A Privacy Concern: Bioinformatics and Storing Biodata

D'Nae Ferguson
Department of Computer Science
Hampton University
Hampton, VA

***Abstract* – Security and privacy, regardless of the instance, are preponderating topics for most organizations. Bioinformatics and the study of computational biology are no exception. The premise of this report is to discuss the many different privacy concerns as it pertains to the field of bioinformatics, as well as the usage and storage of personal biodata. With the varying threats that target average users of technology, is the capability and infrastructure currently in place to protect users against a leakage or breach in personal data? This study discusses the different concerns surrounding the field of bioinformatics, how the data and personal information is currently stored, and will make recommendations on how to mitigate the risks associated with the usage and storage of personal biodata. This study includes interviews from bioinformaticians and industry professionals, a survey of adults who have the potential for impact, and current legislature that exists to address personal data protection.**

***Index terms*—bioinformatics, biodata, computational biology, data mining, data storage, DNA database, DNA profiling, genetic privacy, Personal Data Protection Act 2010, privacy**

I. INTRODUCTION

Human development throughout history has largely been associated with the development of their expansive technology. Technology has seen many iterations, from being simple hunting tools to powerful high-power computing machinery today. Among the powerful advances in technology is the study of bioinformatics and computational biology. These special topics within the computer science domain, fuse biological data, genomics, and cyber-technology together to create the field of bioinformatics. In short, bioinformatics is defined as the sum of the computational approaches to analyze, manage, and store biological data. Bioinformatics involves the analysis of biological information using computers and statistical techniques, the science of developing and utilizing computer databases and algorithms to accelerate and enhance biological research. Bioinformatics is also used in analyzing genomes, proteomes (protein sequences), three-dimensional modeling of biomolecules and biologic systems, etc. Traditionally, training in informatics requires backgrounds in molecular biology and computer science, including database design and analytical approaches. This study is an examination of the privacy concerns currently affecting the storage of patient biodata and the bioinformatics community.

A. Problem Statement

Technology is constantly developing, and as such, the world is exposed to a surge of new technologies. As more companies begin to store patient biodata, and utilize it for research, attackers have the potential to sequester this information and exploit patient biodata.

II. METHODOLOGY

The methodology for responding to the aforementioned problem statement will consist of various strategies. The first methodology form is research gathering attained through literature review. Next, research was attained through interviews with industry professionals, researchers and major contributors to the field of bioinformatics. Lastly, research was conducted

through current legislature that protects patient biodata and personal information. Each stage of our methodology is explained as follows:

A. Literature Review

This method of analysis will discuss the concerns pertaining to the field of bioinformatics, including ethical, privacy and security concerns, associated with bioinformatics and the storage of patient biodata. This form of research will be conducted through the use and synthesis of research reports authored by professional researchers and experts within the field of bioinformatics, as well as scholarly articles. Through reading these various reports and articles, this will act as a foundation and support for the other methods of research conducted within this study. References to news articles will also be used in order to relate our research findings to current events related to bioinformatics and privacy.

B. Interviews

This study collects data about the topic by interviewing several individuals who are proficient in the field of computational biology, bioinformatics, and genomics. The interview process will involve an assortment of experts, researchers and professors, and during the interviews, they will share their personal knowledge and experience on bioinformatics, as well as the current privacy concerns and where they believe to be trending in the near future. Notable intended interviewees are Lior Pachter, Wolfgang Huber, and Serafim Batzoglou. Dr. Lior Pachter is a computational biologist. He currently works at the California Institute of Technology, where he is the Bren Professor of Computational Biology. His research primarily lies within the domains of genomics, combinatorics, computational geometry, machine learning, and scientific computing.

Dr. Wolfgang Huber studied physics at the University of Freiburg, obtained a Ph.D. in theoretical physics on stochastic models and simulation of open quantum systems. He moved to California in 1998 to do postdoctoral research in cheminformatics of small, drug-like compounds at IBM Research Almaden in San José. In 2000, his interest in cancer genomics and microarray analysis led him to the German Cancer Research Centre (DKFZ) in Heidelberg. In 2004, he joined EMBL to start a research group at its European Bioinformatics Institute (EBI) in Cambridge. In 2009, he took up a position in the newly formed Genome Biology unit of EMBL in Heidelberg, and in 2011 became EMBL Senior Scientist.

Lastly, Serafim Batzoglou is the Chief Data Officer at Insitro. He Was Vice President of computational genomics at Illumina, and professor of computer science at Stanford University between 2001 and 2016. His lab focused on computational genomics with special interest in developing algorithms, machine learning methods, and systems for the analysis of large-scale genomic data. He has also been involved with the Human Genome Project and ENCODE.

