Annual Poster Session 2009
Titles and Presenters

Peer Leader: Ana Betancourt

1) **Hector Lopez**
   Title: DNA Methylation-Based Early Lung Cancer Diagnosis: Preliminary Analysis of an Emerging Technology
2) **Javier Ordonez**
   Title: The Role of DNA and Genome Banking for the Conservation of Endangered Animal Species
3) **Marco Rodriguez**
   Title: Improving the Accuracy of Protein Secondary Structure Prediction Using Structural Alignment

Peer Leader: Ketaki Bhide

4) **Kyle Bolton**
   Title: Protloc: A *Shewanella* Protein Location Program
5) **Spyridon Moraros**
   Title: General Overview of Human Papillomavirus (HPV)
6) **Srilaxmi Nerella**
   Title: Proteomics Analysis of the Role of Calcium in *Bacillus subtilis*
7) **Rahul Vegesna**
   Title: Pharmacophore and Docking Studies on Tyrosine Kinase Inhibitor - FGFR1

Peer Leader: Anurag Gautam

8) **Prasanna Kolli**
   Title: Genetic Variation in the *Lecane luna* from Populations in the Chihuahua desert.
9) **Daniel Rodarte**
   Title: An Overview of Influenza A Virus Subtype H1N1
10) **Sravya Tamma**
    Title: Permutation-Based Genetic Algorithm to Predict the Secondary Structure of RNA

Bell Hall 130A
Friday, November 20, 2009, 10:30 AM – 12:30 PM

#1 DNA Methylation Based in Early Lung Cancer Diagnosis

Hector Lopez
Bioinformatics Program, The University of Texas at El Paso

Background

It is widely understood that early lung cancer diagnosis should entail a multi-modal approach. Combining two extensively used non-invasive screening technologies will appeal to medical professionals who are currently using these systems and to researchers interested in feasibility and population studies. Based on past experiences in the field of computer-aided diagnosis (CAD) of lung cancer, combining technologies like Multi-detector Computed Tomography (MDCT) and DNA methylation analysis of sputum samples will result in higher specificity and sensitivity acceptable for clinical use. Correlation between the localized and globalized information from the above-mentioned methodologies is obvious. Our interest is in the non-correlated information that gives us the hidden characteristics missed so far.

Methodology

There are three different disciplinary approaches to the early diagnosis problem:

- CAD based on MDCT
- Biological aspect of DNA methylation of sputum samples
- Bioinformatics aspect of DNA methylation data

Challenges and Objectives in Bioinformatics

Feasibility studies can be done using the data from Gene Expression Omnibus (GEO) repository. An example that can be used is available at
http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2771

This data set (22215 x 192) encompasses 192 samples from large airway epithelial cells obtained from cigarette smokers with suspect lung cancer. Our challenges are as follows:

- Dealing with high-dimensional data (under-determined problem)
- Cluster genes into sets correlating with different stages of cancer (High-risk, Early symptoms, Stage I, II, III, IV)
- Managing the continuous range of methylation values

The final goal of the project is achieving peer-reviewed publications through the following objectives:

- Review extensively literature in DNA methylation and high-dimensional data
- Cluster high-dimensional data to different sets based on assumed constraints even though it does not have any clinical significance at this point
- Propose an approach for a possible publication: it can be either comparing cluster techniques in different categories or improving an existing cluster method. This will help develop a novel algorithm in future.

