Quarterly Report

Prepared for:

Georgia Department of Economic Development

Life Science and IT - Bioinformatics

Q3 2013

Georgia®

Problem Statement

IP2Biz will deliver a report for the Georgia Department of Economic Development on the topic of Bioinformatics for the 3rd Quarter of 2013. Consistent with the Taxonomy in our agreement, iP2Biz will review the selected topic for:

i. Identification and discussion of relevant industry trends

ii. Recent developments and breakthroughs.

iii. Significant new patent issuances

iv. Displays of new research (related to topic) conducted in Georgia, major research, funding rewards, major investments, acquisitions and mergers).

In particular, iP2Biz will look at the instances of Bioinformatics in the areas of genome computation and sequence analysis. For each topic, iP2Biz will illuminate wherever possible the impact of the measured parameter on the companies in Georgia and how the state fares related to the identified broader national trends

iP2Biz Contact

John Bacon
jb@ip2biz.com
404.961.7660

Please do not hesitate to contact us regarding any questions you might have.

Executive Summary

Bioinformatics is the combination of the fields of computer science and biology. Bioinformaticists use computer tools to study the genes that define each organism on Earth. Bioinformatics requires enormous computing resources. As a result, open source and paid software is being developed to advance the field by using existing computing resources more efficiently. Work in bioinformatics is supported by the U.S. federal government because the field is believed to hold great promise: this emerging, multidisciplinary field has the potential to improve human life because advances in bioinformatics can be applied to advance the field of medicine. Bioinformatics is an area of science in which there is much university activity. Georgia universities such as Emory University, Georgia Tech, and Georgia State University have made efforts in bioinformatics. However, in spite of the activities of Georgia universities, Georgia has not yet been recognized as a major center of bioinformatics. Georgia has the academic resources to become such a center should the state wish to pursue such recognition.

Technology Snapshot

Overview

Bioinformatics is the application of computer science to the study of biology. Other names for bioinformatics include computational biology and biomathematics. Bioinformatics university programs prepare students to use mathematics and computer science for biological and medical research. Students take courses in mathematics, statistics, computer science, and biology. They also learn how to perform research and how to program software. Typically, a degree in bioinformatics prepares students to work as researchers in the pharmaceutical and biotechnology industries. Bioinformatics research often has to do with improving computing techniques or methodologies to make them more efficient and more accurate. The emphasis is often not on the biological data themselves, but on computing, technological innovation, and improvement with respect to applications. Bioinformatics centers often use computer software or computer tools to serve their biological data requirements. Bioinformatics tools have the potential to offer more efficient drug discovery and development processes. The field of bioinformatics improves the analysis of genetic data collected from a multitude of biological specimens. For example, the Human Genome Project has allowed deeper insights into genetics by elucidating the topics of protein and gene expression, metabolic pathways, and disease states.

This knowledge can be applied by other fields to develop medical treatments based on the human genetic code. Much of this knowledge has resulted from the use of bioinformatics software tools to manipulate data. Figure 1 depicts a breakdown of the sub-spaces within bioinformatics.

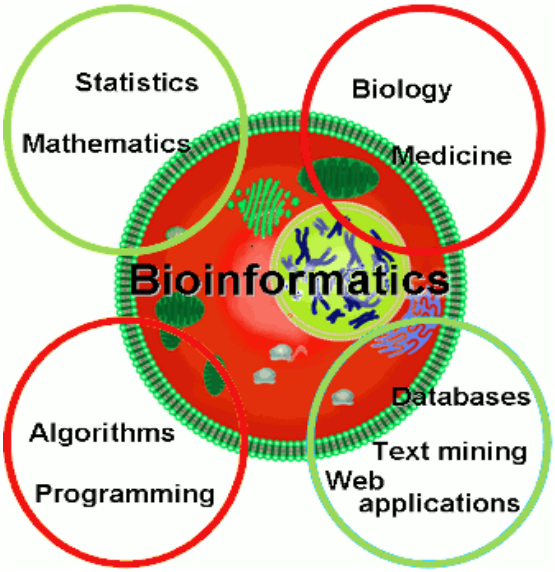


Figure 1: Taxonomy of Bioinformatics.

Genome Computation

Genome computation is the acquisition of a complete set of genetic information for a given living specimen. Genome computation is an activity performed by researchers in the field of genomics. Genomics is a sub-discipline of genetics in which researchers use bioinformatics to sequence, assemble, and analyze the function and structure of genomes. A genome is a complete set of DNA within a single cell of an organism. Living beings store genetic data in long sequences of microscopic pairs of chemically bonded molecules. These molecules are called adenine, guanine, cytosine, and thymine. Guanine can only bond to cytosine and adenine can only bond to thymine. The sequence and identity of these possible pairs determine all of the characteristics of a living being. Just as the language of computers is a sequence of numbers in the binary system (e.g. 0,1,0,0,1,0,1,1,0), the language of DNA is a sequence of molecular pairings (e.g. GC,AT,TA,CG,GC,AT). Sequences of millions of digits of binary code are interpreted by computers to define the colors of the pixels on a display. Similarly, sequences of millions of genetic pairs define the traits of a living organism. Figure 2 and Figure 3 depict the structure of DNA and the combinations of adenine, guanine, cytosine, and thymine from different zoom levels.