C. Current Legislature

This study locates and synthesizes the current legislature that exists surrounding the usage and storage of patient biodata, and limitations or 'checks' within the bioinformatics and computational genomics field. Current legislature is examined from sources both within the United States Federal Government system, in addition to legislature found in the systems of other countries across the world.

III. RESULTS

Concerns regarding bioinformatics and genomic data fall into three separate categories based on data achieved through the methodology in section II: (1) ethical concerns surrounding the use of bioinformatics, (2) concerns regarding health information held by individual organizations, and (3) concerns about the systemic flow of information throughout the healthcare and related industries.

A. Bioinformatics and Ethical Concerns

Many of the novel, cutting-edge ideas are met with scrutiny, and the topics of bioinformatics and computational genomics are no exception. While some believe that the usage and study of bioinformatics and computational genomics is not unethical due to the potential health benefits; others find these studies threatening and invasive to an individual's rights and provides a lack of privacy [6]. The concept of "bioethics" was first developed to handle the application of moral philosophy within medical dilemmas. It emerged out of a need to reflect philosophically on the current issues affecting modern medicine. Computer usage and the spread of internet technologies has impacted the lives of many individuals globally, and continues to alter societies in a similar way that modern medicine has, through its expansion. Biotechnology in conjunction with the usage of computer technology has the ability to impact many aspects of both the physical and social life which often lead to concerns regarding the ethics and security of the machines and these processes [12].

Data mining has the capability to distinguish an individual from a group and identify groups with common characteristics through arranging similar or shared qualities and properties. This type of classification or profiling raises some ethical concerns because it is reliant on utilizing characteristics that can identify individuals and sometimes may be incorrect. In addition, bioinformatics and computational genomics can often determine distinctive facts about individuals and/or groups which makes them liable. For instance, one's personal biodata can be utilized in making decisions or judgements about individuals—these judgements may result in one being denied employment or insurance. Further, data collected in bioinformatics and computational genomic studies is the direct result of educated assent and later receiving consent from human subjects interested in the studies. Assent simply refers to willingness of the participants to participate in the research, and also refers to the agreement of those who cannot give their consent to participate in the study. While consent refers to permission for something to happen or agreement to do something through legal binds. Thus, such data may fail to meet the required conditions for a substantial educated consent due to participants shielding themselves from vulnerability [13].

Regarding privacy, bioinformatics and computational genomics raised a crop of ethical concerns. The ability exists for a person to be identified though his or her genetic data residing within a bioinformatics computer system. This privacy concern could lead to the potential exposure of sensitive medical information or other materials that could be used to harm an individual in the event of a data breach.

The methodology used to conduct bioinformatics and computational genomics research may also be affected by ethical issues, specifically surrounding the consequences regarding the information clinicians deliver to their patients. The contexts of this varies depending on the type of studies conducted. Different study designs often result in different ethical dilemmas. The usage of varied types of biological samples from DNA genotyping to proteomics, may provide results with different consequences to individuals and the population [5].

Another salient issue that arises with the usage of bioinformatics and computational genomics is centered around the ownership and intellectual property of genetic data. Given that most participants in such genetic studies donate samples of their DNA to submit in databases, it is unclear whether or not there is a complete forfeiture of rights regarding the use of the patient's genetic data. Concretely, it is unclear whether the databases have complete control and ownership of the data. Since the study of bioinformatics and computational genomics are relatively new fields, there is a lack of legislature that safeguards patient data. Thus, it is unclear whether the federal government has a position on who the true owner of such genetic data is. Few laws have been enacted to protect the autonomy and privacy of participants in genetic research studies, however they fail to mention who owns the genetic data stored within the databases [13].

The Human Genome Project was the global, collaborative research effort, with the goal of obtaining a complete map and to better understand all of the genes belonging to human beings. The amalgamation of all human genes together, is referred to as the genome. The Human Genome Project researchers were able to decipher the genome in three notable ways—creating maps that mark the locations of genes for prominent sections of all chromosomes; determining the order, or "sequence," of all the bases in our genome's DNA; and producing linkage maps, which indicates which inherited traits (such as hereditary diseases) can be tracked over many generations. The Human Genome Project was revolutionary in the field of medicine and technology. However, ethical concerns arose revolving the privacy and confidentiality of the genetic information, psychological impact, and philosophical debate. Due to government sponsorship of databanks and the supplement to medical research companies, many were concerned with the privacy of their genetic information. When discussing the psychological impact, this refers to the mistrust experienced in reference to race or economic status. For example, many African Americans are mistrustful of the healthcare and medical research studies given the past traumas experienced by slaves and poor blacks in America. Lastly, the philosophical debate is in reference to the common view that many have, relating genetic modification to 'playing God' in any capacity, and whether these actions are considered morally sound [7].

