

CCSC

Consortium for Computing Sciences in Colleges

Southeastern Region

**26th
Annual
Southeastern
Conference**

**Student Research
Contest**

**Extended
Abstracts**

**November 2 and 3, 2012
Southern Polytechnic State University
Marietta, Georgia**

Table of Contents

Using Arena Simulation Software to Test the Efficiency of GGC’s Dining Facilities	4
<i>Kyle Raffa</i> <i>Georgia Gwinnett College</i>	
Restoration and Analysis of Apollo Lunar Data.....	6
<i>Missy Gaddy</i> <i>Wofford College</i>	
Simulation Model to Study College Campus Wireless Service Sustainability.....	8
<i>Mohamed Lahjmaydi</i> <i>Georgia Gwinnett College</i>	
The Search for a Causative Mutation of Osteochondritis Dessecans and Evaluation of RNA Sequencing as a Method of Mutation Detection.....	10
<i>Alissa Williams</i> <i>Wofford College</i>	
Comparative Transcriptome Analysis to Identify Genes Regulating Elastogenesis.....	12
<i>Sharon Guffy</i> <i>Wofford College</i>	
A Ubiquitous Non-Binding Fitness Application for Mobile Devices Using Body Sensors	14
<i>Katie Moore</i> <i>Stetson University</i>	
A Humanoid Robot to Counter Childhood Obesity.....	16
<i>Daria Jordan</i> <i>Spelman College</i>	
Bioinformatics Tutorial Using Python.....	18
<i>Jazette Johnson</i> <i>Spelman College</i>	

Solving Bioinformatics Challenges Using a Command-Line Interface.....19
Zazie Lumpkin
Spelman College

Parallel Branch and Bound: Applying an Asynchronous Multi-pool Approach to
Cyclic Best First Search.....20
Nartezya Dykes
Spelman College

Using Arena Simulation Software to test the Efficiency of GGC'S Dining Facilities

Kyle Raffa

Georgia Gwinnett College

Lawrenceville, GA

kraffa@ggc.edu

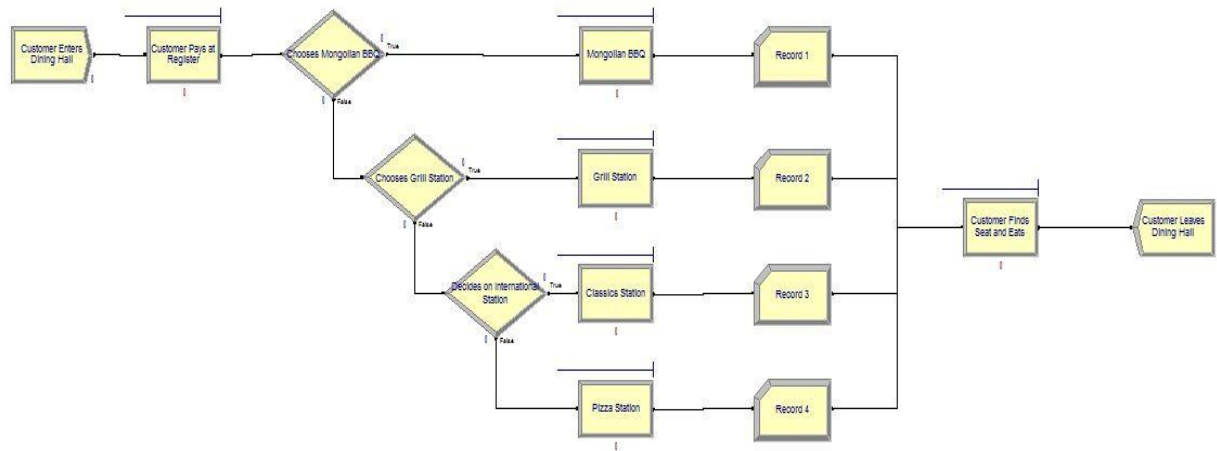
Faculty Sponsor: Dr. Anatoly Kurkovsky, Computer Science

In the last three years Georgia Gwinnett College (GGC) has increased its student population from 3,000 students to nearly 9,000. While this rapid growth has been of great benefit to the school financially, it has created several obvious logistical problems as traffic on the roads surrounding the school has increased dramatically, and classroom space is becoming scarce. There is also the fact that the school's dining facilities are becoming overcrowded, which could present a serious problem to the school's growth in the future.

