward specific immune responses, and 2) to determine the concordance of *in vitro* immune responses to that observed *in vivo*. The key determinant in this system would be the ability to track these various responses in individual

humans, rendering a much higher degree of comparability in results. If this system proves viable, it would allow a "pre-screening" of vaccine candidates *in vitro*; most drugs or vaccines fail in the crucial Phase 1 study, and elimination of candidates prior to this stage would save both time and money. Moreover, combination of this information with the genetic phenotype data would explain when certain individuals might not respond, preventing companies from killing candidate vaccines or drugs that might otherwise be suitable for most of the population.

3.2 Developing a More Effective Response to Emerging and Re-emerging Diseases, Including Biodefense and Pandemic Influenza

Vaccines and therapeutics (mostly antibiotics) have been developed, and continue to be developed, for most of the routine infections including various childhood illnesses. In fact, it has not been so very long since the need to develop new antibiotics was declared – prematurely as it turns out – to be unnecessary. However, the increased mobility of the human population as well as climatic and geopolitical disruptions, has resulted in a new constellation of disease threats. More ominously, the ability to weaponize diseases adds another threat level not amenable to epidemiological or public health surveillance or control. The rapidity of disease emergence and spread, the high pathogenicity of some of these "new" diseases, and the rarity and unfamiliarity of some diseases all represent challenges that can be mitigated to some extent by the power of bioinformatics. First, some definition of (even more) terms is in order:

3.2.1 Emerging diseases

Emerging diseases can take the form of zoonotic diseases or geographically-displaced diseases. Zoonotic diseases are those that usually infect animals but through various routes infect humans, for example due to close contact not previously encountered. Hantavirus is but one example of this, in which humans were infected with the virus due to close contact with rodent urine. Filoviruses such as Ebola or Marburg are examples in which neither the exact host in nature or the route of infection in humans is entirely certain. Geographically displaced diseases occur as a result of widespread human travel, taking the disease (and sometimes its vector) to a new location.

3.2.2 Re-emerging diseases

Perhaps the most relevant example of a re-emerging disease is pandemic influenza, or "bird flu". Pandemic influenza has caused multiple rounds of pandemic disease throughout history, most recently 1917-1918 (the "Spanish flu"). The influenza virus occurs in multiple varieties (serotypes) and the precise serotype, and a population's past exposure to it, determines the severity of the disease. Pandemics generally occur every 70 to 100 years, so there is a high probability that another will occur soon (Solorzano et al., 2007). The morbidity and mortality associated with such pandemics would represent an enormous challenge, and governments around the world are preparing for this eventuality (Morens and Fauci, 2007). One of the ways that bioinformatics has already assisted in pandemic preparedness is by the re-creation of the 1918 virus, allowing medical scientists to study this supposedly extinct virus for clues to mitigate the next pandemic. (Garcia-Sastre, A. and Whitley, R.J., 2006). (Ironically, the technology that

allows us to recreate such virulent organisms also would allow the creation of pathogens for nefarious purposes, *vide infra*.)

3.2.3 Biodefense-related diseases

While mankind has worked prodigiously to produce weapons to kill its own kind, Nature has had millions of years of head-start. As man learned to recognize the basics of disease and contagion, he began incorporating disease organisms into his armamentarium. Within the past century, such efforts were consolidated into state-supported offensive, and more recently defensive, programs. With the increased accessibility of advanced microbiology, it has now become feasible that smaller organizations can develop biological weapons as mass casualty agents; biological warfare has morphed into bioterrorism (Khardori, 2006). Some of the various types of biological threats are listed below:

"Traditional" agents: these are naturally occurring microorganisms that may or may not be natural diseases of man, but are highly pathogenic and in some cases have no known treatment. Examples include plague, anthrax and smallpox. Traditional agents are classified as Category A, B or C depending on their potential danger, with Category A being the agents most expected to be used as offensive agents.

"Enhanced" agents: these are microorganisms that gain pathogenicity by more-or-less natural means, particularly mutation or other forms of natural selection, although they can also result from inadvertent selection such as the overuse/misuse of antibiotics. Such agents may gain a new host range (such as animal diseases gaining an ability to infect humans), resistance to antibiotics (best exemplified by methicillin-resistant *Staphylococcus aureus*, commonly called MRSA), or other enhancements in their ability to survive and produce disease.

"Advanced" agents: these are mostly hypothetical at present, but arguably represent the greater danger. Advanced agents are those that have been deliberately modified to have highly specific pathogenicity such as stealth characteristics or the ability to evade vaccination. Given their novelty (much like so-called "designer drugs"), these agents would be difficult to identify at first since normally innocuous microorganisms might be engineered as weapons.

3.2.4 The role of bioinformatics

With the Amerithrax letter terror campaign of 2001, the threat of bioterrorism has been proven to be possible, although the true practicality of mass casualty has yet to be established. Regardless, an entire industry has arisen in response to this perceived threat, with billions of dollars in research. Yet, in many cases, this huge influx of funding has had limited practical results in making the world safer. In most cases, vaccines and therapeutics are still being developed as they have been for decades (or in the case of vaccines, centuries). With emerging/re-emerging diseases such as SARS, routine epidemiology and disease control were shown to be quite effective, yet next time the world may not be so lucky.

There are at least two major roles that new technologies, including bioinformatics, can play in helping to mitigate the dangers posed by infectious diseases in this context: development of

advanced diagnostics and monitoring, and development of new vaccines and therapeutics, with the nexus being identification of specific molecular signatures and the host response to infection. The immune system in animals has evolved to recognize a variety of molecules or patterns of molecules on pathogenic organisms that constitute a danger to the host, and is able to differentiate these organisms from harmless or commensal organisms. Examples of such signals are pathogenicity-associated molecular patterns (PAMPs), Toll-like receptor agonists, pathogenicity islands, and a wide variety of virulence factors. These signals are recognized by host receptors such as ficolins, defense collagens, Toll-like receptors, and other mechanisms. Triggering of these receptors activates the innate immune system, a rapid-response defense system that works to quickly and non-specifically neutralize such dangers, and which activates the adaptive immune system to mount a specific, effective and long-lasting immunity.

In the pre-genomics era, development of vaccines and therapeutics was based to a large extent on trial and error or serendipity (the discovery of penicillin stands as the best example of the latter). Often, thousands of compounds would be screened for antibacterial or antiviral activity, and the mechanism of action would be determined later. Moreover, antibacterial drugs often kill beneficial organisms, an undesirable side-effect. Traditional antibiotics (and to a lesser degree, other antimicrobials) can lose effectiveness as bacteria develop resistance trough mutation and natural selection (Biswas et al., 2008). Finally, individual antibiotics are not always effective against multiple microbes, so treatment must be tailored to fit the infection, rather than the disease. It is here that the methods of genomics serve well, since full-genome analysis can identify sequences common to pathogens but absent in harmless organisms. In addition, proteomic analysis will yield even greater detail as it relates to gene expression and interaction of the pathogen with the host. However, full genomic analysis of all potential pathogens and their associated proteomes must be turned into useable information, which is done using bioinformatics tools (Drake et al., 2005; Khan et al., 2006; Christen, 2008; Holzmuller et al., 2008). Specifically, by identifying the various danger signals associated with human and animal pathogens, drugs and other therapeutics can be developed that 1) target only pathogens, sparing harmless or beneficial organisms, 2) act on a wide variety of pathogens, and thus treat many diseases at once, and 3) have less chance of losing efficacy since the drugs are targeted to physiological processes that confer the ability to produce disease. This concept is generally referred to as “one drug, many bugs” and is a key concept of the U.S. Department of Defense’s Transformational Medical Technologies Initiative, TMTI).