Ethical concerns regarding computational genomics and bioinformatics does not solely impact humans. In fact, there are ethical concerns with regard to animal genomics, and plant genomics. Advocates for animal rights often argue that all species—not just homo sapiens—deserve the inherent, natural right to be void of genetic altering or manipulation in any capacity. Thus, we see a shift toward more cruelty-free products. With regard to plant genomics, ethical concerns are

centered around the "naturalness" of plants. The debate surrounding the safety of Genetically Modified Organisms (GMOs) has ensued for some time, and many question the safety of these organisms for human consumption.

Bioinformatics is a field that is increasing in popularity and the effects of the work and research has lasting impact on individuals and humanity. For example, the Human Genome Project was a revolutionary study and has enabled a collective understanding of the human genome. In fact, the finishing phase yielded 99% of the human genome in its final form. This form contained 2.85 billion nucleotides, with a predicted error rate of 1 event per 1000,000 bases sequenced in the human genome. Despite these feats in human development, the biomedical and bioinformatics community is still largely at risk.

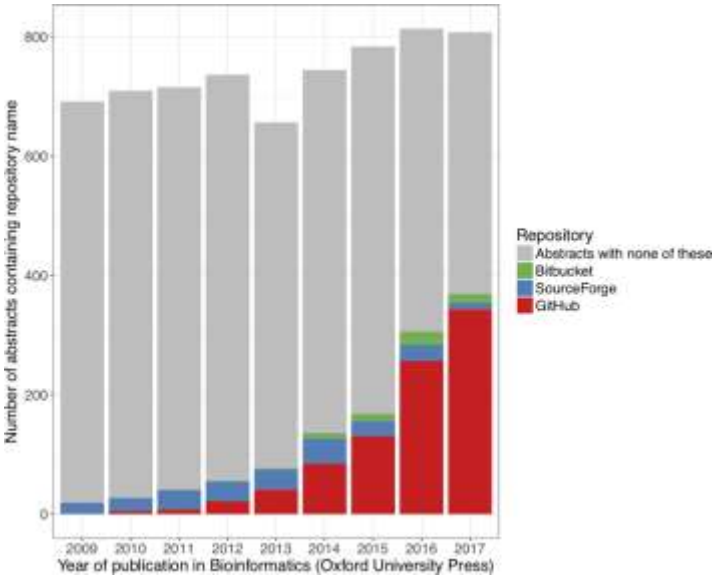

Figure 1: Source code repositories in the journal 'Bioinformatics' between the years 2009 and 2017

Here the term "repository" refers to online code hosting services. The journal *Bioinformatics* publishes new developments in bioinformatics and computational biology. If a paper focuses on software development, authors are required to state software availability in the abstract, including the complete URL. URLs for software hosted on the popular services such as GitHub, Bitbucket, and SourceForge contain the respective repository name except in rare cases of developers referring to the repository from a different URL or page. The figure shows the results of PubMed searches for the repository names in the title or abstract of papers published in *Bioinformatics* between 2009 and 2017. The category "Abstracts with none of these" captures all remaining articles published in *Bioinformatics* for the year, and likely includes many software projects hosted on organization websites or featuring their own domain name, as well as any articles that did not publish software. This is a testament to the staggering increase in popularity among the field and research of bioinformatics.

In the midst of a pandemic, researcher Bob Diachenko discovered a database owned by a medical software company leaking the personal and private details of over 3.1 million patients. This database was left exposed online without the need for a password or other forms of authorization. This 'leaky' database was owned by vendor Adit—a developer of online booking and patient management software for both medical and dental practices. The search engine BinaryEdge indexed the unsecured database on July 12, 2020, which was discovered by Diachenko on the following day. Despite email attempts between Diachenko and Adit about the findings, the company failed to return his emails. The database contained full patient names, email addresses, contact information, sex, marital status, and practice names. This information put patients at greater risk because cybercriminals can utilize this information to launch targeted phishing attacks to gain more information for later fraud or to scam patients. Even more alarming, the data was then destroyed ten days later, on July 22, 2020 and was potentially stolen by a malicious bot known as 'meow bot'.

