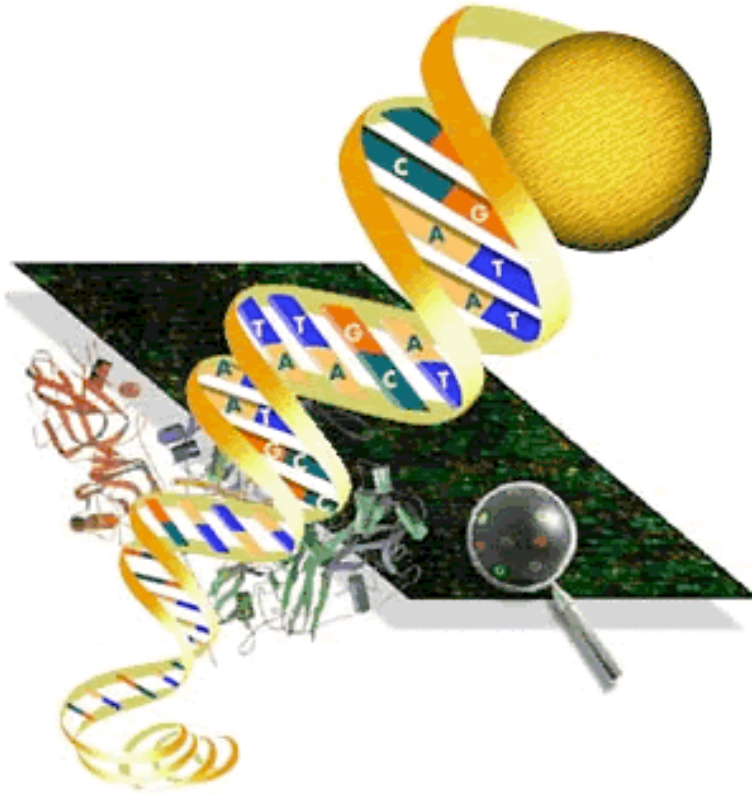
Explaining Genomics and Bioinformatics to High School Biology Students



Amanda Knowles[1], Sharon Schulze[2], Thomas Mitchell[3], David Haase[2], April Cleveland[2], and Ralph Dean[3]

[1]University of North Carolina at Pembroke
[2]The Science House, North Carolina State University
[3]Fungal Genomics Laboratory, North Carolina State University

Contents

Title: Explaining Bioinformatics and Genomics to High School Biology Students

Outline:

I. Introduction
   A. Definition of Genomics and Bioinformatics
   B. Thesis: Explaining Genomics and Bioinformatics to high school biology students.

II. History
   A. Genomics
   B. Bioinformatics

III. Genomics
   A. Types of Genomics
   B. Where used in society

IV. Bioinformatics
   A. What it is and how it works
   B. Where used in society

V. Relevance
   A. Genomics applied to their (high school biology students) lives
   B. Bioinformatics applied to their (high school biology students) lives

VI. Results
   A. Middle school students
   B. Biology Curriculum

VII. Activity for classroom
   A. How activity was devised
   B. Where to apply when teaching biology

VIII. Conclusion
   A. Further ways to use methods
   B. Additional information

Explaining Genomics and Bioinformatics to High School Biology Students

Amanda Knowles[*], Sharon Schulze[1], Thomas Mitchell[2], David Haase[1], and April Cleveland[1]

*University of North Carolina at Pembroke
[1]North Carolina State University; The Science House
[2]North Carolina State University, Fungal Genomics Laboratory

Even though technology and information is increasing in biological sciences, many students are being left behind. Two of the leading sciences have become genomics and bioinformatics; however, students are not being properly informed of the opportunities in these fields. Therefore, science teachers need ways to teach these subjects to their students. Activities for high school biology students on genomics and bioinformatics should be inquiry-based and relevant to their lives. Before activities can be completed, a brief history of the subjects is needed. In addition, the basic background information of genomics and bioinformatics is presented. Applications in science and in their lives is shown to allow students to understand relevance of genomics and bioinformatics. The activities devised began with students using a chromatogram to obtain a gene sequence of about fifty base pairs. After obtaining their gene, the student complete by hand a worksheet in which they match their gene to the one out of twenty-five example genes. This activity is devised to allow the students to fully appreciate that the computer can accomplish in a matter of seconds when humans take hours to complete. Afterwards, they use the actual bioinformatics computer search tool to seek a match to their gene sequence. Once they have found a close match, they report on the structure and function of their gene. These activities are devised to allow the students to appreciate what scientists do and perform the same tasks scientists do everyday in an actual lab setting.