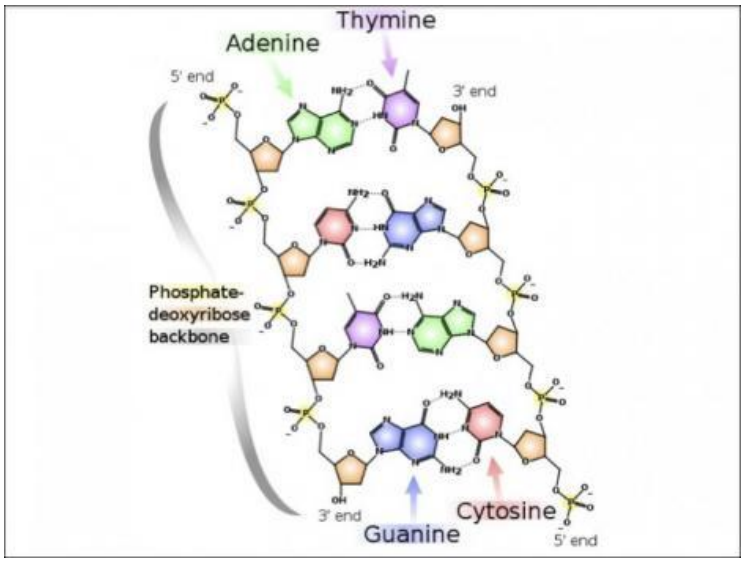


Figure 2: Close view of pairs of adenine, guanine, cytosine, and thymine.

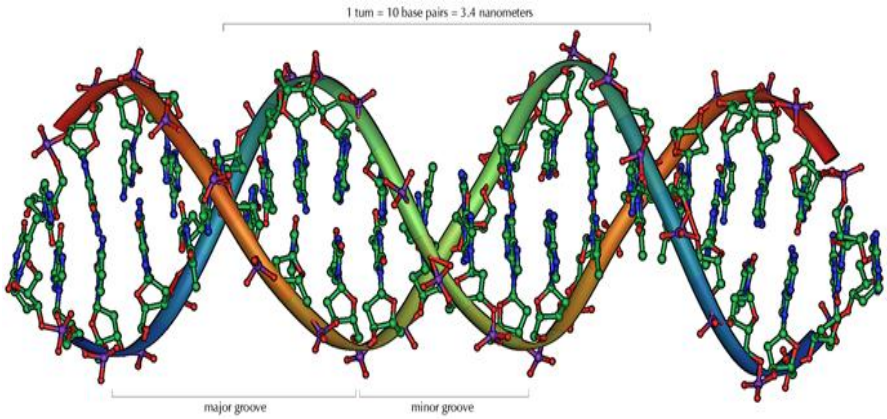


Figure 3: View of double helix DNA structure from further away. Pairs of adenine, guanine, cytosine, and thymine form the inner links between the doubly helical outside strands.

Genomics i of genetic data being gathered. Because genetic data defines who we are, the desire to acquire more and more

of such data is very strong in the scientific community. The genetic sequence of an organism contains an enormous amount of data, and the scientific community is responding by developing ways to effectively use and manage this data. Genomes can range anywhere from 4,000 pairs to 670,000,000,000 pairs. Furthermore, organisms that reproduce sexually have two or more copies of the genome. Humans have two copies of their inherited genome of 3,200,000,000 pairs each. Full sequence data have been archived for many thousands of species, and more than 3000 humans have been sequenced to some substantial extent and reported in the scientific literature; new sequencing data gathering efforts are expanding at an exponential pace.

Sequence Analysis

Sequence analysis is the identification of patterns and functions in the genetic information of one or more species. Sequence analysis helps scientists make sense of the genetic data gathered by genome computation. While the challenge of genome computation is to acquire genetic data, the challenge of sequence analysis is to understand such data. The pace of innovation in genomic data creation is higher than the rate with which humanity can manage and understand these data. This capability gap must be addressed before the field of genomics can make a significant leap forward. Such a leap could lead to accelerated advances in the understanding of evolution and medicine. Figure 4 compares humanity's ability to process and store data with the volume of genetic data being generated. The green line represents humanity's ability to store data on hard drives. The purple line represents humanity's ability to process data using CPUs. The blue line represents the volume of data being generated. After 2007, the amount of genetic data generated surpassed humanity's abilities to store and process it.