A related benefit of using bioinformatics to identify key elements of pathogens for drug and vaccine development is that these same molecular signatures would theoretically serve as ideal platforms for developing highly sensitive diagnostics and environmental sensors. In the context of some of the diseases mentioned here, a key advantage is that specific identification of organisms would not be necessary immediately. Rather, these signatures would detect pathogens of importance for humans and animals, and broad-spectrum treatment could be initiated. Perhaps the most valuable advantage would be the ability of such diagnostics or sensors to identify advanced biothreat agents, since at present there are no practical technologies to quickly identify such agents. A related approach has been developed in which single plasmids have been engineered to contain molecular signatures of multiple pathogens, rather than common pathogenic sequences (Carrera and Sagripanti, 2009). The use of specific genetic signatures to

determine whether an organism of interest has been deliberately engineered (rather than natural mutation) has recently been demonstrated (Allen et al., 2008).

An increasing number of databases and on-line tools are available for addressing the bioinformatics needs of this topic, including biodefense and influenza (for example, Greene et al., 2007; Glasner et al., 2008; Hari et al., 2008; Van Brabant et al., 2008; Zhang et al., 2008). As greater emphasis is placed on this approach to responding to these threats, the need for more sophisticated and powerful databases will only grow.

3.3 Personalized Medicine

For the entire history of medicine up until the last decade or so, human medical science has been based on empirical evidence of disease collected on a macro scale, followed by a deductive and often trial-and-error approach to determine an appropriate response. Over time and with consistent advances in biomedical science, both diagnosis and treatment have improved. However, the basic paradigm of "shoot and see" has remained the norm. This is primarily because biomedical research is based on averages…most people have a certain response to a certain disease, and most people will have a particular reaction to a particular treatment. Although this paradigm has clearly worked well for humans as a group, the exceptions to these average responses represent an unmet medical need. More specifically, many treatments must balance therapeutic efficacy with toxicity. Quoting Paracelsus, "Alle Ding sind Gift, und nichts ohn Gift; allein die Dosis macht, daß ein Ding kein Gift ist." ("All things are poison and nothing is without poison, only the dose permits something not to be poisonous."). Rather, the goal of personalized medicine is to develop therapies that increase the probability of success while decreasing the probability of toxicity.

A related concept is *translational medicine* (supported by translational research), in which many different novel pharmacology tools, biomarkers, clinical methods, clinical technologies and study designs are evaluated in a systems biology context to gain a greater understanding of disease processes and outcomes, increase confidence in drug targets and drug candidates, understand the therapeutic index, and enhance decision making in clinical trials (Littman et al., 2007). A major part of such a holistic approach would be the use bioinoformatics.

Although knowledge for knowledge's sake is important, arguably the greatest value of bioinformatics will be derived from its ability to improve human (and animal) health care. To date, our ability to acquire advanced biological data has outstripped the ability to transform the raw data into relevant medical knowledge, and thus the ability to develop products and practices of clinical applicability. While this nascent biomedical knowledge will eventually find its way into general medical practice, the exquisite specificity that will derive from individual genomics (in essence, an individual's "blueprint") should eventually make the concept of personalized medicine a reality. Such advances will be not only therapeutic, but improvements in preventive medicine such as the development of highly specific biomarkers (Collins et al., 2006), and estimation of susceptibility to various infectious and genetic diseases (as well as potential response to various medications) (Janssens and van Duijn, 2008; Tanaka, 2008).

While the technical advances necessary to make this promise a reality already exist, many ethical and administrative developments will be necessary as well before bioinformatics data can take their place alongside conventional data (Phillips et al., 2008; Shabo, 2008). For example, data maturity and standards (as previously discussed) will have to be consistent and well-developed since an incomplete understanding of the significance of such multivariant data will have immense ramifications once they become a component of standardized electronic medical records.

Physicians have already begun to employ genomic data to tailor diagnosis and treatment, particularly in the area of oncology. Full implementation of personalized/translational medicine as the standard paradigm has the potential to effect significant, industry-wide changes in how drugs are developed, but for now there are multiple ethical, legal, and technical challenges to overcome (Fitzgerald, 2005).

Another interesting consideration is the distinction in personalized medicine of product versus practice; to date, medical "products" have (as mentioned above) been intended to work for the greatest percentage of the population, and variations in efficacy and safety must be balanced for the greater good. With personalized, genomics-based medicine there are at least four areas that must be considered: validation of clinical claims for tests used in targeting therapies; developing and implementing appropriate restrictions on off-label use; promoting consistent concepts of clinical utility for use in regulatory, reimbursement, and judicial contexts; and communication of clear information to guide clinicians in appropriate use of targeted therapeutics (Evans, 2007). Another business consideration inherent in personalized medicine is the question of how the intellectual property associated with discoveries should be protected, especially as related to patenting. The U.S. patent system rewards innovation in medicine and other arts and sciences by granting inventors the right to exclude others from using their inventions for a defined period of time. Exclusive use of such technology may inherently limit the application of benefits to a wide range of individuals (Solomon and Sieczkiewicz, 2007). Moreover, there may be privacy issues associated with discoveries based on the exquisitely personal nature of genetic testing. Finally, an aspect that is receiving increased notice is the role of education of practicing physicians in the sometimes arcane art of molecular biology, forming a bridge between basic research and applied practice (Konstantinopoulos et al., 2008).

In the following sections, we will examine some of the areas in which bioinformatics is already being used to implement personalized medicine. This discussion represents only a survey of current practice, and it is expected that advances on these beachheads will encourage further implementation.

3.3.1 Individualized diagnosis and treatment of cancer

Treatment, and to a lesser degree diagnosis, of cancer has proven to be a nearly intractable problem. Part of this of course is that "cancer" is not a single disease but rather a constellation of many similar diseases. Perhaps more importantly, cancer is a biologically complex problem, and effective measures have not been forthcoming due to this complexity. To date, diagnosis of various types of cancer has relied on pathological, histological and morphological evaluation of resected tissues; while this method is sufficient to define many types of cancer (and thus the

prescribed course of therapy), it does not work for all types of cancer and tends to be retrospective after the cancer has already occurred. More importantly, this gross evaluation gives no insight into what are considered to be the six essential alterations in cells common to most human tumors; these include self-sufficiency in growth signals, insensitivity to growth-inhibitory signals, evasion of apoptosis, unlimited potential to replicate, sustained angiogenesis, and tissue invasion and metastasis (Dietel and Sers, 2006). A variety of technologies are now being applied to cancer research, including microarrays, proteomics, and epigenomics, to understand cancer at the cellular and molecular level as never before (Rivenbark and Coleman, 2007).