The project is being completed over two semesters, with the first half being completed in the spring of 2012, and the second half during the fall. The goals during the first semester were to identify a problem facing GGC, and to come up with a conceptual model of the problem. The sustainability of GGC's dining facilities seemed like a good area for further study, and so these facilities became the subject domain for my project's problem. The conceptual model created during the first semester also helped identify the input, output and control variables that would be used in the second semester model.

In order to build a working model of the school, Rockwell's ARENA simulation software was deemed the best tool available to complete this task. ARENA was chosen for this project for three reasons: first, it is incredibly flexible and can be easily manipulated to create a model at a high or low level of granularity. Second, it is very user friendly and only takes a few weeks to learn the basics of its operation, thus allowing the user to spend more time testing the model instead of wasting time consulting the manual trying to get it to perform correctly. Finally, ARENA specializes in modeling queuing systems, and can track an entity's time traveling through a system. Thus, ARENA is perfect for modeling the lines for the various dining facilities at GGC.

The goals of the fall semester are to first construct the actual model in ARENA. Both a 2-D and 3-D version will be made, the 2-D for general modeling purposes and the 3-D for specifically modeling the school's dining hall. The second goal is to analyze a series of test scenarios to help identify what the maximum efficient operating capacity is for the school's dining facilities. The third and final goal is to develop additional scenarios for future study by GGC students taking simulation and systems analysis courses.



Restoration and Analysis of Apollo Lunar Data
Missy Gaddy, Wofford College
David Williams, NASA Goddard Space Flight Center

The Lunar Ejecta and Meteorites (LEAM) Experiment on Apollo 17 was designed to measure hypervelocity particles that collide with the surface of the Moon. The main goal was to measure the speed, direction, kinetic energy, and momentum of these particles. The instrument had three 4 in. by 4 in. sensors, one located on the Up, East, and West side of the instrument. Initial data analyses in the 1970s yielded unusually high numbers of low-velocity impacts during the passage of the terminator, or the boundary between the sunlit side and the dark side of the moon. Around 60 hours before sunrise, a large number of events would register on LEAM sensors and continue between 30 and 60 hours after sunrise.

After the initial analyses in the 1970s, the LEAM data was used as support for dust levitation theories. The basic idea of these theories are that dust is lofted off of the surface of the Moon during the passage of the terminator due to changing electric fields. The fact that LEAM saw a lot more dust activity at the terminator supports this theory. Until last year, LEAM was seen as a strong piece of evidence for the dust levitation theory. However, in 2011, a question was raised about the reliability of the LEAM data. According to some researchers, the LEAM instrument may have registered nonexistent events during periods of electrical activity. Many of the other Apollo experiments had heaters, and the possibility was raised that electrical events when they were turned on or off may have caused the LEAM sensors to register a hit. Also, the LEAM instrument was the closest experiment to the Apollo central station (about 7.5 meters), so it could be susceptible to noise from electrical events on the central station as well. The current analysis was designed to conduct correlational studies between the LEAM events and electrical events on other instruments to see if electrical interference was a possibility.

In the initial analyses, scientists counted the number of events registered on the LEAM sensors during a three-hour period. For the current analysis, graphs showing activities for a three-hour period is not accurate enough. Better resolution is needed to see exactly when events occur and if there is a housekeeping event immediately preceding the registered event. To do this, graphs were created with a single dot to represent an event. The x-axis represented time, in days, and the y-axis was a single position to mark an event.

At first, computer code was written to find changes in the accumulator count for a given sensor. If the accumulator value changed, the time of the change was recorded and then graphed. The data contained frequent blips, though, or spurious signals that would quickly return to the original count. To prevent the program from counting these as accumulator counts when they clearly are not, criteria were put into place to examine the values around the change and verify that it is indeed a true hit. If the value of the channel changed, the program looked at several of the values before and after the change to make sure there was not a single blip in the data.

Surprisingly, very clear patterns in the front film of the sensor were found in the graphs after running this program. On the UP sensor graphs, it was found that the points were equidistant from each other, and it was confirmed that these events were happening exactly 0.6415 days (15.4 hours) apart from each other. Upon closer inspection, we realized that each event was actually registered as two events; the resolution of the graph had prevented us from seeing these multiples at first. These multiples happen exactly 0.0025 days (which is roughly 3.6 minutes) apart. So the Up sensor had two patterns: a three-minute multiple and a 15-hour cycle. On the EAST sensor, clear patterns

appeared as well. On the front film of the EAST sensor, it was found that the vast majority of the hits occurred every 0.0025 days (roughly 4.5 minutes) apart. The “event” always occurs at the beginning of a frame, and it changes every six frames. The rear film did not show these patterns.