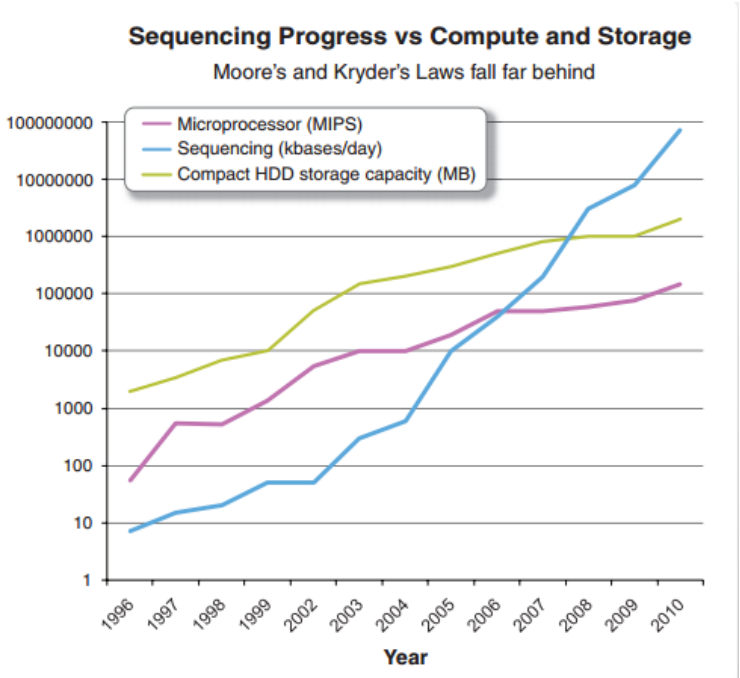


Figure 4: Ability to store and process data versus data volume.

Recent and future developments

For years, the primary concern in most sectors was collecting enough genetic information. Now, the problem has reversed: many bioinformatics companies have more data than they know what to do with. This is especially true in life sciences. In recent years, the amount of data that has become available to researchers has increased almost exponentially. Since the decoding of the human genome, every corner of biological sciences has produced its own genomics-related subfields: epigenomics, transcriptomics, metabolomics, proteomics, interactomics among hundreds of others. This means that people in industry and academia are keener than ever to make use of enormous data sets to investigate molecular mechanisms with greater accuracy than ever. Organization, analysis and interpretation of this information is required to apply it to useful ends and create knowledge. In the pharmaceutical industry, such data is useful in computational systems biology. Using the vast amounts of available information, it will be possible to create algorithms and computer models to accurately predict drug - metabolism and toxicity interactions between drugs and clinically relevant targets. These in silico experiments could lead to faster, cheaper, safer drug development.

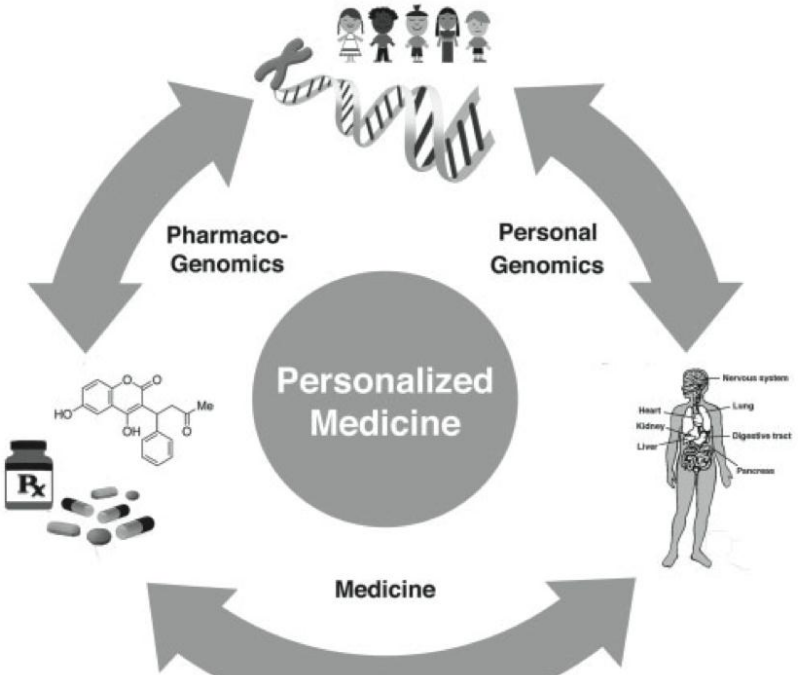


Figure 5: Ailments of the patient's body are treated by patient-specific medicines that are designed based on the patient's genome. The interrelated fields of medicine, personal genomics, and pharmacogenomics seek to achieve this lofty goal.