Cancer has long been recognized as having multiple etiologies, with the common feature that cancer cells tend to have inappropriate cellular signaling and gene expression. Functional genomics is revealing that cancer cells often have dysregulation of multiple genes (Dopazo, 2006) and the ability to evaluate such multiple genetic malfunctions is facilitated by bioinformatics tools. The ultimate aim of such specificity is to develop ***biomarkers***, which are measurable alterations in cancer cells at the genetic, protein, or metabolite level that 1) distinguishes them from normal cells, thus assisting in diagnosis, and 2) suggest or provide therapeutic options (Jain, 2007; Manning et al., 2007; Foekens et al., 2008). A few of the techniques in use or under development for diagnosis of cancer, as well as monitoring treatment modalities, include the following:

- DNA microarrays, in which the expression of various genes can be monitored and compared to normal gene expression profiles. This is the most widely used technique at present. Although this is a potentially powerful technique, it is subject to some limitations such as statistical power.
- Tissue microarrays, in which tissue samples (rather than cDNA transcripts) are arrayed in blocks and then interrogated using immunohistochemistry or fluorescence *in situ* hybridization (FISH).
- Comparative genomic hybridization (CGH), in which DNA from tumor tissue and normal tissue is differentially stained and then directly compared for differences.
- Expression of single nucleotide polymorphism (SNP, called “snips”), which are differentially expressed genes that vary among the human population and may provide clues to why some individuals are susceptible to disease, including certain cancers (Chorley et al., 2008).

3.3.2 Improving drug efficacy through pharmacogenetics/pharmacogenomics

Pharmacogenetics refers to the study of individual (genetic) variation as a factor in response to therapeutic drugs, and has existed as a discipline since the late 1950s. However, early studies were based on “old school” genetics based primarily on phenotypic expression not yet grounded in molecular biology. With an increasing understanding of the human genome, researchers discovered that specific human genotypes were associated with a differential response not only to drugs (generally for reasons due to metabolic differences), but with the course of various diseases (Nebert et al., 2008). Thus, a more complete understanding of an individual’s genetic makeup would be expected to be predictive, to some degree, of how s/he would respond to treatment with various therapeutics. Tanaka (2008) has characterized the ability to translate genomic/proteomic data to personalized care in three generations. In the first generation, care is

based on the polymorphism of germline genome sequences. An example of this would be the use of medication based on individual genetic differences of pharmacodynamics/phamarcokinetics or estimation of genotype relative risk for individual's disease occurrence. The second generation is based on the information pattern from diseased somatic cells, which brings about detailed classification, early diagnosis and prognosis of the disease. Finally, the third generation is based on a systematic understanding of complex diseases, thus enabling a holistic understanding of disease mechanisms.

Naturally, understanding the complex interactions between the host, the disease syndrome, and the therapeutic under investigation is far more complex than simply knowing an individual's genetic makeup. It is here that the power of bioinformatics can truly be realized, since this dynamic is best understood from a systems biology perspective (Yan, 2008).

3.3.3 Improving drug safety through toxicogenomics and systems biology

As important, and arguably more important, as improvements in the efficacy (and prediction of efficacy) of drugs promised by pharmacogenomics would be improvements in safety of drugs. At present, a huge component of the drug development process is the number and diversity (and consequently high cost) of *in vitro* and animal studies necessary to evaluate potential toxicity of any product prior to its introduction into humans in a Phase 1 clinical trial, which is itself generally an initial safety study as well. In addition, later-stage human trials (Phase 3 in particular) are sometimes enormous with tens or hundreds of thousands of volunteers in order to evaluate whether rare drug effects are likely. This adds years of time and many millions of dollars to the cost of developing drugs. Yet, side effects still show up, often severe enough (although generally very rare) to warrant discontinuation of testing since animal testing is never a true indicator of human biology. Consequently, this enormous investment in testing results in high prices for drugs due to the need to recoup investment costs. Clearly, what is needed is a better approach.

Safety testing is confounded by several factors:

1. Nonhuman animal physiology is not always analogous to humans, with the exception of nonhuman primates (which are not totally predictive). Thus, human response must always be extrapolated from animal data to some degree.
2. Unlike most laboratory animal models, humans are genetically diverse and thus quite variable in their response to foreign agents such as drugs. Thus, any findings (even in human studies) represent only an expected average of human response.
3. *In vitro* testing can only assess single, highly specific targets of drugs. However, toxicity is usually the result of multiple interactions, either by multiple targets of a single drug or multiple down-stream consequences following a single toxic insult (such as liver damage). Thus, animals models have to date been impossible to completely phase out.

Increasingly, *in silico* (so-called "virtual") methods and computational toxicology tools are being developed to increase sensitivity of toxicological assessment. The exquisite specificity of genomic and transcriptomic assessment promises to not only provide a greater degree of detail in terms of mode of action of drugs, but should allow for tailoring toxicity assessment at the genetic

level rather than at population level (Harrill and Rusyn, 2008; Muster et al., 2008). Even greater improvements in toxicity assessment, including the long-held promise of replacing the use of nonhuman animals with *in vitro* and *in silico* models will be realized once these approaches are coupled with such methodologies such as cytochrome P450 metabolism, blood-brain-barrier permeability studies, central nervous system activity, blockade of the hERG-potassium channels, homology models and quantum chemical calculations (Hutter, 2009). Finally, as with several of the areas discussed in the present review, a greater reliance on the inclusive paradigm of systems biology, rather than the currently prevailing reductionist approach, is expected to revolutionize toxicology (Edwards and Preston, 2008).

Perhaps one of the most significant challenges associated with this new approach to safety testing will be navigating the regulatory compliance process. When findings are reported to regulatory agencies, the relevance of such findings in the context of prior knowledge is incumbent upon a drug developer. However, the rapidly evolving nature of the various –omic technologies will increasingly place a burden on drug developers to continually reevaluate any findings with not only historical literature, but even the possibility of reanalysis of data as new methods of data handling (such as new bioinformatics tools) are developed. These data must then be interpreted in light of established and validated safety standards and biomarkers (Muster et al., 2008; Sistare and Degeorge, 2008).

3.3.4 Toward a unified practice: theragnostics

As the field of phamacogenetics matures, it will be useful to add more diverse and dynamic types of data to point of care in personalized medicine. While pharmacogenetics focuses on the use of genetic biomarkers to answer highly defined questions such as susceptibility or resistance to certain drugs, the relatively new field of theragnostics (a fusion of therapeutics and diagnostics) seeks to combine highly specific diagnostics with targeted therapy. This approach combines many of the modern tools of biology such as genomics and proteomics, and bioinformatics forms an integral part of the overall system (Pene et al., 2009).

An unfortunate paradox associated with the advancing development of theragnostics—and personalized medicine in general—is the uncertainty regarding the economics of targeted drug development. The development cost for broad-use pharmaceuticals (including biologics) now averages $800M to $1B, and drug companies contend that current drug prices (and large-scale use) are necessary in order to recoup this investment. When paradigms such as theragnostics have a much more narrow target population, new economic models will almost certainly be necessary (Ozdemir et al., 2006).

3.4 Welcome to the Edge: NBIC Convergence

As cutting-edge as bioinformatics (and its children such as theragnostics) is, the nexus between biology/biotechnology and information science that defines bioinformatics is part of a broader group of technologies roughly termed “converging” technologies. One of the more widely discussed convergences is termed NBIC (nanotechnology, biotechnology, information technology and cognitive science) (Chen and Ho, 2006). Whereas many world-changing promises have been made regarding this convergence, it is still in its conceptual infancy, and

great difficulties will likely be encountered as the various technologies begin to merge. However, inclusion of these additional technologies to the exquisite molecular diagnostic and therapeutic advances from bioinformatic analysis is expected to greatly enhance the potential power of personalized medicine (Yang et al., 2007). Such convergence would not be without its potential dark side, including the blurring of human and nonhuman, organic and inorganic. Many ethical and legal questions will likely be raised as the various new technologies breach traditional boundaries (Gordijn, 2006; Ziegler, 2006).