Explaining Bioinformatics and Genomics to High School Biology Students

Introduction

The terms genomics and bioinformatics are not often heard in the high school biology classroom. When students do hear the terms, some will be given a definition to memorize for a test. An example would be "the definition of genomics is the study of the genome", or all the genes in an organism. In addition, the most common definition of bioinformatics is the convergence of biology and computer science to store, retrieve, and analyze data. These definitions do not explain what genomics and bioinformatics are or how they are changing modern science. As with areas of science, students need history, applications, and inquiry-based activities to fully understand the areas of genomics and bioinformatics. Therefore, activities have been devised in order to explain these subjects to high school biology students.

History

The revolution of genomics and bioinformatics began over a century before the terms were coined. In 1866, Gregor Mendel, "the father of genetics," published his findings of the heredity of pea plants. Another important first step happened in 1869, when DNA was first isolated by Friedrich Miescher. However, modern genetics did not begin until Carl Correns Hugo de Vries Erich von Tschermak verified Mendel's investigations in 1900 (History of Genetics). In 1910, Thomas Morgan proposed the theory of heredity through genes located on chromosomes using the *Drosophilia* fruit fly. To further the advancement of science in the area of genomics and bioinformatics, Arne Tiselius introduced electrophoresis in 1933 (Short History). The next progression toward genomics and bioinformatics was the discovery of the double-stranded DNA helix by Rosalind Franklin, James Watson, and Francis Crick between

1951 and 1953. Fredrick Sanger sequenced the first protein, bovine insulin, in 1955 (Short History).

Along with the advances being made in biology, there were also advances in computers which led to the areas of genomics and bioinformatics. One such advancement was the first integrated circuit. It was created at Texas Instruments by Jack Kilby in 1958 (Short History). More progress in biology came in 1961 when mRNA was isolated and studied by Jacob, Monod, and Lwoff (The Researchers). In 1966, a gigantic leap for geneticists occurred. Marshall Nirenberg and Gobind Khorana discovered mRNA occurs in triplets to form a codon, which codes for an amino acid. They discovered all twenty amino acids (History of Genetics). In 1969, Stanford and UCLA began to link their computers to create ARPANET (Short History). The Needleman-Wunsch algorithm was published in 1970 in order to compare sequences (Short History). Yet another advancement in biology occurred when Paul Berg created the first recombinant DNA molecule (History of Genetics). Other advancements in the 1970s included new methods of sequencing DNA and the founding of the first genetic engineering company, Genetech. Rapid advancement in biology and computer science occurred in the 1980s. In 1980, a 5386 base pair gene, which coded for nine proteins, was sequenced. In the same year, IntelliGenetics, Inc. was formed in California (Short History). Another step toward genomics and bioinformatics happened the following year, when the Smith-Waterman algorithm, used for sequence alignment, was published (Short History). Two important advances, FASTP algorithm published and Polymerase Chain Reaction (PCR) described by Karl Mullis, occurred in 1985. To further the genomics revolution, Thomas Roderick coined the term genomics and used it as the title of his journal in 1986 (Short History). Also, to fuel the bioinformatics revolution, the Department of Medical Biochemistry of the University of Geneva and the European Molecular

Biology Laboratory produced the SWISS-PROT database in which sequences can be compared (Short History). The first human genetic map was constructed in 1987 (The Human Genome Project). Three major progressions for bioinformatics and genomics occurred in 1988. The most publicized event, which began in 1988, was the Human Genome Initiative, to sequence the human genome. Another advancement that year was the foundation of the National Center of Biotechnology Information (NCBI) at the National Cancer Institute. A final step in 1988 was the publication of the FASTA algorithm, which is a fast approximation of the Smith-Waterman algorithm (Short History).