After conferring with Mihaly Horanyi from the University of Colorado about these unusual occurrences, it was decided to look at the PHA scores instead of the accumulator counts. However, because of time limitations, this has not been performed as of this writing. Kent Hills, a coinvestigator for another Apollo mission, suggested that the Apollo scientists intentionally included the patterns in the accumulator. LEAM engineers may have put a test pulse at the regular intervals to make sure that the accumulator was working properly. There is no documentation that we have seen of test pulses thus far, but that may be a possibility. Further examination of these patterns may reveal electrical interference, and additional channels of Housekeeping data can be studied to continue the investigation for spurious signals in the LEAM data.

Simulation Model to Study College Campus Wireless Service Sustainability

Author: Mohamed Lahmaydi

Georgia Gwinnett College

Lawrenceville, Georgia

mlahmayd@ggc.edu

Faculty Sponsor: Dr. Anatoly Kurkovsky, School of Science and
Technology

Abstract

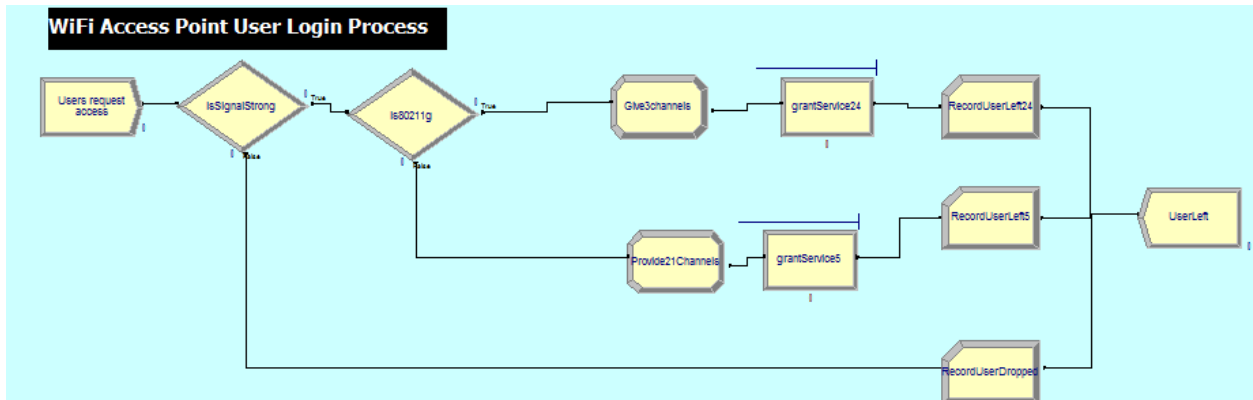
Georgia Gwinnett College, (GCC) the campus of tomorrow, has become a mecca, an attraction to many students due to its great, innovative curriculum. The college demographic has grown exponentially, getting faculty, staff, and students scrambling for resource that cannot sustain the high demand. The inevitable deadlock requires college administrators, managers to look for alternative solution to increase the capacity in many areas, namely efficient parking, proper class rooms scheduling, wireless access point, and many more tools necessary for educating students. The college requires students to use their own laptops because most classwork can conveniently be done online. Hence, every time a large classroom gets in session some users cannot sign in to Wi-Fi network because the high request triggers an unintended denial of service attack.

Within this project, we want to explore further Wi-Fi availability and formulate some guideline for IT support team and managers to help them use the tool and come up with proper solution to alleviate the bottleneck. The simulation technology could save them time and money, and help them react faster to the ongoing growing pain. The conceptual model of our system comprises internal component within the system boundary, such as wireless access point, controller and bridge, and the wired bandwidth infrastructure that supports Wi-Fi cloud. Also the model consists of external users that benefit from the system, such as students, faculty and staff. The wireless spectrum frequency has limitation, and the ranges of frequencies do not provide adequate bandwidth and throughput to support a large number of users crammed in small area.

First, in this study we are translating an established process of physical site survey, walking several buildings, going round from room to room to find the perfect spot where access point are going be installed. Also there are many commercial and open source wireless simulation models such as OPNET Modeler Wireless Suite, openWNS, and MIMIC Simulator from Gambit. By using these tools, and applying the same principle, we can custom build a virtual GGC campus using the new Arena 3D modeling tool to establish a visual snapshot of the network representing the front end of the system. In the back end, Arena tool will provide a thorough statistical analysis using its powerful high level language SIMAN. The basic idea of the study is to find certain limits, such as the maximum number of access points, and the number of users per location.