Scientists in the genetics, bioengineering, bioinformatics, and medicine departments at Stanford University have recently reviewed important trends in the field of bioinformatics. Specifically, these scientists have reviewed the newly discovered potential for using bioinformatics to improve human life by applying bioinformatics advances to the field of medicine. Although bioinformatic tools can be used to understand any living species, studying the human genome with the hope of ending human diseases has captured the interest of many experts in bioinformatics. The Stanford scientists predict that bioinformatics will usher in an age of "personalized" medicine. Personalized medicine will grant physicians the ability to sequence a patient's genome upon admittance into a medical office. Scientists are presently attempting to

sequence medical patients' genomes so that each patient's body's characteristics can be completely understood before administering medicines. When understood, a patient's genome reveals sensitivities to drugs and propensities for diseases that can be detected and prevented. Significant medical research efforts have recently been made to advance personalized medicine with the help of bioinformatics. The era of personalized medicine will start because of advances in genome computation that have drastically reduced the cost of sequencing the genome of a human. Scientists postulate that "soon, full genome sequencing will cost less than $1000." At this cost, genome sequencing could become a routine medical procedure. As shown in Figure 4 and postulated by scientists "the world will be inundated with individual genomic data in the coming years" because of advances in genome computation technology.

As mentioned before, the large influx of genetic data from genome computation techniques has led to efforts that seek to understand and make use of that data via sequence analysis. Although not as successful as genome computation efforts, sequence analysis efforts have led to the discovery of "thousands of DNA variants [that] have been identified [as] associated with diseases and traits." Such identifications form the beginnings of the fulfillment of the promise of personalizing medicine via bioinformatics: in the case of chemotherapy, incidents of adverse events have been reduced by first checking patient's genotype for susceptibility to drugs like abacavir, carbamazepine, and clozapine. Figure 5 displays the relationships and categories of study involved in making the next major bioinformatics-enabled advance in medical care. In Figure 5, bioinformatics plays a very important role because bioinformatic tools are used to advance both the field of personal genomics and the field of pharmacogenomics. Specifically, experts working in the sub-fields of sequence analysis and genome computation will have to be responsible for the advances that enable the medical revolution that could be brought about by the field of bioinformatics.

Bioinformatics must develop further before it can improve the lives of most people. Today, even if doctors had access to each patient's genome, only a small percentage of that genome data could be used. Specific areas of bioinformatics have been identified to be in the greatest need

of further development. One area of bioinformatics that must see improvement is data accuracy. Although bioinformatics has succeeded in producing lots of genetic data from living specimens quickly, bioinformatics has not succeeded in producing these data accurately. Today, every 100 kb of data gathered contain 1 error. This error rate has serious implications for humans because it implies that, sequenced today, the data representing a typical, 3 GB human genome will contain 30,000 errors. Thus, the high-throughput genome computation techniques used today must be improved before they can be safely implemented to help humans. Bioinformtics experts postulate that new, less erroneous data algorithms must be developed. One source of the errors in genetic data is the method of data interpretation used to process genetic data. In order to process genetic data more quickly, scientists typically depend on a reference genetic sequence that is used to interpret the new genetic data: new genetic data is "mapped" to a reference sequence that is assumed to be similar to the new sequence. Errors therefore exist because of "reference bias," the situation in which the assumption of similarity between the reference sequence and the newly calculated sequence does not hold true. Unfortunately, the practice of reference mapping cannot be eliminated by scientists using today's algorithms because reference mapping, although erroneous, drastically accelerates genome interpretation: a genome sequence that would take days to compute with the assistance of reference mapping takes years to compute without reference mapping. Algorithm development is the responsibility of bioinformatics researchers who specialize in computer science. As a result, numerous bioinformatics algorithms are under development.

Developments in Georgia

Advancing bioinformatics is a nationwide goal. Although Georgia has several centers of bioinformatics knowledge, iP2Biz believes that the contribution to major advances in bioinformatics made by Georgia universities is somewhat lagging behind those of other U.S. states. Based on publication impact, university research centers of very high achievement in bioinformatics include Stanford University in Stanford, CA and Johns Hopkins University in Baltimore, MD. Harvard Medical school is also a major bioinformatics research center. A Stanford University review of major progress in the field of bioinformatics that covered high-impact contributions published in prestigious scientific journals such as Science and Nature did not include contributions made in the state of Georgia. Table 1 displays university rankings in the field of bioinformatics; no Georgia university has achieved a top ten ranking in this space. The potential for the improvement of Georgia universities in the bioinformatics space is expressed in Table 2 and Table 3. Table 2 displays the rankings of selected Georgia universities in the field of biological sciences. Table 3 displays the rankings of Georgia universities in the field of computer science. Since bioinformatics is a combination of biological sciences and computer science, Georgia can increase its future influence on bioinformatics.