B. Concerns Regarding Health Information Held by Individual Organizations

While there a multitude of ethical concerns that affect the bioinformatics and computational genomics studies, there is also a concern regarding health information and patient biodata that is held by individual organizations. External agents often seek to violate the direct security and confidentiality policies of a specific organization which makes the storage of electronic health records at individual organizations vulnerable. Conversely, internal agents are comprised of authorized system users who abuse their privileges by gaining access to information for inappropriate uses or reasons—whether it is to view records of family members, friends, coworkers, or neighbors or to spread information for malicious intents. External agents are comprised of those outside the network, who are unauthorized to utilize the Information Technology (IT) system or access its information. However, these individuals still attempt to gain access or manipulate the data to render these systems inoperable. Hospitals and other healthcare organizations have ventured to

counter external agents in an effort to protect patient's paper health records. Healthcare administrators and receptionists have less experience in protecting patient data and the network against technical attacks from external agents. In fact, until recently, many hospitals and healthcare organizations were connected to a network that was accessible to the public [13].

As it stands, there exists little evidence to accurately determine the vulnerability and provide a threat assessment of electronic health information and patient biodata to external attacks. As a part of the research process, most of the sites visited, reported no cases in which damaging intrusions by an external agent were detected. However, hospitals and the healthcare industry currently have no mechanism for reporting incidents and intrusion detections. Nonetheless, there is evidence and a history of computer break-ins that have occurred within this industry. In one incident, the self-proclaimed "414" group intruded a machine at the National Cancer Institute in 1982, even though no damage was detected as a result of the break-in. The "414s" were a group of teenage hackers that broke into high-profile systems, most notably in the years 1982 and 1983. Concerns regarding technical attacks by external agents and safety are increasing in a multitude of other industry sectors, and even the government. Providing commentary on a recent study conducted by the Federal Bureau of Investigation and the Computer Security Institute, Director Patrick Rupalis stated, "The information age has already arrived, but most organizations are woefully unprepared . . . [making] it easier for perpetrators to steal, spy, or sabotage without being noticed and with little culpability if they are." Resulting in the surveying of different sites, it was found that 42 percent of the sites had experienced and intrusion or unauthorized access and usage within the past year—nearly half of the sites surveyed, 20 percent of the participants were unaware of the intrusion, only 17 percent of those who suffered an attack notified authorities and in fact, most organizations did not have a written policy in the event of network intrusions. A current estimate conducted by the Defense Information Systems Agency indicates that Pentagon networks and machines suffered over 250,000 attacks by intruders in 1995. In fact, this figure continues to double every year, and in roughly 67 percent of these attacks, threat actors were able to gain entry to the computer network. Lastly, a RAND Corporation study on information war scenarios would suggest that terrorists using malware and hacker technologies would be increasingly detrimental to computer-based systems. Thus, undermining the efforts of 911 emergency telephone services, banking and securities systems, information broadcast and news channels, electric power distribution networks, train and rideshare services, pipeline and septic systems, as well as other parts of the information infrastructure.

While they do not identify and describe the exact threats posed to healthcare organizations, the research indicates an increasing vulnerability to information technology systems, especially ones that are connected to public infrastructure, such as the internet. Thus, this research suggests that the increased desire to use electronic health information linked through modern networking technologies, could induce exposure to sensitive health information to an assortment of external and internal threats, which will require adequate addressing.

C. Systemic Concerns Regarding Health Information

This study has explored ethical issues regarding bioinformatics, concerns regarding the storage of patient biodata, and now explores the systemic concerns regarding health information. Systemic concerns regarding the privacy of patient specific health information are largely based on the usage of such information in a manner that acts against the interests of the individual or patient involved. These interests can vary from identifiable inauspicious consequences—an increase in difficulty obtaining insurance or employment—to the less noticeable ones—personal embarrassment or discomfort. To better understand public concerns about the usage of patient biodata, it is important to first examine the current exchanges of patient data and health information throughout the healthcare system.

Health information, both in its paper and electronic forms, is used for a multitude of purposes by an array of individuals and organizations internal and external to the healthcare industry. The primary users of such data include doctors, nurses, physicians, clinics and hospitals that provide care for patients. Secondary users are often those who utilize and organize this health information for an assortment of business, societal and government purposes—outside of providing care. These users include organizations that pay for healthcare benefits, such as government programs like Medicaid and Medicare, managed healthcare providers, and traditional health insurance companies. These secondary users and payer organizations also conduct analyses on the quality of healthcare provided by such organizations relative to its costs, as a part of their management functions. Other secondary users may include social science and medical researchers, social

welfare and rehabilitation programs, pharmaceutical companies, public healthcare services, marketing firms, the judiciary system, and even the media. These entities use health information for the astute purposes of:

- Researching the costs and benefits to alternative treatment plans
- Determining one's eligibility for social programs
- Understanding the current state and local health needs
- News reporting
- Targeting possible markets for new and existing products.