Compared to the advancements in the 1980s, the information in the 1990s is overwhelming. To begin the 1990s, a program which the public can use and somewhat understand is created (Short History). BLAST, Basic Local Alignment Search Tool is a program designed to compare data sets from many databases. The following year, a huge advancement in modern technology occurred. The World Wide Web was created by the CERN institute in Geneva. The same year genomics received a big boost when Craig Venter created expressed sequence tags (ESTs). Although there is much debate over when the term bioinformatics was first used, the first time it was published in literature was 1991. Craig Venter also founded The Institute for Genomic Research (TIGR) in 1992. Many other companies were formed during the 1990s to carry out research on genomics and bioinformatics, such as Gene Logic, Paradigm Genetics Inc., and GeneFormatics. The first bacterium genome, *Haemophilus influenzea,* was sequenced in 1995. Also the *Mycoplasma genitalium* genome was sequenced the same year (Short History). To add to the genome database, *Saccharomyces cerevisiae* was sequenced in 1996. In addition to many genomes being sequenced during the 1990s, many databases were being constructed to house all the information, such as the Prosite database and the PRINTS

database (Short History). The *E. coli* genome was sequenced in 1997. In 1998, two organizations were formed that are important to genomics and bioinformatics. The first was the Swiss Institute of Bioinformatics. The second, established by Craig Venter, was Celera. Many other genomes have been sequenced over the years, including the human genome, which was completed in 2001 and contained 3000 million base pairs. The work of geneticists and genomic-based companies has created many applications and uses to the public.

Genomics

Throughout the years, genomics has revolutionized to provide much information. As stated, genomics is the study of the genome, or all the genes that make up an organism. Many genomes of various organisms have been sequenced, from the very simple, like bacterium, to the most complex, human. Genomics can be broken down into three categories: structural, functional, and comparative. Structural genomics is defined as "the assignment of three-dimensional structures to proteomes (which define the protein compliment to the genome) and the investigation of their biological implications" (Structural Genomics). As the term implies, structural genomics is used to determine the structure of a protein. The structure of a protein is valuable for determining how the protein works and where it can bind to cause reactions. To determine the structure, scientists use Nuclear Magnetic Resonance, NMR, X-ray crystallography, or prediction by looking at known homologous proteins (Structural Genomics). One main application of structural genomics is drug design. In treating certain diseases, protein structure is important. Drug companies can design drugs to fit the protein, so it will not harm the organism.

Functional genomics is the "the development and application of global experimental approaches to assess gene function by making use of the information and reagents provided by

structural genomics" (Hieter and Boguski, 601). In simple terms, functional genomics deals with the function of the genes and proteins. Functions of single nucleotide polymorphisms and non-coding regions or "junk" DNA are some objectives researchers use functional genomics for.

The final category of genomics is comparative genomics. This category is the product of the other categories. Comparative genomics is exactly what the phrase implies; it compares genomes of different organisms. This includes comparing genes, genomes and proteins. Much of this field is concentrated in comparing organisms to humans; however, it does include all other comparisons. When comparative genomics is used, one can determine how closely related two species are or if two species have similar proteins. A main outcome of comparative genomics is determining the evolutionary tree of living organisms.

The implications of genomics in general can be seen everywhere. An obvious application is medicine. The medical field benefits tremendously from genomics. One major benefit is in treating genetically inherited diseases. Other applications of genomics include the food industry. Through genomics research, scientists have been able to produce better and longer lasting food products such as the FlavrSav tomato. Another implication involves making new catalysts for chemical reactions (Broderick). From an environmental standpoint, genomics may possibly help re-diversify endangered species. With all these implications, it is hard to fathom that a number of people, including high school students, think that the study of genomics has no effect on their lives. Students need to know how the study of genomics can and does affect their everyday lives.

Bioinformatics

As defined earlier, bioinformatics is the convergence of biology and computer science to store, retrieve and analyze data. However, the field of bioinformatics encompasses much more

than this simple definition. Data, in bioinformatics, pertains to nucleotide sequences and protein sequences. Bioinformatics does involve biology and computer science, but it also involves mathematics and statistics. Most of the mathematics and statistics are hidden in the computer science aspect of bioinformatics. The mathematics involved mainly deals with algorithms. An algorithm is a step-wise method for solving a problem. A statistical aspect of bioinformatics involves the e-value. The Expectation value is "an assessment of the statistical significance of the score" (Claverie and Notredame, 66). Also, the e-value is a determinate of how random a match is or how much chance is involved. Therefore, it is better to get a low e-value to have good results.