Table 1: Top ten U.S. universities in bioinformatics.

Rank	University	Location
1	Stanford University	Stanford, CA
2	Massachusetts Institute of Technology	Cambridge, MA
3	Harvard University	Boston, MA
4	University of California - Berkeley	Berkeley, CA
5	University of Washington	Seattle, WA
6	Washington University in St. Louis	St. Louis, MO
7	Johns Hopkins University	Baltimore, MD
8	University of California - San Francisco	San Francisco, CA
9	Yale University	New Haven, CT
10	Princeton University	Princeton, NJ

Table 2: Selected Georgia university rankings in biological sciences.

Rank	University	Location
34	Emory University	Atlanta, GA
46	University of Georgia	Athens, GA
56	Georgia Institute of Technology	Atlanta, GA
144	Georgia Regents University	Augusta, GA
160	Georgia State University	Atlanta, GA
220	Morehouse College of Medicine	Atlanta, GA

Table 3: Selected Georgia university Rankings in computer science.

Rank	University	Location
10	Georgia Institute of Technology	Atlanta, GA
91	University of Georgia	Athens, GA
99	Georgia State University	Atlanta, GA

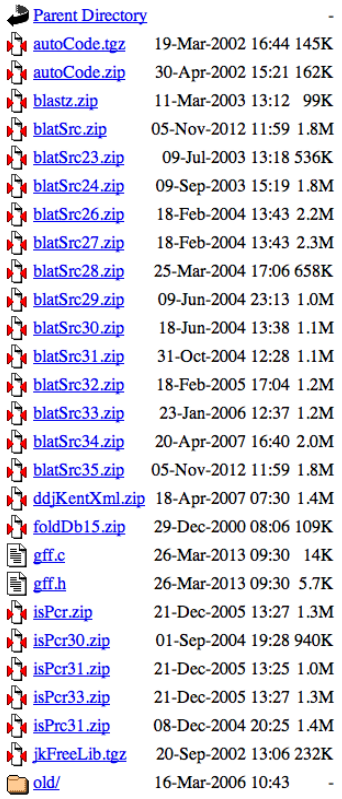
There are more faculty members focusing on bioinformatics at Georgia State University than at any other school in Georgia. At Georgia State, bioinformatics is studied as part of the computer science program. Bioinformatics became a major focus at Georgia State when funding from the National Institutes of Health (NIH) was given to Dr. Yi Pan, the chair of the Georgia State computer science department. NIH funding was given to Dr. Pan as part of the NIH's goal to advance the study of bioinformatics nationwide. The NIH places a very strong emphasis on bioinformatics and distributes funding accordingly. The NIH states that "Computer applications have become an essential part of biomedical research to analyze the often huge amounts of data." The NIH is currently accepting funding applications for researchers working in the field of bioinformatics. Georgia State holds an annual conference called the International Symposium of Bioinformatics Research and Application to attract bioinformatics researchers to Georgia State's campus.

In 2013, Georgia Tech and Emory University continue the tradition of biennial International Conferences in Bioinformatics by inviting leading scientists to present the latest advances at the forefront of genomics and bioinformatics. The conference features the prestigious Margaret O. Dayhoff lecture, the third since its inauguration in 2009. The Institute for Integrative BioSystems is the home for several research centers involved in systems biology, computational biology and bioinformatics. Personalized medicine is of particular interest to the NIH and Georgia Tech: recently, the NIH granted professor Craig Forest $4.3 million to advance personalized medicine by gathering biological data via robotic means. Lastly, Healthcare@EI2 works directly with the State of Georgia and industry partners to advance the research and deployment of healthcare information technologies.

Technology Forecast

Future progress in bioinformatics will lie in the ways in which researchers use the vast troves of data generated by genome computation. Specifically, researchers will develop new computer algorithms that process genetic data more efficiently by using fewer "CPU-hours": new algorithms reduce the time required for a calculation to be completed by a computer. There already are dozens of algorithms and software packages that bioinformatics researchers have made freely available to their colleagues in their field. Algorithms are distributed via the Internet in an open source fashion as shown in Figure 6. Figure 6 displays the freely accesible download page the "BLAT" algorithm developed by Jim Kent at the University of California,

Snata Cruz in 2002. The generation of new, more efficient successors to algorithms like BLAT will be a focal point of bioinformatics research in the future. Although improvements in computer hardware will accelerate progress in bioinformatics, improvements in the software that harnesses the power of the hardware in better ways is needed much more.