Vendors of health-related products and marketing firms also receive and analyze health information in an effort to help them target particular types of patient for direct marketing. The types of data and information received by primary and secondary users vary greatly among individual organizations. The exchange of data within and across these organizations are dynamic and highly complex. Nonetheless, the amount of patient data that these organizations obtain is vast [16].

Furthermore, the United States Federal Government often collects data provided under Medicaid and Medicare for reimbursement purposes, however states also collect an expansive amount of patient-identifiable information for outside purposes. Agencies and state-health organizations are able to provide services and collect private identifiable data about each patient just as providers in private healthcare organizations would. In the provider capacity, state health organizations would release identifiable data and personal information, with patient consent, to insurers and separate providers who may be privy to that information. These agencies collect data for the purpose of analyzation and dissemination of information on health status, personal health complications, quality of provided services and availability of health resources. However, this comes at a cost to patients, as their personal information and identifiable biodata is handled my many entities, and exist in several databases beyond their control. The categorization of data collected are dependent upon the services and functions each health department possesses within its authority. Professional and facility licensing, Medicaid, environmental services, alcohol and drug abuse, and/or mental services are not located and utilized consistently in all state health departments across the country. This further highlights the concern with many entities having access to proprietary patient data.

Ordinarily, state health departments will collect a patient's identifiable data related to health service utilization and costs, personal health status and risk—surveilling health data, alcohol and drug abuse services, and mental health services, in addition to other health-related categories. The types of data systems related to each of these categories are often extensive.

Typically, the databases that are created for these purposes have a designated administrator who is responsible for managing the uses and protection of patient data. These types of data are released in an identifiable form only in select situations: 1. Research purposes for which there has been an approved human subjects review and a data-sharing agreement that outlines restrictions on the use of data, destruction of data at the end of research, and the penalties for violating the agreement; 2. The investigation of a reportable disease or condition for the purposes of protecting the public's health.

In the latter case, identifiable data are released to specially authorized public health investigators or private physicians who are responsible for care of the person believed to have a reportable condition or disease (e.g., measles, sexually transmitted disease, tuberculosis, birth defect, cancer). The steward of the database determines which staff members are allowed to access identifiable data for the purposes of analyzing them. Finally, state laws include penalties that prohibit improper release of data by a state government employee[2].

IV. ANALYSIS

The cost savings to companies from the gathering of data via a computerized modeling system, rather than traditional wet-bench biology, led to a dramatic increase in the formation of bioinformatics companies beginning in the 1990s. However, this rapid increase in the types and sources of bioinformatics data meant that the data collected by these companies were a new source of security and privacy concerns for individuals and corporate entities trying to protect their interests.

One challenge intrinsic to data privacy and security in the field of bioinformatics is that a large proportion of bioinformatics solutions have been developed in open source software, such as Perl and Unix. This was welcomed by groups concerned with the cost of obtaining software code from proprietary corporate databases and by developers (often academics, e.g.,

students and researchers) who shared a philosophical belief in the widespread sharing of data[4].

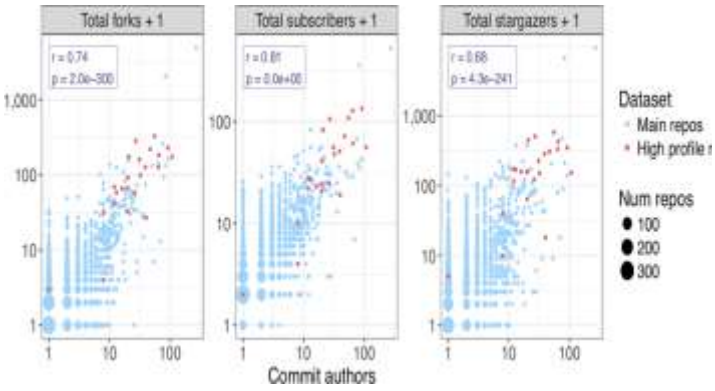

Figure 2: Size of Developer Community

This figure represents the size of the developer community based on free / open source softwares such as Github, SourceForge, and Bitbucket. This further exemplifies the reliance on open source communities within the field of bioinformatics and computational genomics.