#2

The Role of DNA and Genome Banking for the Conservation of Endangered Animal Species

Javier Ordonez
Bioinformatics Program, The University of Texas at El Paso

Scientists agree that life on Earth is facing the most severe episode of extinction after the event that drove dinosaurs extinct. The estimation is that around 27,000 animal species become extinct each year (3 species every hour)1. According to the "Earth's Endangered Creatures" list, there are 6,160 endangered known animal species as of October 23, 2009 (this list is updated on a daily basis). DNA and genome banks play a very important role in the efforts of collecting and preserving information stored in the genetic material of endangered species. DNA and genome banks are repositories of biological material. This material consists of samples of sperm, ova, embryos, tissue and serum2. The biological material needs to be stored at optimal conditions: temperature (Freezing in nitrogen at -196□C), and humidity (replacing water with a cryo-protectant fluid)2. The main objectives of these banks are to: Reduce the extinction risk, identify populations with genetic health problems, resolve taxonomic uncertainties, provide molecular genetic analysis for conservation planning, and contribute to the implementation of laws for the protection of endangered species3. The all-encompassing benefit of DNA and Genome Banking will be to avoid as much as possible the loss of species that resulted from millions of years of evolution, and going further, to be able to bring back species that became extinct before humans could even do anything to remediate it.

#3 Improving the Accuracy of Protein Secondary Structure Prediction Using Structural Alignment

Marco Rodriguez
Bioinformatics Program, The University of Texas at El Paso

Currently, understanding protein structures has become a major topic of research in today's era. New tools are emerging that can be used to quantify data about protein structures, but how is it possible to know which tool or in this case; which web server is more accurate or closer to what bioinformaticians and scientist might want? A web server named PROTEUS is currently in use and based on studies by Montgomerie, Sundararaj, Gallin and Wishart, PROTEUS will be a great benefit for many people due to the fact that it will try to use all the data that is currently presented in the Protein Data Bank, also known as PDB. Montgomerie, Sundararaj, Gallin and Wishart (2006) believe that 3D to 2D mapping approach could be a way to exploit secondary structure prediction. By combining a PDB-based structure alignment with prediction programs, a higher Q3 score could be achieved. Q3 score is the score used to see how accurate a secondary structure prediction is (most web-servers have a score of about 75%). Another form in which protein secondary structure prediction combines or is combined with PDB is with the implementation the SOV score. The SOV score will look at the alignment in a more detail manner in which it will look at situations where one sequence could produce different outcomes due to the amino acid group arrangement. Mongomerie et. al. made PROTEUS in a way that it will be possible for it to work with other networks that could double check its predicted structure, this method is called the de novo method. By using this approach, PROTEUS could achieve high Q3 scores like 88% on proteins that are being discovered and if part of a structure is on PDB the Q3 score exceeds 90%. If no homologous structures are found the Q3 is above 79%. PROTEUS is made of databases that are continuously updated and that are trying to reach a high secondary structure prediction rates. The databases are PSIPRED, JNET, and TRANSSEC. These databases have the ability to do multiple alignments that could be used to determine if structures have a common origin and how they could be traced. PSIPRED, JNET, and TRANSSEC aid PROTEUS by contributing with secondary structure prediction outputs that can validate the results obtained; they are called the "jury of experts." With the use of PROTEUS, a new way of increasing secondary structure prediction accuracy could be implemented and help in the understanding of proteins.

Based on Montgomerie, Sundararaj, Gallin and Wishart

#4 Protloc: A *Shewanella* Protein Location Program

Kyle Bolton
Bioinformatics Program, The University of Texas at El Paso

Shewanella oneidensis, MR-1, a type of gram-negative marine bacteria, is an important organism for study because of its uniquely versatile metabolism. It is able to use many different electron acceptors including nitrite, nitrate, thiosulfate, iron, manganese, and most interestingly, uranium. As the proteome of *S. oneidensis* grows, researchers have turned their attention to the localization of each protein in the hope of shedding light on the function of these proteins. There are many algorithms used for the purpose of protein localization; however, they each only provide one piece of the information needed to determine the most likely localization of the protein in the bacterium. Before completing our protein location program, Protloc, researchers had to run each of these algorithms by hand and analyze the results. This painstaking process did not provide the throughput necessary to analyze the huge number of proteins in the proteome. We designed Protloc to automate this localization process by combining 14 different programs and gave the most likely localization of these proteins based on the results provided by the *Shewanella Federation.* We also designed a web frontend for easy access with a customized Apache server configuration.