Parent Directory		-
autoCode.tgz	19-Mar-2002 16:44	145K
autoCode.zip	30-Apr-2002 15:21	162K
blastz.zip	11-Mar-2003 13:12	99K
blatSrc.zip	05-Nov-2012 11:59	1.8M
blatSrc23.zip	09-Jul-2003 13:18	536K
blatSrc24.zip	09-Sep-2003 15:19	1.8M
blatSrc26.zip	18-Feb-2004 13:43	2.2M
blatSrc27.zip	18-Feb-2004 13:43	2.3M
blatSrc28.zip	25-Mar-2004 17:06	658K
blatSrc29.zip	09-Jun-2004 23:13	1.0M
blatSrc30.zip	18-Jun-2004 13:38	1.1M
blatSrc31.zip	31-Oct-2004 12:28	1.1M
blatSrc32.zip	18-Feb-2005 17:04	1.2M
blatSrc33.zip	23-Jan-2006 12:37	1.2M
blatSrc34.zip	20-Apr-2007 16:40	2.0M
blatSrc35.zip	05-Nov-2012 11:59	1.8M
ddjKentXml.zip	18-Apr-2007 07:30	1.4M
foldDb15.zip	29-Dec-2000 08:06	109K
gff.c	26-Mar-2013 09:30	14K
gff.h	26-Mar-2013 09:30	5.7K
isPcr.zip	21-Dec-2005 13:27	1.3M
isPcr30.zip	01-Sep-2004 19:28	940K
isPcr31.zip	21-Dec-2005 13:25	1.0M
isPcr33.zip	21-Dec-2005 13:27	1.3M
isPrc31.zip	08-Dec-2004 20:25	1.4M
jkFreeLib.tgz	20-Sep-2002 13:06	232K
old/	16-Mar-2006 10:43	-

Figure 6. Free algorithm download method.

Once new algorithms improve researchers' ability to make sense of genetic data, the next step is to apply the understanding of the data to the improvement of human life. Stanford bioinformatics experts agree that achieving medical relevance should be a key goal of bioinformatics. In the future, bioinformatics researchers will make their work medically relevant by distributing their collective understanding of the human genome to clinics via curated online databases. One challenge of distributing bioinformatic conclusions to the medical community is, again, the sheer volume of data that must be processed. Once the work of bioinformatics researchers is distributed to clinics and medical offices, the last challenge is to integrate an understanding of each patient's genome into each patient's electronic medical record. One example of an effort in this space is the BioBank system at Vanderbilt University that links patient DNA with electronic medical records. Although BioBank cannot be used to actually treat patients in its current state, it provides a platform for researchers to improve upon, and it is a vision of the future of the intergration of bioinformatics with medicine.

Market

The global bioinformatics market was valued at $2.9 billion in 2012 and is poised to reach $7.5 billion by 2017 at a CAGR of 20.9%. The growth of the bioinformatics market is driven by decrease in cost of DNA sequencing, increases in government initiatives and funding, and growing use of bioinformatics in drug discovery and development processes. It is expected that the market will offer opportunities for bioinformatics solutions manufacturers with the introduction and adoption of upcoming technologies such as nanopore sequencing and cloud computing. However, factors such as dearth of skilled personnel to ensure proper use of bioinformatics tools and lack of integration of a wide variety of data generated through various bioinformatics platforms are hindering the growth of the market. Manufacturers of bioinformatics solutions will face further challenges with regard to industry consolidation and management of high volume data.

North America accounted for the largest market share of the bioinformatics market, followed by Europe, in 2012. However, Asian and Latin American countries represent emerging markets,

owing to a rise in research outsourcing by pharmaceutical giants, increasing number of Contract Research Organizations (CROs), rise in public and private sector investment, and growing industry - academia partnerships. The major players in the bioinformatics market are Accelrys, Inc. (U.S.), Affymetrix, Inc. (U.S.), Life Technologies Corporation (U.S.), Illumina, Inc. (U.S.), and CLC bio. (Denmark).

The growth in this area is likely to happen in 3 key areas: databases, analysis, and hardware. Database services to allow companies to collate useful information are central to efforts of employing it in a fruitful fashion and accordingly this market is set to grow from $1.5 billion to $3.4 billion between now and 2017. The software required to analyze the collected information and make it accessible by means of more tractable outputs like visualizations will prove vital to this economy and as such the global market is predicted to grow from $1.1 billion to $2.9 billion in the next 4 years. Many researchers now do not require physical storage devices to store their gargantuan data sets; they make use of cloud computing and remote storage services. In a relatively short time people have become divorced from the notion of physical hard disks, however the reality is that somewhere there has to be a physical storage medium with the relevant information recorded on it. With the other trends in this area it may be unsurprising that the market for hardware infrastructure to support bioinformatics is poised to grow from $0.5 billion to $1.2 billion by 2017.