Most of the biology involved in bioinformatics deals with the input and output of data. Bioinformatics data is acquired through biological methods, mostly genomics and proteomics. In addition all bioinformatics methods are accomplished using computers. The main goal of bioinformatics is to figure out what biological data means. Therefore, the housing or storage of data is important. The data is stored in many databases, such as GenBank, which is currently housing over twenty-two billion sequence records (Description). The information in these databases is submitted by public and private scientists for comparison and is available to the public. It is possible for anyone to submit data to certain databases.

GenBank has many methods of submitting data. One of which is BankIt. This program is designed to accommodate submissions of 50 base pairs or more that are not complex. If a complex sequence is being submitted, another program, Sequin, should be used (Sequin). In addition, there are special submission programs for other kinds of data such as ESTs. Once data sets have been submitted and reviewed, they can be used to compare to other sequences.

There are many ways to compare data sets. One program, located on the NCBI website is BLAST. Using BLAST, many items such as nucleotide sequences or protein sequences can be compared. Not only does bioinformatics offer ways to compare data sets, but it also performs many other tasks as well. One such task is to predict sequence structure and function as well as assemble proteins into families (Bioinformatics Primer). After assembling families, bioinformatics can also help in establishing evolutionary relationships among organisms. In addition, bioinformatics allows scientists to discover single nucleotide polymorphism or SNPs. SNPs are differences in genomes that define individuals as unique. Another technique that bioinformatics has made practical is microarrays (Lynn, et al., 72). Microarrays produce massive amounts of data, too much to compute by hand. Another role of bioinformatics, is to locate protein-coding regions in DNA sequences (Claverie and Notredame, 26). Bioinformatics performs many tasks, and will continue to gain new assignments. The jobs of bioinformatics have applications in research and in everyday life.

Applications of Bioinformatics

The most widespread application of bioinformatics is in the medical field. In particular, the drug design process has become much faster. By using bioinformatics and not the trial and error method, the cost of drug design has decreased as well. A new method for prescription drugs also comes from bioinformatics. Known as pharmacogenomics, this field allows scientists to use bioinformatics to design and prescribe personal medications to individuals. Another advancement in the medical field relates to clinical diagnostics. Doctors, using bioinformatics, have been able to diagnose genetic diseases and other health problems more easily. Other applications outside of the medical field include use in plant pathology and criminal investigations. The use in plant pathology deals with comparing possible toxic genes to other

known toxic genes. Also, bioinformatics make it possible to be certain scientists are studying the pathogen's gene and not the host's gene when necessary. This application saves researchers much time and money. Bioinformatics can also aid in criminal investigations. For example, Dr. Christopher Basten, a Research Associate in Statistics at North Carolina State University, uses bioinformatics to compare canine hair and blood samples from crime scenes to the canine database to determine which breed of dog was at a particular crime scene.

Relevancy

Genomics and bioinformatics are relevant to high school biology students' lives in more ways than just because it is on the test. The relevance of these subject areas is predominant in the medical field. One medical relationship students have is the need for drug design. Most students will be prescribed something at one point in time. Knowing there is a way of making a drug for their individual needs will be helpful to students. Another medical application to students' lives is genetic disease diagnosis. Most students have heard of genetic diseases, such as Down syndrome or trisomy. With the technology available, people can be tested for diseases before they even begin to show symptoms.

Another area of interest and importance to students' lives, not related to the medical field, is genetically engineered products. There is much controversy over genetically engineered food that is being produced using genomics and bioinformatics. Students should understand that some of the foods they eat are produced through mutations. Another way in which genomics and bioinformatics are relevant to students is through understanding how the human genome project currently is or will be affecting their lives. Finally, knowing about genomics and bioinformatics opens up career pathways for students, including research and product development.