4.0 U.S. GOVERNMENT ACTIVITIES IN BIOINFORMATICS

4.1 National Institute of Allergy and Infectious Diseases (NIAID)

As the branch of the National Institutes of Health charged with basic research on infectious diseases, NIAID has a keen interest in the power of bioinformatics to shed light on infectious organisms and the mechanisms of disease and humans' response to those organisms. Below is a small sampling of some of NIAID's ongoing bioinformatics initiatives:

- Office of Cyber Infrastructure and Computational Biology (http://www3.niaid.nih.gov/about/organization/odoffices/omo/ocicb/). The Office of Cyber Infrastructure and Computational Biology (OCICB) manages technologies supporting NIAID biomedical research programs and provides management, technologies development, applications/software engineering, bioinformatics support, and professional development. OCICB works closely with NIAID intramural, extramural, and administrative staff to provide technical support, liaison, coordination, and consultation on a wide variety of ventures. These projects and initiatives are aimed at ensuring ever-increasing interchange and dissemination of scientific information within the Federal Government and among the worldwide scientific network of biomedical researchers.
- Bioinformatics Resource Centers (www.niaid.nih.gov/research/resources/brc/). The eight NIAID BRCs collect, store, display, annotate, query and update genomic and related data for pathogens of interest to NIAID. Sequence data will be integrated with gene expression and proteomics information, host/pathogen interactions and pathways data. The BRCs provide training on using their site and solicit genome annotation from the scientific community.
- Clinical Proteomics Centers for Infectious Diseases and Biodefense (www3.niaid.nih.gov/research/resources/cp/). The two NIAID Clinical Proteomics Centers apply state-of-the-art proteomics technologies for the discovery, qualification and verification of protein biomarkers in well-defined clinical samples as well as providing proteomic technology and expertise to the scientific community.
- Microbial Sequencing Centers (www.niaid.nih.gov/research/resources/mscs/). The Microbial Sequencing Centers provide rapid resources for producing high-quality genome sequences of pathogens and invertebrate vectors of infectious diseases. The scientific community may request microbial genome sequencing services.
- Pathogen Functional Genomics Resource Center (www.niaid.nih.gov/dmid/genomes/pfgrc/). The PFGRC provides scientists with genomic

resources and reagents, such as microarrays, protein expression clones, and bioinformatics services. The Center was established to provide the research community with a centralized resource to aid functional genomics research on human pathogens and invertebrate vectors of infectious diseases.

- Biodefense Proteomics Research Centers (www.niaid.nih.gov/dmid/genomes/prc/). Seven PRCs characterize the pathogen and/or host cell proteome, including the identification of proteins associated with innate and adaptive immune responses to infectious agents.
- Structural Genomics Centers for Infectious Diseases (www.niaid.nih.gov/research/resources/sg/). The two SGCs characterize the three-dimensional atomic structure of targeted proteins using state-of-the-art, high throughput structural biology technologies. Structure determination may be requested by the scientific community.
- Systems Biology Centers for Infectious Diseases (www3.niaid.nih.gov/research/resources/sb/). The four Systems Biology Centers are using a combination of computational and experimental methodologies to analyze, identify, quantify, model and predict the overall dynamics of microbial organisms' molecular networks including their host interactions.
- The Immunology Database and Analysis Portal (ImmPORT) (www.immport.org) archives data from the research community supported by NIAID's Division of Allergy, Immunology, and Transplantation.
- The Pathogen Functional Genomics Resource Center (www.niaid.nih.gov/dmid/genomes/pfgrc/) was established by NIAID, which contracted work out to the J. Craig Venter Institute, including a bioinformatics department. The Bioinformatics department works closely with the staff in all PFGRC programs including DNA microarray and gene expression, the Invitrogen Gateway® Clone Resource, Comparative Genomics, and Proteomics. Their directive is to provide software, technical support and analysis expertise to advance the production and scientific objectives of the PFGRC and the pathogen research community. Several software tools have been released to the community under open-source licenses. The implementation of a microarray annotation pipeline and internal tool streamlined the process of creating current annotation files for every microarray design. A number of comparative analysis tools have been developed to facilitate comparative genomics research.
- The Microbial Genome Sequencing Centers under the Influenza Genome Sequencing Project were established by NIAID in 2004. The goals of this project are to determine the complete genetic sequences of thousands of influenza virus strains and to rapidly provide these sequence data to the scientific community. Genomic sequences, coupled with other biochemical and microbiological information, facilitate the identification of novel and specific targets for improving strain identification and molecular genotyping; developing sequence-based detection technologies and diagnostics; developing therapeutic targets for new drugs and vaccines; comparative genomics (comparing the sequences of different strains, species, and clinical isolates) provides critical data to identify genetic polymorphisms that correlate with phenotypes such as drug resistance, virulence, and infectivity.

4.2 National Cancer Institute (NCI)

As described previously, the complexity of the constellation of diseases collectively referred to as cancer provides an obvious target to apply the power of bioinformatics. Some of NCI's initiatives in this respect are listed below:

- The NCI Center for Bioinformatics (http://ncicb.nci.nih.gov/) helps speed scientific discovery and facilitates translational research by building many types of tools and resources that enable information to be shared along the continuum from the scientific bench to the clinical bedside and back. NCICB offers critical open-source infrastructure components that others can use to develop valuable databases and software tools to meet specific research needs. NCICB's expanding suite of tools is built from these foundational components. Our projects bring tools and partners together to tackle key challenges.
- The NCI Cancer Biomedical Informatics Grid (https://cabig.nci.nih.gov/) project is an information network enabling all constituencies in the cancer community, including researchers, physicians, and patients, to share data and knowledge. Contractors proposing to work on new projects are required to conform to caBIG standards.
- NCI's Clinical Proteomic Technology Assessment for Cancer (http://proteomics.cancer.gov/) program is illustrative of the issues that need to be resolved throughout the scientific bioinformatics community. Current cancer proteomic research is hampered by a lack of standardized technologies and methodologies, which are critically needed to more effectively discover and validate proteins and peptides relevant to cancer (biomarkers). To address this critical need, the NCI established a collaborative network of five CPTAC teams in September, 2006. The CPTAC network's ultimate goal is to enable all researchers conducting cancer-related protein research at different laboratories to effectively use proteomic technologies and methodologies to directly compare and analyze their work. This should lead in turn to improved diagnostics, therapies and even prevention of cancer. The multidisciplinary CPTAC teams are conducting rigorous assessment of two major technologies currently used to analyze proteins and peptides, mass spectrometry and affinity capture platforms. Specific objectives include evaluating the performance of proteomic technology platforms and standardizing approaches to developing applications of these platforms; assessing proteomic platforms for their ability to analyze cancer-relevant proteomic changes in human clinical specimens; establishing systematic ways to standardize proteomic protocols and data analysis among different laboratories; developing and implementing uniform algorithms for sharing bioinformatics and proteomic data and analytical/data mining tools; and developing well-characterized material and bioinformatics resources for the entire cancer research community.

4.3 Miscellaneous Other Programs

Naturally, given the power of bioinformatics to advance biomedical science, many elements of the U.S. Government are beginning to fund programs to harness this power. Below are only two of these additional areas of research. In the future, many additional such programs are certain to emerge.

- The National Centers for Biomedical Computing (http://www.ncbcs.org/) is a cooperative agreement awards that are funded under the NIH Roadmap for Bioinformatics and Computational Biology and includes universities and hospitals.
- The National Institute of Biomedical Imaging and BioTechnology participates in the bioinformatics and computational biology initiative funded through the NIH common fund.