The variables identified in the system of our subject domain consist of input, output, and control elements. The data of every device containing a wireless network adapter is considered to be an input variable. This could be the faculty's, the staff's, or the student's smart phone or notebook. These devices will request services from a remote server such as login or downloading homework during the class hours. The output variable consists of the node response or page download time. The control variable consists of the number quality of service and traffic shaping according to importance and sensitivity of application data to latency. At this point of the project

the model run but has not been finalized yet. It is at the data analysis stage. It will require more time to run Arena input and output analyzer and study the statistical data. The deliverable due date is at the end of this semester.



Arena Data Flow Diagram for Wireless Access Point accepting 802.11g and 802.11n protocols.

The Search for a Causative Mutation of Osteochondritis Dissecans and Evaluation of RNA Sequencing as a Method of Mutation Detection

Alissa Williams, Wofford College

Summer mentor: Dr. Matthew L. Warman, Director of Orthopaedic Research Laboratories at Harvard Medical School-Children's Hospital Boston, Howard Hughes Medical Institute Investigator

Faculty reference: Dr. Angela Shiflet, Wofford College

Osteochondritis dissecans (OCD) is a disorder of the joints; more specifically, it affects the articular cartilage (cartilage of the joints) and its underlying bone. These tissues essentially break off in small pieces and float around in the joint cavity. The fragments cause pain and can further damage the remaining structures of the joint. OCD is relatively rare in humans, and a large majority of cases involve the knees, the elbows, or both. In some patients, particularly in adolescent athletes, overuse of the joints probably contributes to the development of this disorder. However, given that many OCD patients have immediate relatives with similar symptoms, it is likely that there is genetic predisposition at work. With this in mind, we began investigating whether there was a genetic cause of osteochondritis dissecans.

There are many ways in which we can query an individual's genome in search of a disease-causing mutation. In the realm of genetic mutation detection, RNA sequencing is a relatively new technology. As opposed to more traditional methods, which generally provide a profile of an individual's entire genome, RNA sequencing represents only the genes transcribed by a particular group of cells; that is, RNA sequencing allows for a focus on the genes actually translated into protein by a specific sample of cells. Due to the reduction in the number of genes sampled, RNA sequencing is cheaper and less data-intensive than customary genome sequencing. For these reasons, my laboratory began an investigation of a possible genetic cause of osteochondritis dissecans with RNA sequencing.

Thus, a large part of my summer consisted of analyzing RNA sequencing data from OCD patient samples. When I arrived, the RNA from each of eight patient samples had already been extracted and sequenced using Illumina technology. The RNA sequence data for each individual was contained within two fastq files. These fastq files contained the nucleotide sequence of each patient's RNA transcripts. The first step was to merge the two fastq files into a single sam file. This was accomplished using SAMtools, a software package created specifically to interact with sequence data. Next, each sam file had to be converted into the equivalent binary form (a bam file). Each of these bam files was sorted using SAMtools.

Using the final bam files, coverage values were calculated for each gene in each sample. Coverage data informs us about how well each gene was surveyed by Illumina sequencing; a gene must be well-represented, or well-covered, in order to draw conclusions about its sequence data. This step also allowed normalization for the size of the gene. Genes can range from a few dozen base pairs long to a few thousand base pairs long. The more base pairs a gene contains, the greater the chance that the gene is overrepresented in sequence data purely by chance. Thus, the coverage data was standardized to reflect the relative size of each gene. The algorithm used for these calculations was provided by BioConductor, a set of open-source software packages designed specifically for bioinformatics.

After coverage data was calculated, each of the modified bam files was analyzed via VarScan and ANNOVAR. VarScan is a software package that compares sample sequence data to a reference genome—in this case, the human genome—and detects any deviation from that reference. For instance, the sample may contain a different nucleotide at position n than the reference genome. The final output of this software was a list of variant nucleotides found in each sample's sequencing data.

The variants reported by VarScan were then sent to ANNOVAR for further analysis. ANNOVAR is used to annotate, or determine information about, deviations in sequencing data. ANNOVAR cross-referenced the reported variants with a database that contains well-documented and non-disease causing variants common to the human population; because we're looking for a variant that

causes a disease, these previously recorded variants were initially dropped in further analysis. The remaining variants, all of which were novel and possible candidates for a new, OCD-causing mutation, were run a second time through ANNOVAR to determine which were found in protein-producing regions of the genome. Protein-producing regions are of particular interest because proteins carry out nearly every biological function.