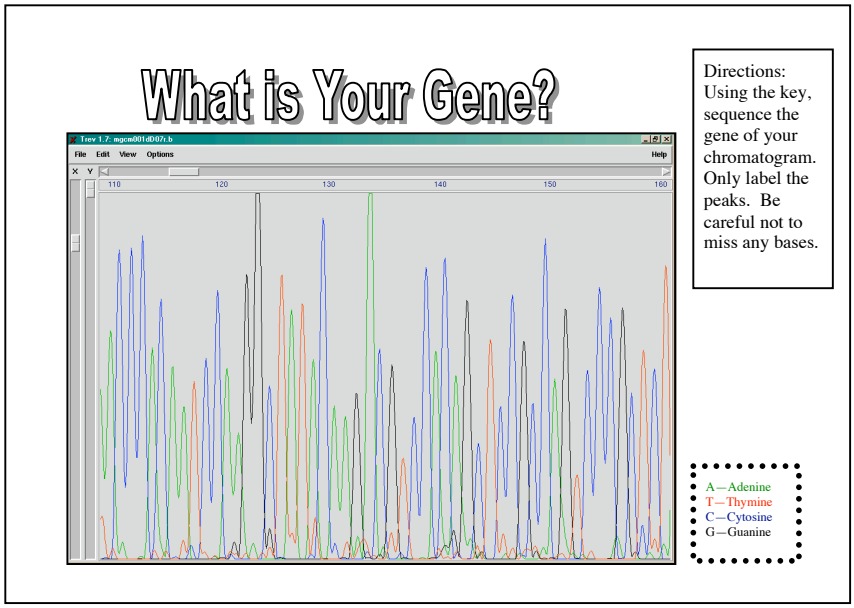


Figure 1

*Activities

The activities were created as part of the K-12 outreach partnership of the Fungal Genomics Laboratory and The Science House, both at North Carolina State University. The activities created were based on a guided inquiry method and assumed teachers have explained the history, basics, and applications genomics and bioinformatics. The chromatogram activity (Figure 1), named "What is Your Gene?" mainly deals with genomics after the wet-lab work. The wet-lab work includes procedures such as DNA extraction, PCR, and using a computer to sequence the gene. Students use the chromatogram to determine their gene sequence. In devising this activity, certain constraints had to be considered. First, the activity would be complete best if in color because each base is a different color*. The chromatogram should contain roughly fifty base pairs and each student's set should be different.

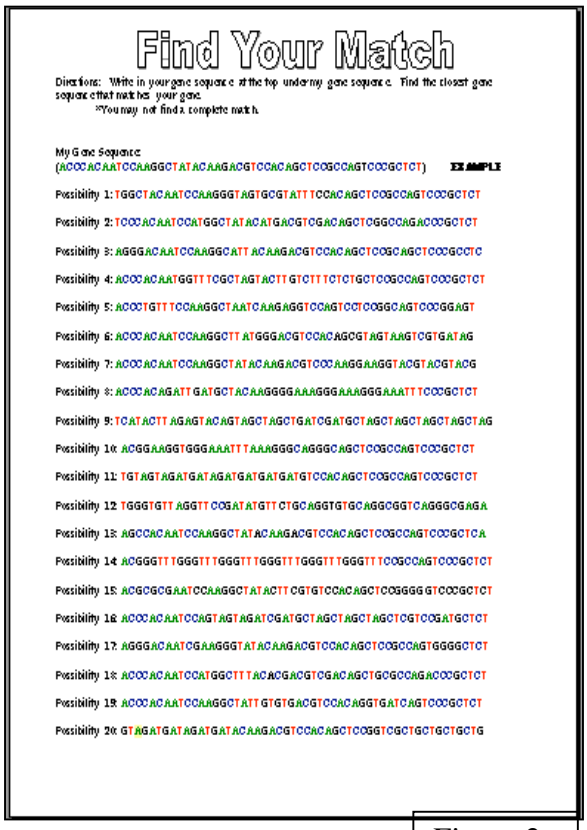


Figure 2

The continuing activity (Figure 2), called "Match Your Gene" is designed to allow students to use the gene they called in an attempt to find the closest match. They are provided twenty possibilities for comparison. Since each student has a different sequence, each student needs a different worksheet*. The worksheet is designed with the bases the same color as in the chromatogram. However, if color resources are not available, the activity can be completed in black and white.

The activity can be used in a classroom that has both academic level and honors level students. Honors students can be asked not only to match their gene, but to also find the complimentary strand which is located in the worksheet*. An additional application for honors level students in the worksheet is to calculate percent difference. Using the percent difference formula [% difference = ((number of bases in original gene – number of bases that match in second gene) / number of bases in original gene) x 100] students show how close their original gene is to the one they found. This can be done to help ex*plain the e-value in bioinformatics searches. The matching activity is designed to allow students to gain an appreciation for the advancements in computer technology.