5.0 THE COMMERCIAL BIOINFORMATICS MARKET

The global bioinformatics industry has grown at a double-digit growth rate in the past and is expected to follow the same pattern in the next several years (2009 to 2012). Presently, the U.S. remains the largest market in the world, but India and China have the fastest growth rate. The biggest opportunity will be in the drug discovery sector. Bioinformatics reduces the overall drug development timeline by 30% and the annual cost by 33% due to fast development of tools and software. Given that the development lifecycle for a new drug or biologic comprises 12 to 15 years and costs approximately a billion dollars, there is significant incentive to reduce the time necessary to develop products. Major U.S. pharmaceutical companies are expected to increase their R&D expenditures in the future; a major portion of this spending is expected to go toward bioinformatics. Global pharmaceutical R&D expenditures in 2006 were $86.9B, and were expected to rise to $105.2B in 2008, a 21% increase over two years. Moreover, expenditures in 2010 are predicted to rise to $127.2B (another 21% increase in just two years). Given recent economic downturns, it is unknown if this trend will hold up.

Broadly speaking, the nascent bioinformatics industry can be categorized into four product categories: content generation and data storage (databases and data warehouses), analysis software and services, and IT infrastructure. The **largest market** at present, and the one expected to remain so for the immediate future, is the content generation market. The scope and activities associated with content generation include ongoing collection of omics-type data, advances in systems biology and synthetic biology, and development of new methodologies (such as assay techniques to advance content collection). The **fastest growing market** is analysis software and services, including artificial intelligence-based systems, new data collection platforms (biochips, etc.), specific database creation and dissemination, and development of analytical tools and algorithms. This segment is estimated to grow to over $1.2 B by 2010, an AAGR of 21.2% from $444.7 M in 2005. Specialized databases will form the major part of bioinformatics content market; the share of specialized databases in the total content market will increase from 67.6% in 2005 to over 75% by 2010 (Source: Bioinformatics: Technical Status and Market Prospects [BCC Research]). IT infrastructure is a fundamental part of the bioinformatics industry and will remain a viable market. This market segment includes data storage and retrieval, software and hardware architecture, and human/machine interfaces. While there is nothing remarkable about the IT infrastructure needed to support bioinformatics from a hardware standpoint, the human/machine interfaces will be crucially important in ensuring that the tools of bioinformatics become more established in the mainstream. A rough schematic of the interdependency of these three market segments is shown in Figure 3 below.

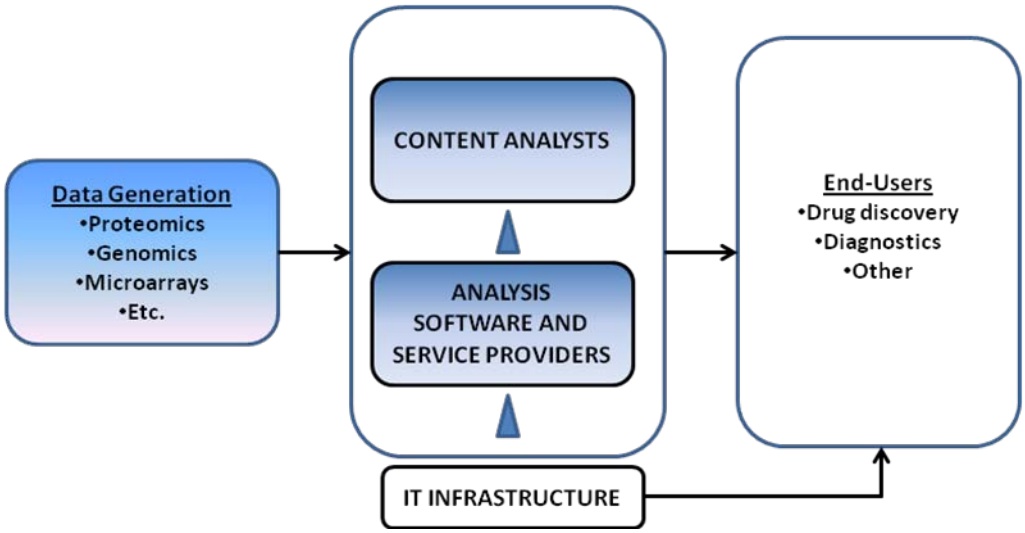

Figure 3 Structure of the Bioinformatics Market, Highly Simplified

6.0 Systems Integration: The Missing Piece to Ensure Success

Currently, the pharmaceutical and biotechnology industries are the primary drivers for growth of the commercial applications of bioinformatics. The recent decoding of the human genome has provided the key for incredible potentials in human medicine, and these industries seek to capitalize on this. As described previously in this paper, advanced biomedical tools are changing the way that pharmaceuticals are being developed. These tools generate enormous volumes of raw data which must be managed, and these data must be further refined to extract information embedded in the genetic code. Although one may logically assume that much of this information has its primary use in drug discovery, in fact the need to manage biomedical data has implications from earliest research through advanced development and licensure. We are truly living in the brave new world of bioinformatics.

Challenges currently facing the bioinformatics industry include, at minimum, the following: lack of interoperability and multiplatform capabilities; lack of standardized formats; difficulties in integrating applications; management of high-volume data; and growing competition from in-house development and publicly available tools (Source: Drug Discovery World, The Current Bioinformatics Analytical Software Landscape, Summer 2004). Many of these challenges have been described to varying degrees in the current paper.

According to a report from the Larta Institute, there are a number of potential impediments to growth of the bioinformatics industry:

- The bioinformatics market is fractured:
  - Characterized by numerous individual companies catering to the particular needs of drug developers
  - Fewer companies focused on providing integrated solutions to broader R&D requirements.
- The lack of standardized applications addressing R&D issues:
  - Limited the growth of bioinformatics and inhibited its development into a full-fledged industry
  - Lack of integration between the various players in the bioinformatics business model (software vendors, database providers, etc.)
  - Lack of integration between the internally developed systems of drug companies and technologies provided by outside vendors.
- The market isn't large enough to support the number of companies involved:
  - The market is fractured and niche-oriented such that standardization and scale associated with an industry will be difficult to establish
  - Smaller companies are unable to provide integrated application – standardization cannot happen.
- The bioinformatics market is yet to mature and create a consistent, predictable, profitable sector for itself:
  - The industry is one with relatively low barriers to entry and increasing competition from larger established IT companies
  - Only the fittest companies that address the standardization and integration issues will survive.

Although this assessment is several years old, there has been little recent progress to address the main underlying conclusion, which is that a "true" bioinformatics industry (as opposed to simply a scientific discipline) will only develop fully when the multiple issues of standardization have been addressed. Unfortunately, there is no clear pathway to integrating the various systems. Rather, in the absence of some form of integration, it is likely that multiple systems will continue to proliferate, compounding the problem. This is where the concept of a systems integrator could fill the gap. In essence, a systems integrator employs a system of systems approach to bring high-level order to very complex problems. This approach has been demonstrated to work successfully in developing certain medical countermeasures (House, 2007) and should work for bioinformatics as well.

One example of a successful integrator approach applied to bioinformatics is the NIAID Bioinformatics Integration Support Contract (BISC). This six-year, multi-million dollar project was primed by Northrop Grumman Information Technology in collaboration with the University of Texas Southwestern Medical Center, Biomind LLC and Science Commons. The goal of the BISC is to advance the discovery and generation of new hypotheses for immune-mediated diseases and to further our understanding of innate and adaptive immunity by providing an

integrated data repository, advanced computer support for handling scientific data, disseminating best practices in scientific data analysis including research data standards for data sharing and ontologies of immunology, disease phenotype and clinical research, and building a platform for integrated research and data sharing. This is being accomplished by the creation of the Immunology and Data Analysis Port (www.immport.org/immportWeb/home/home.do). This data repository houses a variety of data types in the areas of both clinical and basic research, and this integrated resource facilitates the translation of mechanistic data generated from the bench to improve public health and treatment of immune-mediated diseases. It is conceivable that this type of collaboration, on a large scale, would be an excellent platform for laterally integrating many other types of bioinformatic data.