#5 General Overview of Human Papillomavirus (HPV)

Spyridon Moraros
Bioinformatics Program, The University of Texas at El Paso

Cervical cancer deaths are more prevalent among cancer-related deaths in women of the world's developing countries. Clinical research has proven that the Human Papilloma Virus (HPV) is the major cause of all cervical carcinomas. HPV pertains to a family of more than 100 viruses called papilloma viruses due to the nature of the lesions they generate (papillomas or more commonly known as warts). More than 30 of 100 types of HPV are transmitted through sexual intercourse and affect the genital region. Genital HPV infections are very common and are divided into low-risk and high-risk strains according to their oncogenic potentials. Low-risk strains, mainly strains 6 and 11, are the main cause of warts affecting the genitoanal region. High-risk strains, mainly strains 16 & 18, cause malignant neoplasias affecting the cervical region of the female genitourinary system. HPV strains 16, 18, and 31 are the main risk factors for cervical cancer. Early detection of this virus can lead to a great decrease in the mortality rate and enhance treatment success. Detection methods used are: the Papanicolaou (PAP) test smear, the HPV test, and colposcopy. The PAP smear test is a routine outpatient and convenient exam with which an epithelial sample of cervix is acquired and analyzed for any cellular abnormalities. The HPV test is employed whenever a PAP smear test returns an ambiguous result. Colposcopy, an outpatient & mildly invasive procedure, is the only method that asserts the presence of an HPV infection and serves as a treatment for the said infection. This method is only employed upon the return of a positive PAP smear or HPV test. Presently, the development of vaccines, such as Cervarix and Gardasil, has made it possible to prevent female genital infection from the main HPV strains. By increasing the awareness in the general population, especially in developing countries, of the procedures that are readily available to combat this disease, it is possible to decrease the morbidity and mortality rates produced by this virus.

#6

Proteomics and Bioinformatics Analysis of the Role of Calcium in *Bacillus subtilis*

Srilaxmi Nerella,[1] Delfina C Dominguez,[1,2] and Rosanna Lopes[2]
[1]Bioinformatics Program and [2]Clinical Lab Sciences, The University of Texas at El Paso

Abstract:While the role of calcium binding proteins (CaBPs) in cell signaling pathways and homeostasis is well established in eukaryotic cells, the physiological function of CaBPs in prokaryotes is unknown. We hypothesize that CaBPs play an important role in Ca^{2+} homeostasis and that Ca^{2+} ions are involved in several processes in bacterial cells. The purpose of this work was to map changes in protein expression in *Bacillus subtilis* cells as a function of cytosolic Ca^{2+} levels, and to identify CaBPs involved in the bacterial response to changes in extracellular calcium using proteomic analysis approach. The proteins that had showed significant changes in expression with $CaCl_2$ were analyzed with various bioinformatics tools. The expression of 284 proteins (209 and 75, BAPTA/EGTA respectively) was significantly modified after chelator treatment while 57 proteins changed after the addition of $CaCl_2$. Among the 57 proteins analyzed, 3 proteins were found to have calcium binding domains in them.

#7 Pharmacaphore and Docking Studies on Tyrosine Kinase Inhibitor FGFR-1

Rahul Vegesna,[1,2] Ruchi Yadav,[3] and Kishore Kumar[3]
[1]Bioinformatics Program, The University of Texas at El Paso, [2]Department of Bioinformatics, Satyabhama University, India, and [3]GVK Biosciences Pvt Ltd, India.