During the dot.com decline of 2000, many companies preferred and encouraged the open source movement in bioinformatics, due to cheaper utilization costs. Other supporters argue that open source software is more reliable and better developed as broad usage and diffuse expertise allow for optimization [9]. The hope is that, in exchange for the tools needed to conduct research, researchers will freely contribute to ongoing projects. However, the ready availability of open source code allows easier hacking of the information developed from these bioinformatics systems. It also became difficult to define and protect intellectual property and commercial interests with universally available data. One especially pertinent example of potential concerns facing privacy, especially in the health informatics subdiscipline of the bioinformatics field, is the increasing utilization of large-scale genomic databases. These data sets are used to study the association between genomic composition and molecular, organ, and tissue-level systems, a study that has proven essential to understanding the genetic predisposition to complicated medical disorders. This information, which allows researchers to make advances in the knowledge and treatment of disease, brings fear of identification and discrimination for those individuals carrying medically stereotyped genetic information.

Questions of patient privacy are complicated by the nature of the data: Genomic information by definition is the ultimate identification tool, which carries the further risk of implicating family members. To protect personal privacy, steps are taken to anonymize or pseudonymize the data, if identity is not required (as in a health care setting). This is often done automatically on collection by assigning each genome a randomized ID, but further precautions must be taken with the genetic information itself to prevent potential reidentification of samples. Potential solutions include deleting or altering incriminating sequences, adding extra "noise" sequences, or providing only short, nonincriminating sequences relevant to the researchers' query. Though, these approaches have yet to be pursued on a larger scale[7].

Sharing genomic data must be done in a secure way, and this is one example of the potential application of the trusted third party, which can function as an encryption system and a further layer of de-identification for the genomes collected. There are still limitations to protecting data in this way, as the nature of the genetic material that allows identification is continually changing as science progresses. This material must accordingly be continually monitored for incriminating sequences. This rather extreme example highlights potential difficulties as well as the warranted necessity to protect data and the inevitable compromises made to open-access data to maintain the level of privacy warranted.

The scientific community has marked a significant milestone in the study of genes, the completion of the "working draft" of the human genome. This work, which was recorded in special issues of the journals Nature and Science in 2001, heralds a new beginning for advances in the prevention, diagnosis, and treatment of many genetic and genomic disorders. The availability of this wealth of raw data has a significant effect on the field of bioinformatics, with a great deal of effort being spent on effectively and efficiently storing and accessing these data, as well as on new methods aimed at mining the data in order to make revolutionary medical discoveries [15]. These advances have generated numerous new and exciting challenges with which computing professionals will have to grapple.

A large variety of genomic data sources have emerged, resulting in inconsistent terminology and data formats. Many of these come from independent studies that were organism based, such as for cancer research. While much of this information is publicly available over the Internet, comparison and unification are critical for much of the sequence analysis that remains to be done. However, the fact that these data sources were

developed for different purposes by different researchers using different methods often makes the data difficult to unify. Regarding data standards, the emergence of the macromolecular crystallographic information file (mmCIF) and extensible markup language (XML) provides standards that can produce a common format for data [12]. It is critical that the bioinformatics community either decide on or gravitate toward one common format that will make data sharing vastly easier.

A. An Integrative Framework

Additionally, collaborative research requires conceptualization and implementation of an integrative framework. Apart from standardization of data formats, this will require development of Web-based user interfaces, standards for access to the data and data warehousing capabilities, as well as interoperable software components. The development of a standardized, Web-based, globally distributed view is critical in the light of researchers working together across several languages and countries. A standardized interface to the multiple heterogeneous databases is an important objective for developers. Two distinct approaches have been used for data warehousing. IBM uses a federated database, in which the data remain in the original separate sources and are accessible with a single query. The data from various sources are brought into a data warehouse, where data freshness depends on the frequency of data replication. The issue of which approach is more useful and when is yet to be determined.

Examples of data sources for a federated database or data warehouse are the three primary sequence databases: GenBank (NCBJ), Nucleotide Sequence Database (EMBC), and the DNA Databank of Japan (DDBJ). These are repositories for raw sequence data, but each entry is extensively annotated and has a features table to highlight the important prospects of each sequence. The three databases exchange data on a daily basis [9].

Interoperability among software components is a crucial goal for successful collaborative work. Object management groups (OMG) and a life sciences research domain task force's goal to establish common object request broker architecture (CORBA) as the standard for interoperable software components offer potential.

B. Future Computing Needs

While the knowledge gained from the sequencing of the human genome via bioinformatics is expected to change our lives, more powerful and robust computing is needed to develop the tools for genetically based drug design, medical diagnosis and treatment, and agricultural application, among others. The power and robustness should come from development of both software algorithms and hardware. Many traditional algorithms, including Bayesian statistics, dynamic programming, and Markov chains, have already been used for sequencing.