A large portion of the project was merging each of these files into one master file. This master file included coverage data for each gene, as well as an indication of which genetic variants were present in which samples. The file had to contain all relevant information about each variant while avoiding duplicated entries. The code I wrote to merge the files was in R, a language widely used in bioinformatics due to the ease of creating graphs and tables with its built-in functions.

The final master file contained 28,912 unique and novel genetic variants, which were sorted to filter and prioritize data. Analysis began with 561 variants of interest.

Each of these variants was systematically analyzed. All were visualized using the Interactive Genomics Viewer to eliminate those that were merely the result of sequencing errors. All that passed this initial screen were cross-referenced with other databases that contain known variants. After several levels of analysis and research, none of the 561 variants stood out as an OCD-causing mutation. Next steps will include analyzing variants originally ignored by our stringent filters. A recent paper published in *Nature* by the ENCODE Project Consortium (2012) highlights the importance of this next step, as the project found that segments of DNA previously dismissed as “junk” actually perform crucial functions.

Though no mutation of interest was detected, it is possible that one exists. This project encompassed more than just mutation detection; with this data, we were able to evaluate the usefulness of RNA sequencing as a sequencing method. We noticed throughout our analysis that even within the same gene, there was variability in the representation, or coverage, of base pairs. Some portions of the genome, even in the most abundantly expressed transcripts, were grossly under-covered, and sometimes completely absent. This means that even a fully penetrant OCD-causing mutation could have been easily overlooked if it was in an underrepresented portion of the genome. Even if analysis of all of the RNA sequencing data ends with no causative candidate, it doesn't eliminate the possibility of its existence. This brings us to the importance of multiple assays to confirm results, regardless of the initial outcomes. It would be worthwhile to query the sample genomes using different methods, particularly those that cover the entire genome. Our analysis focused on the most likely portions of the genome in which a mutation would exist; however, a disease-causing mutation could technically be located anywhere in the genome.

Comparative Transcriptome Analysis to Identify Genes Regulating Elastogenesis

Sharon Guffy

Wofford College

Spartanburg, SC

guffysl@email.wofford.edu

Research Mentors: Erin Pardue, Dr. Jeremy Barth, and Dr. W. Scott Argraves

Department of Cell Biology and Regenerative Medicine, Medical University of South Carolina

Faculty Advisor: Dr. Angela Shiflet

Department of Computer Science, Wofford College

Versican is a hyaluronan-binding proteoglycan found in the extracellular matrix which has four splice variants. While the V0, V1, and V2 variants all contain at least one chondroitin sulfate-binding region, the V3 isoform has no binding sites for these side chains, which are thought to interfere with elastic fiber deposition (Wight 2002; Ikeda et al. 2009). It has been shown that arterial smooth muscle cells transduced with V3 show increased elastic fiber formation and increased levels of tropoelastin synthesis (Merrilees et al. 2002), and a microarray analysis of V3-transduced cells shows increased expression of tropoelastin, fibulin 5, fibulin 2, and other genes involved in elastogenesis. In this study, genes which showed significantly different levels of expression in V3-transduced cells were first examined in two gene expression profiling studies examining mouse lung repair after damage by naphthalene injection and mouse skin repair after a punch biopsy. A group of twenty-nine genes including those encoding tropoelastin, fibulin-2, neuraminidase-1, ADAMTS9, tenascin-C, and decorin were consistently either upregulated or downregulated in all three models of elastogenesis, suggesting that these genes may be important in the process of elastic fiber formation. The relative expression of these genes in the V3 model of elastogenesis as compared to control cells is shown in Table 1 below.

To further test this hypothesis, qPCR was conducted on genes of interest (indicated by asterisks in Table 1) using cDNA prepared from the lungs of wild-type mice at several stages of postnatal development. The expression patterns of these genes in the developing lung were then compared to those of tropoelastin, fibulin-5, and fibulin-2, which are known to be involved in elastogenesis. Since the expression of genes identified in this manner shows a consistent correlation with elastic fiber formation in multiple models, they may serve as new targets for investigation in the study of elastogenesis.

Table 1: Fold change values from microarray transcriptome analysis of V3-transduced rat arterial smooth muscle cells for genes which were consistently upregulated or downregulated in all three of the examined studies. Asterisks indicate genes that have been tested in RNA from developing mouse lungs.