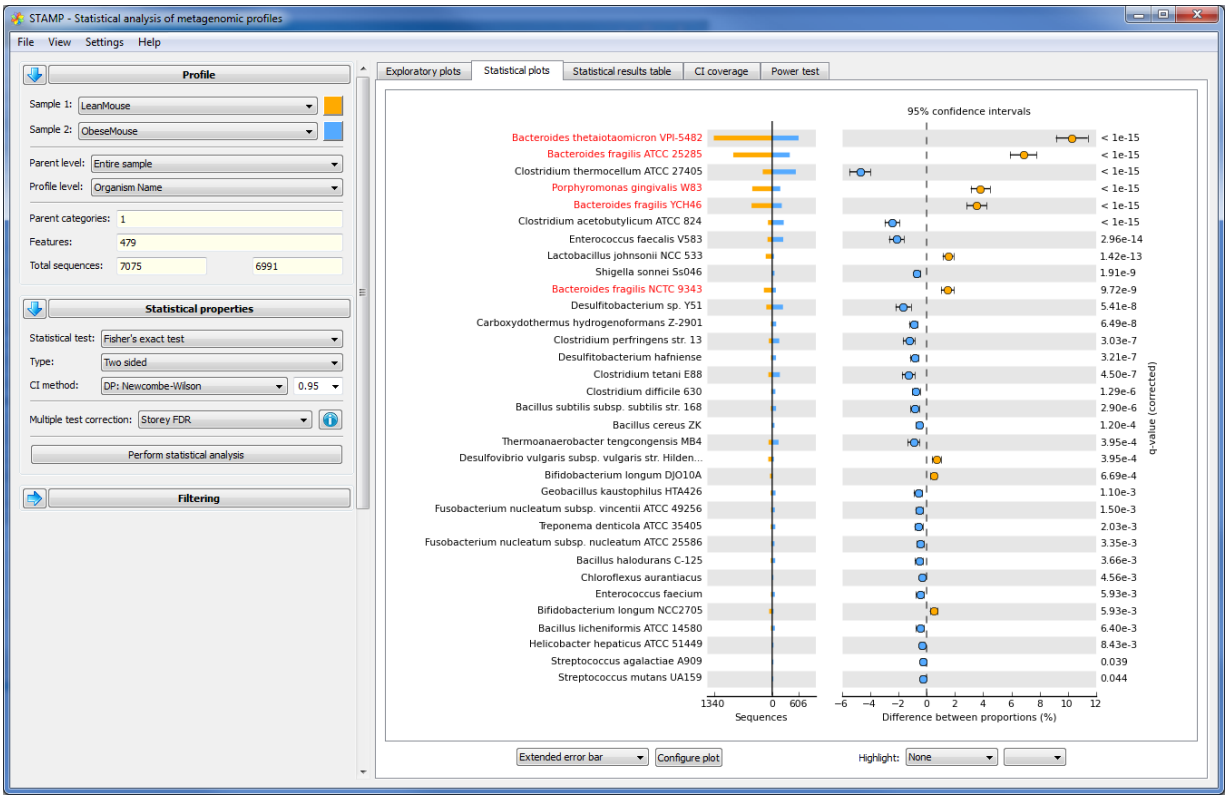


Figure 6: Example of STAMP, an open source bioinformatics software package.

As discussed earlier, bioinformatics is a field in which the open source model of software distribution is common. University researchers often write code for algorithms and then provide

the code for free distribution via the Internet. Such algorithm code may then be copied, modified, and used by anyone else. Such code comes with no guarantee of quality, accuracy, or ease of use. Figure 6 displays an example of open-source bioinformatics software. When open source solutions are insufficient, researchers in the bioinformatics and related bioscience spaces reply on software written and sold by companies. These companies' software is not freely distributed, but has much higher quality standards. Figure 7 displays an example of paid software. Companies in the bioinformatics space are mostly providers of software packages used by researchers. Such companies employ software programmers who write and debug code.