With the enormous size of databases today, the efficiency of these algorithms is critical for successful use. Dynamic programming, for example, can considerably slowdown in multiple sequence alignments because the complexity of the calculations increases for more than two sequences. However, improvements in the algorithms and use of heuristics have improved the situation significantly. Future research should focus on development of such heuristics. Moreover, mining the data for patterns is essential for newer discoveries. Pattern recognition algorithms and neural networks have been applied to bioinformatics research. Neural networks can also be applied to classification as well as decision problems. Other artificial intelligence–based algorithms, like case-based reasoning (CBR), can be useful in this regard[3].

The issue is to embellish the currently available algorithms and heuristics as well as develop new ones to deal with the need for sequencing, prediction, and pattern recognition. Comparative studies of the effectiveness and efficiency of these algorithms are essential for further applications. The term "deep computing" for bioinformatics research, implies the use of powerful machines executing sophisticated software based on innovative algorithms to solve complex problems like mapping, modeling, and visualization. From a hardware perspective, both a supercomputing approach and a distributed computing approach have been used in bioinformatics [5]. Grid computing allows geographically distributed organizations to share applications data and computing resources. While the distributed approach is less expensive, it raises further issues pertinent to distributed processing and data distribution, particularly those over Internet services.

To facilitate access, several tools have been developed or are works in progress. These tools include GeneX, an example of a system that helps with the storage, organized retrieval, and analysis of gene expression data. Among the most important software tools for the

understanding of DNA and protein sequences are sequence similarity and alignment tools such as Basic Local Alignment Search Tool (BLAST) and a sequence alignment algorithm using a flat file format known as FASTA. A user is able to visualize the complexity of the back-end databases and the front-end query tools with which BLAST deals [11]. These tools allow a user to compose an unknown sequence with a database of sequences from other organisms that are better understood. These programs report the hit in the database, along with the estimated statistical significance of the hit.

DiscoveryLink is described as a middleware software product from IBM. It can be used to build a federated database application. A prototype system called MyGrid is being developed at universities in the UK. The new system will allow biologists to analyze information in many databases in a standardized fashion, which until now required many types of custom-built software. It is reported that with MyGrid, biologists will not become programmers, for the team is using software agents to help translate and standardize the contents of conflicting formats. MyGrid should automatically find any information relevant to the study, searching for genomic and proteomic data, regulatory networks, and any other relevant facts. The robustness of data submitted to the primary database is important in the context of bioinformatics software. Much of the progress in bioinformatics is in fact due to the accelerated rate at which sequence data are being produced [17]. Bioinformatics is required at several different stages during DNA sequencing. First, the data produced at every stage of generation and analysis must be captured in real time. Second, sophisticated software algorithms are required to assemble, edit, and compare the sequence data. Genomic databases need to facilitate the storage and analysis of large amounts of data, but also have a user-friendly format and graphical display to allow relevant data to be displayed and analyzed.

Beyond storage and integration, the computing capabilities required for these new scientific developments are diverse, with complex operational requirements:

- Availability—continuous access to the distributed data warehouse and Web sites
- Security—appropriate controls for access and information assurance
- Data protection—loss of data is decidedly unacceptable, and backup is critical
- Data mobility—data need to be available to the right user, at the right time, in the right place
- Data purpose—the same data may have multiple purposes and views
- Data sharing—access to all information by all participants
- Real-time availability—data must be available at all times in a global setting[15]

IBM, a leading vendor in bioinformatics tools, proposes secure access to data from a growing number of increasingly diverse data sources and the ability to put that data to use quickly; simplified sharing of data and functionality among the diverse applications and tools used in different research areas; easier collaboration internally and externally to turn data into knowledge, as well as the ability to manage and share that knowledge more efficiently; secure storage and easier management of data; faster installation of new applications and integration with valuable existing systems, making research and product development more efficient; and smooth integration of outsourced functions [14].

V. RECOMMENDATIONS

Besides hoping to maintain the integrity of the data itself, companies and researchers need to protect their commercial or academic interests in an increasingly competitive field . As such, it has become important to develop more secure methods to store and transfer data. Depending on the privacy needs of both the data and the data user, as well as the method of data sharing, multiple schemes have been proposed.

Another proposed solution is to create a “trusted third party,” which is then used either as a method to transfer encrypted data while maintaining input and query secrecy or as a way to store data in a secure but accessible form. Interestingly, although these proposed solutions increase the level of data protection as desired, the field’s intrinsic wish to maintain data accessibility and sharing is still evident. It is understood that larger and more complete data sets provide higher-quality analyses, and as such the data are still available, albeit in a secure form.