<i>Upregulated Genes</i>		<i>Downregulated Genes</i>	
Gene Symbol	Fold Change	Gene Symbol	Fold Change
Adamts9 *	9.01	Rgs2	-9.28, -8.19
Neu1	8.70, 4.36	Cd55 (DAF)	-8.72, -6.70
Lmcd1	8.05	Adh1	-3.25
Eln *	3.30	Dcn *	-2.95
Cpne8	2.56	Enpp2	-2.83
Notch3	2.55	Col14a1	-2.73, -1.94
Tnc *	2.52	Nt5e	-1.87
Pkig	2.47	Tcea3	-1.85
Csrp2	2.37	Acsl4	-1.65
Cmklr1	2.18	Bbs2	-1.64
Fbln2 *	1.95	Ncl	-1.64
Sdc3	1.92		
Gpx7	1.78		
Itga3	1.74		
Cndp2	1.71		
Sat1	1.62		
Rcan1	1.62		
Bmp1	1.56		

This work was supported from NIH grant HL095067, NSF EPSCoR RII grant EPS-0903795 and a grant from the SC EPSCoR Space Grant Consortium NNX10AM76H.

A Ubiquitous Non-Binding Fitness Application for Mobile Devices Using Body Sensors

Katie Moore

klmoore@stetson.edu

Stetson University

421 N. Woodland Blvd.

Deland, Florida, 32723

Faculty Sponsor:

Dr. Hala ElAarag

helaarag@stetson.edu

In our modern society, the need for physically intensive tasks and activities has all but vanished. This growing sedentary lifestyle is a direct result of the modern conveniences that may help make our lives easier, but at the price of our physical health. With the sudden decline in the society's need to be intensely physically active, people are becoming less physically fit, and consequently, less healthy than in previous years when people as a whole were more physically active. The dangers of this sedentary lifestyle are becoming more and more apparent, as people are gaining an unhealthy amount of weight and developing health problems on a daily basis. The obvious solution to this problem is for people to exercise more often, but this can become a challenge in itself. Where in the past a person's daily activities would suffice to prevent fitness problems, with the advent of modern technology, it has become unnecessary to make the extra effort, making exercise an additional chore. It is now up to the individual to motivate themselves to exercise, and even if they can do that, they run the risk of pushing themselves too hard or not enough; which could result in injury, or making the work-out ineffective.

Our goal in this project is to develop a motivational fitness application that would not only be engaging, but also be able to provide an effective guided exercise for the user. By implementing an application on a smart phone it not only allows the user a ubiquitous medium in which their fitness data is private, it also provides exercise instructions uniquely designed for them; advantages not found in other mobile applications to our knowledge.

Our fitness application is developed for an Android smart phone. Using a wearable Zephyr HxM BT Bluetooth heart rate monitor, the application gathers the user's biological data, delivering the data over the phone's Bluetooth to the application. This data is then used to manipulate a 3D environment developed using OpenGL ES 2.0, and ensure that the user is performing at their target heart rate. The smart phone's screen provides the user with instructions as well as the gathered data along with the 3D environment to allow the user to see in real time how they are performing during their exercise routine.

One of the more advantageous features of this fitness game is the freedom the users have in its use. Because the application only requires the smart phone and the wearable Bluetooth heart monitor, things someone can easily bring with them when they exercise, as well as only requiring the user's heart rate, the user is free to exercise on any piece of equipment they prefer, so long as they can safely look at the screen to receive instructions. This freedom of choice has not been emphasized in other mobile fitness applications to our knowledge, but we believe that it provides a distinct advantage for the user to allow them the ability to exercise as they see fit. The biological data provided on the screen presents another advantage if the user chooses to use manual workout equipment, for example if the user exercised on a manual treadmill, the application could display data such as distance and running speed. This allows the user to have easy access to biological information during exercise that they normally could not obtain except through expensive exercise equipment.

We hope that the ease of use of both the application and the heart rate monitor, and by seeing their biological data, as well as being able to manipulate the character in the environment with their heart rate, users will be motivated to exercise more often and become more physically fit.