7.0 CLOSING THOUGHTS

The powerful tools generically called bioinformatics will be vitally important to help us make sense of the increasing torrent of biological data now being generated in laboratories around the world. These data, when properly stored, organized, manipulated and decoded, promise to not only dramatically accelerate biomedical sciences in general, but to finally usher in an era of personalized medicine—the specific tailoring of clinical practice by use of basic human biology. The most daunting challenge will be in standardizing the techniques for handling these data so that a true standard of practice may evolve.

8.0 REFERENCES CITED

Allen, J.E., Gardner, S.N. and Slezak, T.R. (2008). DNA signatures for detecting genetic engineering in bacteria. Genome Biol. 9(3):R56.

Almaas, E. (2007). Biological impacts and context of network theory. J Exp Biol. 210(Pt 9):1548-1558.

Baudis, M. (2006). Online database and bioinformatics toolbox to support data mining in cancer cytogenetics. Biotechniques. 40(3):269-27

Bernas, T., Grégori, G., Asem, E.K. and Robinson, J.P. (2006). Integrating cytomics and proteomics. Mol Cell Proteomics. 5(1):2-13.

Biswas, S., Raoult, D. and Rolain, J.M. (2008). A bioinformatic approach to understanding antibiotic resistance in intracellular bacteria through whole genome analysis. Int J Antimicrob Agents. 32(3):207-220.

Capecchi, B., Serruto, D., Adu-Bobie, J., Rappuoli, R. and Pizza, M. (2004). The genome revolution in vaccine research. Curr Issues Mol Biol. 6(1):17-27.

Carrera, M. and Sagripanti, J.L. (2009). Artificial plasmid engineered to simulate multiple biological threat agents. Appl Microbiol Biotechnol. 81(6):1129-1139.

Chen, J.M. and Ho, C.M. (2006). Path to bio-nano-information fusion. Ann N Y Acad Sci. 1093:123-142.

Chorley, B.N., Wang, X., Campbell, M.R., Pittman, G.S., Noureddine, M.A. and Bell, D.A. (2008). Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. Mutat Res. 659(1-2):147-157.

Christen, R. (2008). Identifications of pathogens - a bioinformatic point of view. Curr Opin Biotechnol. 19(3):266-273.

Collins, C.D., Purohit, S., Podolsky, R.H., Zhao, H.S., Schatz, D., Eckenrode, S.E., Yang, P., Hopkins, D., Muir, A., Hoffman, M., McIndoe, R.A., Rewers, M. and She, J.X. (2006). The application of genomic and proteomic technologies in predictive, preventive and personalized medicine. Vascul Pharmacol. 45(5):258-267.

Davies, M.N. and Flower, D.R. (2007). Harnessing bioinformatics to discover new vaccines. Drug Discov Today. 12(9-10):389-395.

De Groot , A.S. and Martin, W. (2003). From immunome to vaccine: epitope mapping and vaccine design tools. Novartis Found Symp. 254:57-72.

De Groot, A.S., McMurry, J. and Moise, L. (2008). Prediction of immunogenicity: in silico paradigms, ex vivo and in vivo correlates. Curr Opin Pharmacol. 8(5):620-626.

Dietel, M. and Sers, C. (2006). Personalized medicine and development of targeted therapies: the upcoming challenge for diagnostic molecular pathology. A review. Virchows Arch. 448: 744-755.

Dopazo, J. (2006). Bioinformatics and cancer: an essential alliance. Clin Transl Oncol. 8(6):409-415.

Drake, R.R., Deng, Y., Schwegler, E.E. and Gravenstein, S. (2005). Proteomics for biodefense applications: progress and opportunities. Expert Rev Proteomics. 2(2):203-213.

Edwards, S.W. and Preston, R.J. (2008). Systems biology and mode of action based risk assessment. Toxicol Sci. 106(2):312-318.

Evans, B.J. (2007). Distinguishing product and practice regulation in personalized medicine. Clin Pharmacol Ther. 81(2):288-293.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 37-54.

Fitzgerald, G.A. (2005). Opinion: anticipating change in drug development: the emerging era of translational medicine and therapeutics. Nat Rev Drug Discov. 4(10):815-818.

Foekens, J.A., Wang, Y., Martens, J.W., Berns, E.M. and Klijn, J.G. (2008). The use of genomic tools for the molecular understanding of breast cancer and to guide personalized medicine. Drug Discov Today. 13(11-12):481-487.

Garcia-Sastre, A. and Whitley, R.J. (2006). Lessons learned from reconstructing the 1918 influenza pandemic. J Infect Dis. 194 Suppl 2:S127-132.

Gibson, D.G., Benders, G.A., Axelrod, K.C., Zaveri, J., Algire, M.A., Moodie, M., Montague, M.G., Venter, J.C., Smith, H.O. and Hutchison, C.A. 3rd. (2008). One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic Mycoplasma genitalium genome. Proc Natl Acad Sci U S A. 105(51):20404-20409.