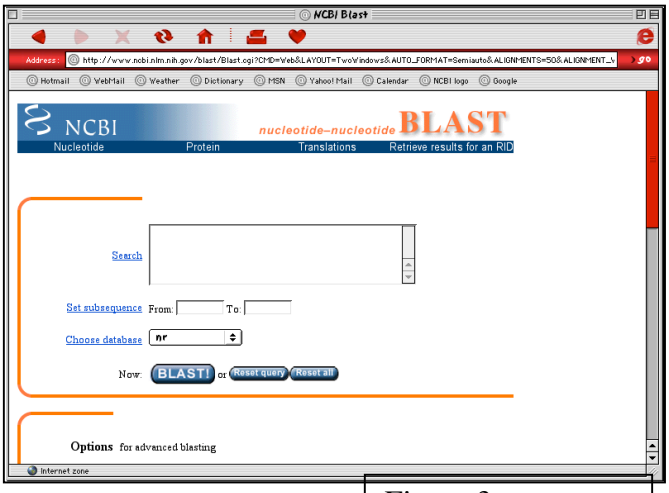


Figure 3

The next activity, a BLAST search, is designed to further aid in this task. In completing a BLAST search, students use the gene they called in the chromatogram activity. In order to use BLAST, students need a computer and connection to the Internet. Located on the NCBI website

(http://www.ncbi.nlm.nih.gov/BLAST), BLAST can be used by anyone. After typing in the web address, students click on the link named Standard nucleotide-nucleotide BLAST [blastn]. The next screen (Figure 3) that appears is where students type in their gene. After typing in the gene, students hit the BLAST! button. Another screen will appear, students click on FORMAT! and the search begins. Depending on the length of the gene and time of day searches can take from seconds to minutes. If the search runs for more than ten minutes, students may need to start over.

Figure 4

After completing the search, students receive a list of possible matches. A description of each component of the results' screen is shown in Figure 4. From the closest match, students report on their findings. The report can include many items such as the following: the e-value, where the database obtain the results, name of organism(s) where gene is found, when gene was first posted, what database used, and a PubMed article if available. The BLAST report activity can be completed in many different ways according to a teacher's style.

Results

Only one of these activities has been used with students. The chromatogram activity was completed by a group of middle school students during a seminar at the Fungal Genomics Laboratory. The students seemed to have no trouble completing the activity and finished in approximately ten minutes.

Although information provided and activities have not been tested with high school biology students, they still apply to the North Carolina Standard Course of Study. Specifically the information applies to Competency goal 2, "The learner will develop an understanding of the continuity of life and the changes of* organisms over time," and Competency goal 3, "The learner will develop an understanding of the unity and diversity of life" (Science Curriculum). Under Competency goal 2 is objective 2.04, "Assess the application of DNA technology to forensics, medicine, and agriculture" (Science Curriculum). All the activities can be applied to this objective.

In explaining genomics and bioinformatics, the teacher can teach the applications of DNA technology to medicine and agriculture by giving the provided examples. By completing the activities, students actually apply the technology. Objective 2.05, "Analyze and explain the role of genetics and environment in health and disease," can be expanded to include how genomics and bioinformatics play a role in health and disease detection (Science Curriculum).

Both genomics and bioinformatics can help in executing the objectives 2.06, "Examine the development of the Theory of Biological Evolution" and 3.01, "Relate the variety of living organisms to their evolutionary relationships," which are meant to explain evolution and evolutionary relationships (Science Curriculum). A final objective in which genomics and

bioinformatics can help explain is 3.02; "Classify organisms according to currently accepted systems" (Science Curriculum). Genomics and bioinformatics work to establish evolutionary classification systems.

Conclusion

To produce further results, The Fungal Genomics Laboratory will continue to use the activities in workshops offered by The Science House. Additionally, the activities will be available at The Science House website (http://www.science-house.org). In addition to using the information and activities in the biology curriculum, they can also be used in the mathematics curriculum, especially Discrete Mathematics and Advance Placement Statistics*. The activities provide an ideal opportunity to integrate mathematics and science education. The activities are also suited to collaboration among computer science and biology teachers. In the Computer/Technology Skills Curriculum, Competency goal 3 recommends integration with science. In collaborating with a biology teacher, a computer/technology skills teacher could design a lesson on genomics and bioinformatics that would fit under Competency goal 3. In society today, the uses of technology are rapidly increasing and improving. Teachers need to work to stay informed on new technologies to be able to inform students of the many opportunities available. Through explaining genomics and bioinformatics to students, teachers give students a head start into the opportunities available. The information and activities provided can help teachers accomplish this task.