Fibroblast Growth Factor Receptor 1 (FGFR1) gene encodes a tyrosine kinase receptor protein that is part of the fibroblast growth factor and growth factor receptors family. Defects in FGFR1 are a cause of Pfeiffer Syndrome (PS), also known as acrocephalosyndactyly type V. PS is characterized by craniosynostosis (premature fusion of the skull sutures) with deviation and enlargement of the thumbs and great toes. Recent studies showed that when *FGFR1* is specifically amplified and over expressed, the *FGFR1* signal thus obtained is important for the survival of the cancer cell line in Breast cancer patients with *FGFR1* gene amplification. So, this gene may be a potential therapeutic target for specific breast cancers. Tyrosine kinase activity has been observed in many different malignancies and diseases. Thus, there is a therapeutic need for specific inhibitors of tyrosine kinase activity. FGFR1 contains an Ig-like domain and a tyrosine kinase domain. This receptor has multiple distinct isoforms and is a Type I membrane protein widely expressed in specific tissues. FGFR1 binds fibroblast growth factor and induces mitogenesis and cellular differentiation. 52 human FGFR1 inhibitors were collected with activity data spanning from research articles and 3D structures of all molecules were constructed by using Cerius2 and then submitted them to Catalyst program to generate a Pharmacophore model using Hypogen algorithm.Then Docking studies were done on FGFR1 and the binding sites were found. Then results from docking and pharmacophore where compared to validate the pharmacophore obtained.

#8 Genetic Variation in the Rotifer *Lecane luna* from Populations in the Chihuahuan Desert

Prasanna Kolli[1] and Elizabeth Walsh[1,2]
[1]Bioinformatics Program and [2]Department of Biological Sciences
The University of Texas at El Paso

Many rotifer species are found to be present globally in variety of habitats. Recent studies have reported substantial genetic variation among population of these globally distributed species. However, little attention has been given to rotifer populations in the Chihuahua Desert. Here we sequenced 12 populations of the rotifer *Lecane luna* to determine if cryptic speciation is found. *L. luna* was found in 52 locations in the Chihuahuan desert (US & México). Isolates from some of these populations were cultured for DNA sequencing. Based on preliminary sequence results from the cox1 gene, levels of genetic variation in Chihuahua Desert populations ranges from 4 - 26%. These results indicate that cryptic speciation may be present in some populations in the Chihuahua Desert. However, ITS sequences showed little genetic divergence (0-2.3%). Because of the differences in levels of genetic variation, future studies are needed to confirm whether there is cryptic speciation in this species.

#9 An Overview of Influenza A Virus Subtype H1N1

Daniel Rodarte
Bioinformatics Program, The University of Texas at El Paso

Many rotifer species are found to be present globally in variety of habitats. Recent studies have reported substantial genetic variation among population of these globally distributed species. However, little attention has been given to rotifer populations in the Chihuahua Desert. Here we sequenced 12 populations of the rotifer *Lecane luna* to determine if cryptic speciation is found. *L. luna* was found in 52 locations in the Chihuahuan desert (US & México). Isolates from some of these populations were cultured for DNA sequencing. Based on preliminary sequence results from the cox1 gene, levels of genetic variation in Chihuahua Desert populations ranges from 4 - 26%. These results indicate that cryptic speciation may be present in some populations in the Chihuahua Desert. However, ITS sequences showed little genetic divergence (0-2.3%). Because of the differences in levels of genetic variation, future studies are needed to confirm whether there is cryptic speciation in this species.

#10 Permutation-Based Genetic Algorithm to predict the Secondary Structure of RNA

Sravya Tamma,[1] Olac Fuentes,[1,2] and Ming-Ying Leung[1,3]
[1]Bioinformatics Program, [2]Department of Computer Science, and
[3]Department of Mathematical Sciences, The University of Texas at El Paso

The paper presents a permutation-based genetic algorithm for predicting RNA secondary structure. It is practicable and can be used to predict real RNA molecules. The concept of permutation is introduced, which is the starting point of our algorithm. Individual is represented as a permutation of stem list. Crossover operator, mutation operator, and selection strategy are designed to be compatible with such an individual representation. At the end of the paper, a comparison between our result and that from RNA structure is outlined.