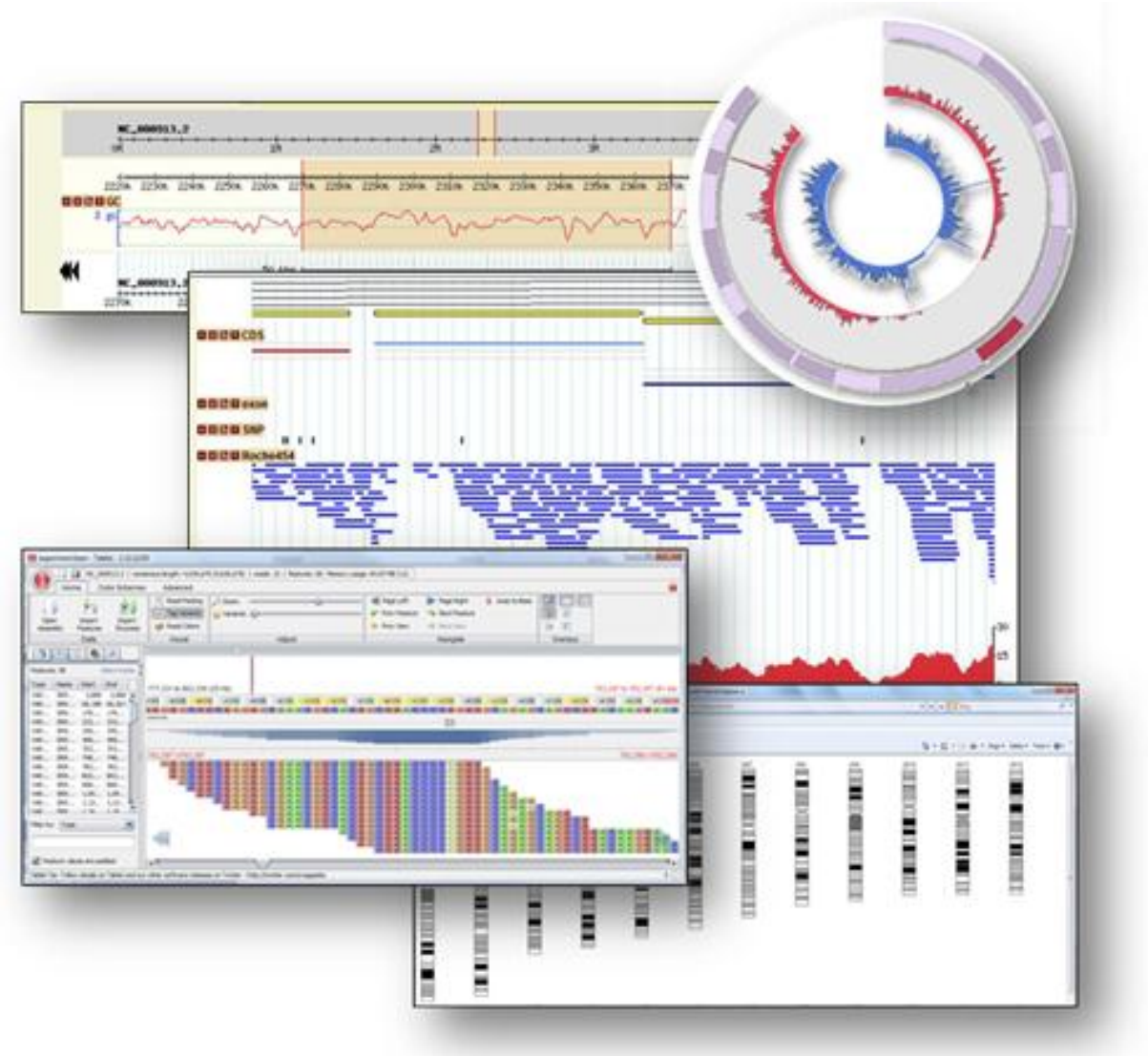


Figure 7: Example of NGS Collection a paid bioinformatics software package produced by Accelrys.

Most software companies in this space are relatively small, and they generally specialize in providing software packages written specifically for use in fields such as molecular biology, genetics, or computational chemistry. In the genetics space, such software is used to compute

and analyze genomes. In other spaces, such software is used to simulate experiments that would be prohibitively expensive to conduct. Some larger companies write software for more than one field of science. National software provider companies include Accelrys, a publicly traded company based in San Diego, CA that sells software packages with a number of different purposes intended for a number of different fields of science including bioinformatics. Accelrys has more than 700 employees and generates $155 million in revenue per year. Smaller, more specialized companies include BioDiscovery, a privately held company based in El Segundo, CA that provides software packages specifically for bioinformatics researchers. Software packages include "Nexus Copy Number," an application that helps researchers study variations in gene sequences. BioDiscovery also provides consulting services in this space. Most generalized and specialized bioinformatics companies are based in Europe or in California. Some non-software companies hire bioinformatics researchers and specialists. The Centers for Disease Control and Prevention as well as consulting companies such as Booz Allen Hamilton hire bioinformatics analysts that may use software created by the aforementioned software companies. Although Georgia has hundreds of software companies, iP2Biz was not able to identify one that claims to service the bioinformatics space. There is great potential for bioinformatics companies to be created in Georgia because of organizations such as Venture Atlanta and Atlanta Tech Village, which are home to dozens of software startups.

Thoughts for Georgia

Georgia has the resources to become a center of advancement in bioinformatics because Georgia has strengths in the areas that combine to form bioinformatics. Emory University is known for its strength in the biological sciences and medicine, Georgia Tech is a strong in computer science and computer engineering, and Georgia State University has attracted bioinformatics funding from the National Institutes of Health. Combinatory efforts and investments to establish Georgia as a player in the bioinformatics space alongside California would be appropriate. In spite of previous efforts however, Georgia is not presently recognized as a major hub for bioinformatics.