Works Cited

Bioinformatics Primer. 2003. University at Buffalo Center of Excellence in Bioinformatics. 15 July 2003. <http://www.bioinformatics.buffalo.edu/ current_buffalo/primer.html>.

Broderick, Andrew. Genomics. 2003. SRI Consulting Business Intelligence. 14 July 2003. <http://www.sric-bi.com/Explorer/GEN.shtml>.

Claverie, Jean-Michel, PhD., and Cedric Notredame, PhD. Bioinformatics for Dummies. New York: Wiley Publishing, 2003.

Description of BLAST Services. 2003. NCBI. 15 July 2003. <http://www.ncbi.nlm.nih.gov/blast/html/BLASThomehelp.html>.

Hieter, Philip, and Mark Boguski. "Functional Genomics: It's All How You Read It." SCIENCE, 278. (1997). 601-602. 18 Jun 2003. <http://www.sciencemag.org>.

History of Genetics Timeline. Jo Ann Lane. 1994. Access Excellence. 10 July 2003. <http://www.acessexcellence.org/AE/AEPC/wwc/1994/geneticstlm.html>.

Human Genome Project Timeline, The. 2003. 11 July 2003. <http://www.genome.gov/Pages/Education/Kit/main.cfm?pageid=1>.

Lynn, David J. MSc., Andrew T. Lloyd, PhD., and Cliona O'Farrelly, PhD. "Bioinformatics: Implications for medical research and clinical practice." Med. Clin exp., 26.2 (2003). 70-74.

Researchers, The: Historical Highlights: The Pioneers. 8 May 2003. Canadian Museum of Nature. 10 July 2003. <http://www.nature.ca/genome/03/e/03e_30_e.cfm>.

Science Curriculum. 2003. NC Public Schools. 3 July 2003. <http://www.ncpublicschools.org/curriculum/science/biology.html>.

Sequin—A DNA Sequence Submission and Update Tool. 28 April 2003. NCBI. 15 July 2003. <http://www.ncbi.nlm.nih.gov/Sequin/index.html>.

Short History of Bioinformatics, A. 2002. Allen B. Richon. Network Science. 23 July 2003. <http://www.netsci.org/Science/Bioinform/feature06.html>.

Structural Genomics: Protein sturcture determination, classification, modelling and docking. Arthur Leak and Manuela Helmer-Citterich. 2003. Functional Genomics.org.uk. 14 July 2003. <http://www.functionalgenomics.org.uk/ sections/programme/structural.html>.

***Notes to Teachers**

*A set of thirty chromatograms can be obtained by emailing Amanda Knowles at ajk001@uncp.edu.

*The teacher can use the same matching worksheet for all students. The students just cannot use their own gene in completing the exercise. They would have to use the example provided. Additionally, this worksheet can be obtained by emailing Amanda Knowles at ajk001@uncp.edu

*The answers to the worksheet provided are number 13 is the closest match and number 12 is the complimentary strand.

*The chromatogram activity should take about ten minutes to complete. In addition, the matching should take about twenty-five minutes to complete.

*The BLAST search time depends on Internet speed and length of sequence.

*For Discrete Mathematics, the activities apply to objectives 2.01 a and b, and 2.02.

*For Advance Placement Statistics, the activities apply to objectives 3.03 and 3.05.

Acknowledgements

Michael Clinkscales, Fungal Genomics Laboratory, North Carolina State University

Judy Day, The Science House, North Carolina State University

Joyce Hilliard-Clark, The Science House, North Carolina State University

Phyllis Hilliard, Fungal Genomics Laboratory, North Carolina State University

Bonnie Kelley, University of North Carolina at Pembroke

Michael Smith, The Science House, North Carolina State University

Cherrie Tchir, The Science House, North Carolina State University

Michael Thon, Fungal Genomics Laboratory, North Carolina State University

Carol White, Fungal Genomics Laboratory, North Carolina State University

National Science Foundation