Glasner, J.D., Plunkett, G. 3rd, Anderson, B.D. et al. (2008). Enteropathogen Resource Integration Center (ERIC): bioinformatics support for research on biodefense-relevant enterobacteria. Nucleic Acids Res. 36(Database issue):D519-523.
Gordijn, B. (2006). Converging NBIC technologies for improving human performance: a critical assessment of the novelty and the prospects of the project. J Law Med Ethics. 34(4):726-732.
Greene, J.M., Collins, F., Lefkowitz, E.J., Roos, D., Scheuermann, R.H., Sobral, B., Stevens, R., White, O. and Di Francesco, V. (2007). National Institute of Allergy and Infectious Diseases bioinformatics resource centers: new assets for pathogen informatics. Infect Immun. 75(7):3212-3219.
Haoudi A. and Bensmail, H. (2006). Bioinformatics and data mining in proteomics. Expert Rev Proteomics. 3(3):333-343.
Hari, K.L., Goates, A.T., Jain, R., Towers, A., Harpin, V.S., Robertson, J.M., Wilson, M.R., Samant, V.S., Ecker, D.J., McNeil, J.A., and Budowle, B. (2008). The Microbial Rosetta Stone: a database system for tracking infectious microorganisms. Int J Legal Med. 123(1):65-69.
Harrill, A.H. and Rusyn, I. (2008). Systems biology and functional genomics approaches for the identification of cellular responses to drug toxicity. Expert Opin Drug Metab Toxicol. 4(11):1379-1389.
Holzmuller, P., Grébaut, P., Brizard, J.P., Berthier, D., Chantal, I., Bossard, G., Bucheton, B., Vezilier, F., Chuchana, P., Bras-Gonçalves, R., Lemesre, J.L., Vincendeau, P., Cuny, G., Frutos, R. and Biron, D.G. (2008). Pathogeno-proteomics". Ann N Y Acad Sci. 1149:66-70.
House, R.V. (2007). Systems integration: an effective and innovative response to emerging biological threats. Vaccine. 25(16):3170-3174.
Hutter, M.C. (2009). In silico prediction of drug properties. Curr Med Chem. 16(2):189-202.
Jain, K.K. (2007). Cancer biomarkers: current issues and future directions. Curr Opin Mol Ther. 9(6):563-571.
Janssens, A.C. and van Duijn, C.M. (2008). Genome-based prediction of common diseases: advances and prospects. Hum Mol Genet. 17(R2):R166-173.
Kaushik, D.K. and Sehgal, D. (2008). Developing antibacterial vaccines in genomics and proteomics era. Scand J Immunol. 67(6):544-552.
Khan, A.S., Mujer, C.V., Alefantis, T.G., Connolly, J.P., Mayr, U.B., Walcher, P., Lubitz, W. and Delvecchio, V.G. (2006). Proteomics and bioinformatics strategies to design countermeasures against infectious threat agents. J Chem Inf Model. 46(1):111-115.
Khardori, N. (2006). Bioterrorism and bioterrorism preparedness: historical perspective and overview. Infect Dis Clin North Am. 20(2):179-211, vii.
Konstantinopoulos, P.A., Karamouzis, M.V. and Papavassiliou, A.G. (2008). Educational and social-ethical issues in the pursuit of molecular medicine. Mol Med. 15(1-2):60-63.
Littman, B.H., Di Mario, L., Plebani, M. and Marincola, F.M. (2007). What's next in translational medicine? Clin Sci (Lond). 112(4):217-227.
Loging, W., Harland, L. and Williams-Jones, B. (2007). High-throughput electronic biology: mining information for drug discovery. Nat Rev Drug Discov. 6(3):220-230.
Manning, A.T., Garvin, J.T., Shahbazi, R.I., Miller, N., McNeill, R.E. and Kerin, M.J. (2007). Molecular profiling techniques and bioinformatics in cancer research. Eur J Surg Oncol. 33(3):255-265.
Mooney, S.D. and Baenziger, P.H. (2008). Extensible open source content management systems and frameworks: a solution for many needs of a bioinformatics group. Brief Bioinform. 9(1):69-74.
Moore, J.H. (2007). Bioinformatics. J Cell Physiol. 213(2):365-369.
Mora, M., Donati, C., Medini, D., Covacci, A.and Rappuoli, R. (2006). Microbial genomes and vaccine design: refinements to the classical reverse vaccinology approach. Curr Opin Microbiol. 9(5):532-536.
Morens, D.M. and Fauci, A.S. (2007). The 1918 influenza pandemic: insights for the 21st century. J Infect Dis. 195(7):1018-1028.
Muster, W., Breidenbach, A., Fischer, H., Kirchner, S., Müller, L. and Pähler, A. (2008). Computational toxicology in drug development. Drug Discov Today. 13(7-8):303-310.
Nebert, D.W., Zhang, G. and Vesell, E.S. (2008). From human genetics and genomics to pharmacogenetics and pharmacogenomics: past lessons, future directions. Drug Metab Rev. 40(2):187-224.
Nicholson, J.K., Holmes, E., Lindon, J.C. and Wilson, I.D. (2004). The challenges of modeling mammalian biocomplexity. Nat Biotechnol. 22(10):1268-1274.
Ozdemir, V., Williams-Jones, B., Glatt, S.J., Tsuang, M.T., Lohr, J.B. and Reist, C. (2006). Shifting emphasis from pharmacogenomics to theragnostics. Nat Biotechnol. 24(8):942-946.
Pene, F., Courtine, E., Cariou, A. and Mira, J.P. (2009). Toward theragnostics. Crit Care Med. 37(1 Suppl):S50-S58.

Phan, J.H., Quo, C.F. and Wang, M.D. (2006). Functional genomics and proteomics in the clinical neurosciences: data mining and bioinformatics.
Prog Brain Res. 158:83-108.
Phillips, K.A., Liang, S.Y., Van Bebber, S.; Canpers Research Group. (2008). Challenges to the translation of genomic information into clinical practice and health policy: Utilization, preferences and economic value. Curr Opin Mol Ther. 10(3):260-266.
Poland, G.A., Ovsyannokova, I.G., Jacobson, R.M. and Smith, D.I. (2007). Heterogeneity in vaccine immune response: the role of immunogenetics and the emerging field of vaccinomics. Clin. Pharmacol. Ther. 82:653-664.
Rivenbark, A.G. and Coleman, W.B. (2007). Dissecting the molecular mechanisms of cancer through bioinformatics-based experimental approaches. J Cell Biochem. 101(5):1074-1086.
Scarselli, M., Giuliani, M.M., Adu-Bobie, J., Pizza, M. and Rappuoli, R. (2005). The impact of genomics on vaccine design. Trends Biotech. 23:84-91.
Schnackenberg, L.K. (2007). Global metabolic profiling and its role in systems biology to advance personalized medicine in the 21st century. Expert Rev Mol Diagn. 7(3):247-259.
Shabo, A. (2008). Integrating genomics into clinical practice: standards and regulatory challenges. Curr Opin Mol Ther. 10(3):267-272.
Shulaev, V. (2006). Metabolomics technology and bioinformatics. Brief Bioinform. 7(2):128-139.
Sistare, F.D. and Degeorge, J.J. (2008). Applications of toxicogenomics to nonclinical drug development: regulatory science considerations. Methods Mol Biol. 460:239-261.
Solomon, L.M. and Sieczkiewicz, G.J. (2007). Impact of the US Patent System on the promise of personalized medicine. Gend Med. 4(3):187-192.
Solorzano, A., Song, H., Hickman, D. and Pérez, D.R. (2007). Pandemic influenza: preventing the emergence of novel strains and countermeasures to ameliorate its effects. Infect Disord Drug Targets. 7(4):304-317.
Squires, B., Macken, C., Garcia-Sastre, A., Godbole, S., Noronha, J., Hunt, V., Chang, R., Larsen, C.N., Klem, E., Biersack, K. and Scheuermann, R.H. (2007). BioHealthBase: informatics support in the elucidation of influenza virus host pathogen interactions and virulence. Nucleic Acids Res. 36(Database issue):D497-503.
Stajich, J.E. and Lapp, H. (2006). Open source tools and toolkits for bioinformatics: significance, and where are we? Brief Bioinform. 7(3):287-296.
Tanaka, H. (2008). Bioinformatics and genomics for opening new perspective for personalized care. Stud Health Technol Inform. 134:47-58.
Van Brabant, B., Gray, T., Verslyppe, B., Kyrpides, N., Dietrich, K., Glöckner, F.O., Cole, J., Farris, R., Schriml, L.M., De Vos, P., De Baets, B., Field, D. and Dawyndt, P. (2008). Genomic Standards Consortium. Laying the foundation for a Genomic Rosetta Stone: creating information hubs through the use of consensus identifiers. OMICS. 12(2):123-127.
Wingren, C. and Borrebaeck, C.A. (2009). Antibody-based microarrays. Methods Mol Biol. 509:57-84.
Yan, Q. (2008). The integration of personalized and systems medicine: bioinformatics support for pharmacogenomics and drug discovery. Methods Mol Biol. 448:1-19.
Yang, J.Y., Yang, M.Q., Arabnia, H.R. and Deng, Y. (2008). Genomics, molecular imaging, bioinformatics, and bio-nano-info integration are synergistic components of translational medicine and personalized healthcare research. BMC Genomics. 16;9 Suppl 2:I1.
Zhang, C., Crasta, O., Cammer, S., Will, R., Kenyon, R., Sullivan, D., Yu, Q., Sun, W., Jha, R., Liu, D., Xue, T., Zhang, Y., Moore, M., McGarvey, P., Huang, H., Chen, Y., Zhang, J., Mazumder, R., Wu, C. and Sobral, B. (2008). An emerging cyberinfrastructure for biodefense pathogen and pathogen-host data. Nucleic Acids Res.36:D884-D891.
Ziegler, A.S. (2006). Regulation: threat to converging technologies. Ann N Y Acad Sci. 1093:339-349.