VI. CONCLUSION

Bioinformatics is the utilization of computer software to study biological information and methods. Some examples of the diverse data produced by the field include analysis of genomic, proteomic, and

metabolomics sequencing; computational biology models; biodiversity measurements; and records and models of protein expression, regulation, and structure. The potential application of such computational analysis is limitless, but work is focused primarily on creating ways to effectively store, process, and manipulate large data sets, on deriving statistical or mathematical analysis from such data, and on creating and analyzing models of important molecular, physiological, and ecological systems. Bioinformatics has potential medical, agricultural, and biological applications, both commercially and academically, as the patterns derived from samples and modeling can be used to better understand, develop, and optimize treatments, products, and crops. Bioinformatics, commonly used by the health care system to manage large amounts of patient data, is now being used in international collaborations focused on understanding disease states and normal physiology for commercial purposes as well. Bioinformatics is a promising field with the potential to be developed further into a larger opportunity for both computer scientists and biologists. Excellent working examples have been developed and is in use such as the GenBank and the PubMed databases. It is accessible but possess the risk of loss and misuse. Genomics as a field has already made huge impacts on society including the Human Genome Project and in selective breeding in animals and plants. It has advantages medically and economically, but it also has disadvantages related to the environmental and consumer health. There have also been ethical concerns of privacy, mistrust in racial or economic status, animal rights, plant "naturalness", and the intrusiveness of "playing God" voiced, while exploring these different innovations. In summary, bioinformatics and computational genomics have drastically improved the exploration of hereditary qualities, biotechnology, and medicine. We now have the advancements needed to discover new medications and drugs. However, these advancements and improvements to human life often come at the cost of exposing personal biodata.

REFERENCES

[1] Attwood, Theresa K. "Genomics. The Babel of Bioinformatics." *Science 290*, no. 5491. Retrieved September 25, 2017.

[2] Gibson, Greg and Spencer V. Muse. A Primer on Genome Science. Sunderland, MA: Sinauer Associates, Inc., Publishers, 2002.

[3] Hlodan, Oksana. "For Sale: Iceland's Genetic History." Available at the American Institute of Biological Sciences' Web site at www.actionbioscience.org/genomic/hlodan.html.

[4] Baxevanis, Andreas. D. and B. F. Francis Ouellette (Editors). Bioinformatics, 2nd edition. New York: John Wiley & Sons, Inc., 2001.

[5] Westhead, David R., J. Howard Parish, and Richard. M. Thyman. Bioinformatics. Oxfordshire, UK: BIOS Scientific Publishers, 2002.

[6] Swope, William. C. "Deep Computing for the Life Sciences." *IBM Systems Jou*rnal 40, no. 2 (2001): 248-262.

[7] Mount, D.W. Bioinformatics, Sequence and Gene Analysis. New York: Cold Spring Harbor Laboratory Press, 2001.

[8] Lesk, A. M. Introduction to Bioinformatics. Oxford, UK: Oxford University Press, 2002.

[9] Head-Gordon, Teresa and John C. Wooley. "Computational Challenges in Structural and Functional Genomics." *IBM Systems Journal* 40, no. 2 (2001): 265-291.

[10] Westhead, David R., J. Howard Parish, and Richard M. Twyman. Bioinformatics.

[11] Graham-Rowe, Duncan. "Software Agents Could Tackle Human Genome Data Explosion." *New Scientist* 179, no. 2407 (2003): 22.

[12] Goble, Carole A. et al. "Transparent Access to Multiple Bioinformatics Information Sources." *IBM Systems Journal* 40, no. 2 (2001): 532-551.

[13] Sensen, Christoph W. (Editor). Essentials of Genomics and Bioinformatics. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co., 2002.

[14] Regalado, Antonio and Leila Abboud. "New Genetics Map to Explore Links to Ailments." *The Wall Street Journal* October 30, 2002, p. D4.

[15] Ma W, Chen L, Zhou Y, Xu B. What Are the Dominant Projects in the GitHub Python Ecosystem? 2016 Third International Conference on Trustworthy Systems and their Applications (TSA). 2016. pp. 87–95. https://doi.org/10.1109/TSA.2016.23

[16] Sheoran J, Blincoe K, Kalliamvakou E, Damian D, Ell J. Understanding "Watchers" on GitHub. Proceedings of the 11th Working Conference on Mining Software Repositories. New York, NY, USA: ACM; 2014. pp. 336–339. https://doi.org/10.1145/2597073.2597114

[17] Spotlight on Bioinformatics. NatureJobs. Nature PublishingGroup;2016; https://doi.org/10.1038/nj0478