A Humanoid Robot to Counter Childhood Obesity

Daria Jordan

Spelman College

Faculty Member Reference: Jakita Thomas (Jthoma41@spelman.edu)

Childhood obesity affects one-third of all children, thus making it an extremely important societal issue. Through an unconventional approach, BeyonceBot, a Nao humanoid robot may be the best method for ending or slowing the childhood obesity epidemic in the future. A humanoid robot will be cutting edge technology to work alongside with children to help decrease the risk of becoming obese or help the child to lose weight. The development and use of humanoid robots in the near future is of importance because the lack of similar technology in America compared to nations such as Japan. Human-to-robot interactions can give a child a sense of having a new friend. Dancing, exercising, and playing sports are the most common ways that the humanoid robot could interact with the child to help lower the growing number of children that are obese in the United States. Therefore, this project will examine all of the following: what interventions are currently being used to reduce the childhood obesity epidemic? Does the change in a child's diet affect their health? Could the use of an exercising humanoid robot increase physical activity for children? Would the use of an exercising humanoid robot be effective? How an exercising humanoid robot would actually work?

The approach included partnering with the Morehouse School of Medicine to develop a calisthenics plan and partnering with the Boys & Girls Club to implement the dances and exercises with children. The callisthenic work out plan would be as follows:

Warm-up/Free Dance (5 min)

Jumping Jacks (2 min)

Jog in Place (3 min)

Alternate Leg Kicks (3 min)

Lunges (2 min)

Squats (2 min)

Cool Down (3 min)

The plan includes eight participants from the Boys & Girls Club wearing pedometers to track their fitness. This method is just for initial testing purposes; the ideal circumstance is that there is one robot per child or per household. However, due to the cost of the robot the more feasible plan is one robot per classroom. The use of sensors on the robot such as bumpers, tactile, hand and chest are all used to aid the robot in recognizing the child. Along with these, vision is a more important tool in aiding with facial recognition so that the robot can recognize the different children and recall each of them. Communication is another component to help in making the robot as human like as possible.

Other similar research projects include Clara the exercise coach. Clara is a robot, but is not as human like and fully functional as the Nao robot. Clara, belonging to the University of Southern California, is unable to move joints and physically exercise or dance. Another example is Autom, belonging to the Massachusetts Institute of Technology, who is a stationary robot that performs as a

weight loss coach. Autom possesses a tablet, which the user uses to maintain track of their eating and exercise habits. Snackbot, belonging to Carnegie Mellon University, is used for navigation and delivery of snacks to humans. All of these robots serve similar purposes, but none can perform to the ability as the BeyonceBot, which has the ability to perform as Clara, Autom, and Snackbot. There is a need for a robot such as BeyonceBot because it can help reduce childhood obesity.

Keywords: Childhood Obesity Epidemic, Exercising Humanoid Robot

This work is supported by the National Science Foundation HBCU-UP grant HRD-0714553, ASPIRE, and Broadening Participation in Computing grants CNS-1042454, SpelBots, and CNS-0742252, ARTSI, along with support from the Boeing Company, and General Motors.

Bioinformatics Tutorial Using Python

Jazette Johnson

Spelman College

Advisor: Dr. Alfred Watkins

Computer science is the study of algorithms including their formal and mathematical properties, hardware realization, linguistic realization, and their applications. Biology and computer science is put together to form bioinformatics. Bioinformatics is the science of using information to understand biology. It's a division of a larger field of computational biology, which is an application of quantitative analytical techniques in modeling biological systems. Bioinformatics makes it possible for biologist to do exciting things such as compare DNA sequences and generate results that are potentially significant. The bioinformatics tool also gives biologist the opportunity to over interpret data and assign meaning where none really exists.

The goal of this research project is to develop a web-based bioinformatics tutorial that would be a guide for students studying biology to learn the wide range of computational techniques to analyze biology data. Professors are sometimes unable to help the visual learner as well as the kinetic learner, so the website will incorporate audio, video, pictures and text to appeal to diverse learning styles, such as: visual learners, kinetic learners, and audio learners. The website will also be used to make it easier for biology students to understand how computers make it very efficient to analyze DNA sequences. Biology students understanding the background of the programming language python and computational biology will lead to further exploration of using Bioinformatics to help biologist results of analyzing DNA sequences more accurate.

Solving Bioinformatics Challenges Using a Command-Line Interface

Zazie Lumpkin	Alfred Watkins
Spelman College	Spelman College
350 Spelman Ln SW	350 Spelman Ln SW
Atlanta, GA 30314	Atlanta, GA 30314
678.913.8261	404.270.5768
zlumpkin@scmail.spelman.edu	AlWatkins@spelman.edu

According to Mount, Bioinformatics is 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. Biology is becoming more computationally intensive, causing biologist to lack the training and information-processing skills needed to tackle expanding datasets. Therefore, Computer Science plays a role in helping biologist to analyze data by using software tools and technologies to: help biologist with computational problems, organize data sets, edit data sets, and decrease the time in accomplishing tasks that would take longer to do by hand. The ultimate goal of this research is to create a tutorial in the form of a website that will assist biology students in learning Unix commands.