9.0 APPENDIX 1: COMPANIES INVOLVED IN BIOINFORMATICS

Table 2 Companies Involved in Bioinformatics

Note: Underlined text in this table is hyperlinked to the corresponding Web site.

Company	Web link	Notes
Accelrys	www.accelrys.com	The leader in simulation and informatics software for the pharmaceutical and chemicals process industries (formerly MSI).
Affymetrix	www.affymetrix.com	
Agilent	www.agilent.com	
Applied Biosystems	www.appliedbiosystems.com	
Astellas	www.astellas.com	
AstraZeneca	www.astrazeneca.com	
Beckman Coulter	www.beckmancoulter.com	
BioChem Pharma	www.biochemgroup.com	
BioDiscovery	www.biodiscovery.com	An established company providing comprehensive software solutions exclusively for gene expression research bioinformatics.
Bioinformatics Solutions		
Biomax Informatics	www.biomax.com	Provides computational solutions for decision making and knowledge management in the life science industry.
Bio-Rad	www.biorad.com	
Capgemini	www.capgemini.com	
CeleraScience	www.celera.com	Provides genomic data and analysis tools based on a robust delivery.
CLC bio	www.clcbio.com	Bioinformatics software and consulting for DNA, RNA and protein analysis, focusing on custom data analysis and specialized bioinformatics algorithms.
Cogenix	www.cogenix.com	
Compaq	www.compaq.com	
Compugen	www.compugen.com	
Confirmant		
CuraGen	www.curagen.com	
diaDexus	www.diadexus.com	Has utilized genomics and bioinformatics to identify thousands of disease-associated molecular targets. San Francisco, CA.

Table 2 Companies Involved in Bioinformatics (Continued)

Company	Web link	Notes
DNAStar	www.dnastar.com	
Double Twist	www.doubletwist.com	
DuPont	www.dupont.com	
Eli Lilly	www.lilly.com	
Entigen Corporation		
EraGen Biosciences	www.eragen.com	Has patents and tools in interpretive proteomics, DNA diagnostics, and pharmacogenomics, including MasterCatalog™ proteomics data mining platform. Madison, WI.
Exelixis	www.exelixis.com	
Expression Analysis	www.expressionanalysis.com	Providing total Affymetrix GeneChip® microarray processing and analysis.
Fujitsu	www.fujitsu.com	
GE Healthcare Bio-Sciences	www.gehealthcare.com	
Gene Codes	www.genecodesforensics.com	
Gene Network Sciences	www.gnsbiotech.com	Creates dynamic computer models of living cells and next generation data-mining tools for pharmaceutical and biotechnology companies.
Genelogic	www.genelogic.com	Offering gene expression databases derived from clinically important tissues.
Genetics Squared	www.genetics2.com	Offering genetic programming-based tools for Bioinformatics.
Genomatix	www.genomatix.de	Offering PromoterInspector on-line tool for prediction of mammalian promoters; MatInspector for for the identification of transcription factor binding sites in genomic DNA; and other tools.
Genome Therapeutics	www.genomecorp.com	Focused for commercialization of genomics-based pharmaceutical and diagnostic products.
Genomica		
Genomining	www.genomining.com	Specializing in discovery, interpretation and management of data in biology, with strong expertise in data-mining, and access to large databases.
Genomix Corporation		

Table 2 Companies Involved in Bioinformatics (Continued)

Company	Web link	Notes
Google	www.google.com	
Hewlett-Packard	www.hp.com	
Hitachi	www.hitachi.com	
Human Genome Sciences	www.hgsi.com	Pioneer in the use of genomics, the study of human genes, and the development of new pharmaceutical products. Has 6 drugs in human clinical trials. Headquarters: Baltimore, MD, USA.
Hybrigenics	www.hybrigenics.com	
IBM	www.ibm.com	
Incyte Pharmaceuticals	www.incyte.com	Provides genomic databases, bioanalysis software, biological reagents, and microarray services.
IDBS	www.idbs.com	
In Silico Discovery	www.insilicodiscovery.com	
Informax		
Invitrogen	www.invitrogen.com	
Johnson & Johnson	www.jnj.com	
Kiran		
Language and Computing	www.landcglobal.com	Offers information analysis, document mining, information retrieval and extraction, and terminology management solutions to healthcare and pharmaceutical companies.
Media Cybernetics	www.mediacy.com	
Merck & Co	www.merck.com	
Metalife	www.metalife.de	Offering many tools for automation of functional analysis of biological data, including visualization and text mining of scientific literature.
Millennium Pharmaceuticals	www.mlnm.com	
Motorola	www.motorola.com	
Myriad Genetics	www.myriad.com	
NetGenics		
NextGen Sciences	www.nextgensciences.com	Developing technology platforms for genomics, transcriptomics, and proteomics.
Novartis	www.novartis.com	

Table 2 Companies Involved in Bioinformatics (Continued)

Company	Web link	Notes
NovaScreen BioSciences		
Novo Nordisk	www.novonordisk.com	
Ocimum Biosolution	www.ocimumbio.com	Offering lab information and knowledge management systems, genomics, proteomics, bioinformatics and custom contract services (spin-off of Gene Logic in Gaithersburg, MD; India and USA).
OGS		
Oracle	www.oracle.com	
Pfizer	www.pfizer.com	
PharmaDM	www.pharmadm.com	A global enabler of Discovery Informatics, the combination of bio-, chemo- and clinical informatics.
PREMIER Biosoft	www.premierbiosoft.com	Molecular biology software for PCR, real-time PCR, microarray design, glycan mass fingerprinting and molecular cloning.
Protein Lounge	www.proteinlounge.com	
Reel Two	www.reeltwo.com	Providing text and data mining solutions for pharmaceutical and biotech companies.
Roche Applied Science	www.roche-applied-science.com	
Rosetta Inpharmatics	www.rii.com	
Sanofi-Pasteur	www.sanofi-pasteur.com	
Sigma-Aldrich	www.sigma-aldrich.com	
Solexa		
Spotfire	http://spotfire.tibco.com	
Structural Bioinformatics	http://www.sbg.bio.ic.ac.uk	
Sun Microsystems	www.sun.com	
The Boston Consulting Group	www.bcg.com	
Viaken Systems	www.viaken.com	
Wyeth	www.wyeth.com	
Yamanouchi		

Sources:
http://www.kdnuggets.com/companies/bioinformatics.html#W
http://www.marketresearch.com/product/display.asp?productid=1391621&g=1

10.0 APPENDIX 2: UNIVERSITIES WITH BIOINFORMATICS PROGRAMS

Table 3 Universities with Bioinformatics Programs

Universities with Bioinformatics Programs	
University of California (UCLA, UCSC, UCSI, USD)	Ohio State University
University of Chicago	University of Pennsylvania
Columbia University	Purdue University
George Mason University	University of Pittsburgh
Harvard University	Stanford University
University of Illinois	Towson University
Johns Hopkins University	Vanderbilt
McGill University	Virginia Tech
University of Michigan	University of Wisconsin

Sources:
http://www.amia.org/meetings/stb08/panels.asp
http://www.bioitcoalition.org/