The objective is to construct a website tutorial for CIS115 that teaches Unix commands. Unix is a multi-user, multitasking operating system that allows people to use the computer at the same time and let each user run multiple jobs at once. CIS115: Introduction to Computing and Informatics is a new course that will be offered fall 2012 at Spelman College for biology students. This tutorial will assist the students in learning the rudiments of the shell (a tool for interacting with the computer through typed instructions at the command line).

In order to conduct research there were two tools that were needed to successfully create the tutorials: Xcode (version 3.1) and iTerm (version 0.9.0). Xcode is an Integrated Developments Environment containing a suite of software development tools developed by Apple Inc. iTerm is a program that when launched, provides a user interface to control the UNIX based operating system.

Before implementing my website, I acquired knowledge of a wide range of Unix commands, such as cd, ls, pwd, mkdir, mv, etc. After learning the Unix commands I created a web-based tutorial that teaches Unix Commands and provides exercises to enhance and ensure that the Unix commands were learned. The tutorial helps a user navigate through files and directories, edit files and directories, redirect input and output, use wildcards, and other useful Unix commands.

Although the website was successfully completed, in the future I would like to implement video and audio to appeal to diverse learning styles. I would also like to create a way to allow users to either submit their completed exercises to their professors for an assignment or online for immediate feedback.

Parallel Branch and Bound: Applying an Asynchronous Multi-pool Approach to Cyclic Best First Search

Joshua Gluck (Swarthmore College), Nartezya Dykes (Spelman College), Dr. Sheldon Jacobson, David Morrison, Jason Sauppe (University of Illinois)

Proven to be an acceptable strategy for solving integer programs, the branch and bound method consists of decomposing the set of feasible solutions into subsets and creating upper and lower bounds for the subproblems in order to eliminate options to find an optimal solution. Thus, the branch and bound algorithm ultimately builds a tree, with the root being the original problem and a search is performed on the tree to discover the optimal solution. Since subproblems are independent of one another and can be explored in any particular order, many algorithms have been developed and tested to parallelize the search. The branch and bound (b & b) algorithm has been used to find optimal solutions to a range of discrete programming problems since A.H. Land and A. G. Doig first proposed it in 1960. Until recently, the most efficient algorithm for searching and solving the problems generated by b & b was best first search (bfs). BFS was implemented using a basic priority queue, where for each parent node, the sub-problem of the parent node which had the best objective value would be the next to be explored.

More recently, a new and more efficient search algorithm, cyclic best first search (cbfs), has been developed. Cyclic best first search is implemented on a cyclic queue of priority queues, in which each individual priority queue handles a given depth of the tree generated by the b & b algorithm. CBFS cycles through the priority queues, essentially performing BFS once on each successive level of the tree one until the last level is reached, then starting again at the highest level with an unexplored node.

CBFS has shown great potential in solving discrete programming problems, whereas BFS falls short in memory demands due to keeping a partially expanded search graph. Thus, this makes the search tree impractical for very large applications. There has been little to no research done on the potential gains of CBFS applied in parallel. These gains could make solving currently infeasible problems possible in a fraction of the time. However, there are a number of problems with attempting to make any b & b algorithm parallel. These include the fact that sub problems are generated dynamically, meaning load balancing of sub problems is difficult to predict in advance. Additionally, a major concern is that the solution to any sub problem can affect the need to compute every other sub problem, but there is no algorithmic process to determine when this will occur, which causes a range of communication problems.

Since different algorithms have been used to explore a tree in parallel using fairly common search algorithms, the motivation for this project stems from trying these methods on the cyclic best first search to compare run times. From the possible methods this projects demonstrates one implementation of the method deemed most efficient. Specifically, it is an asynchronous multi-pool approach, which has shown significant efficiency gains. The 8-way parallel implementation was tested on Knapsack problem instances with 95 to 120 variables. The implementation generated a significant reduction in wall time amounting to at least a factor of 10 speed up for all problems. Given an 8- thread implementation, our achieving greater than an 8x speed gain may seem surprising; it is accomplished via the reduction in insertion times into priority queues enabled by the multi-pool approach. This implementation has the potential to be useful to solve larger problems in optimization.