BINF 704 Fall 2009 Colloquium

Instructor - Dr Jeff Solka (jsolka@gmail.com)

Meeting Place - Prince William Bull Run Hall Rm 130
Meeting Time - 4:30-6:00 pm Tuesdays

Course Webpage
http://binf.gmu.edu/~jsolka/fall09/binf704/Fall_2009BINF_704_colloquium_Syllabus_rev1.html

Course Description:
This course will provide an opportunity to learn about ongoing bioinformatics research outside of George Mason University. The students in this class will be exposed to presentations by a number of different researchers in a number of different bioinformatics research areas.

Prerequisites:
Good standing in the Bioinformatics and Computational Biology Department

Required Text:
None

Grading:
Grades will be based on 8 short reports on the speakers, each approximately 2 paragraphs in length. These ½ page summaries of the talks CANNOT be the speaker’s abstract and are due 1 week after the speaker’s presentation. Students are expected to produce these reports for 8 of the planned presentations.

In addition a summary report for one of the talks must be created ahead of time prior to the speaker’s presentation based on the student’s analysis of a minimum of two papers by the speaker. This 3-5 page summary report must be prepared ahead of time and should end with five questions they would like to ask the speaker to answer about their work. All citations have to be given in full, including extra Web sites used. The students are warned to be very careful with regards to plagiarism issues. This particular summary report must be turned in a week prior to the speaker’s planned presentation. A seminar can be covered by both a report and a summary report, since sometimes the speaker does not cover all the materials in their papers in the seminar.

Projected Class Schedule

http://binf.gmu.edu/~jsolka/fall09/binf704/Fall_2009BINF_704_colloquium_Syllabus_rev1.html
Sept. 1, 2009
Student Orientation

Sept. 8, 2009
Dr. Hao-Thu Huong (USA MEDCOM WRAIR)

Title: Comprehensive Full Genomic Sequencing of 2009 Novel H1N1 Viruses by High Throughput “Next-Gen” Sequencing

Background of 2009 Novel H1N1 Pandemic Outbreaks:

The genomes of the last three pandemic influenza viruses (1918 H1N1, 1957 H2N2 and 1968 H3N2) all originated in whole or in part from non-human reservoirs, and the HA genes of all of the pandemic viruses ultimately originated from avian influenza viruses. Novel 2009 influenza A (H1N1) is a new flu virus of swine origin that was first detected in Mexico and the United States in March and April 2009. Since its initial identification and announcement of the unusual outbreaks, the 2009 swine H1N1 virus then quickly spread into Mexico's neighboring country, US via mostly Spring break tourists. Following Mexico and US reported cases, confirmed outbreaks of 2009 swine H1N1 were rapidly proliferated and spread into countries beyond America continental, such as Europe, Asia, Africa, South America possibly through the efficient modern traveling system. WHO then upgraded and announced the novel 2009 H1N1 infections as worldwide flu pandemic infections on June 10, 2009.

The novel H1N1 flu mainly spreads in the same way that regular "seasonal influenza" spreads, which is through the air from coughs and sneezes or touching those infected surfaces. It seems that new cases in the U.S. and most cases throughout the world have so far been mild relative to the initial reported cases in Mexico. But because this is a new virus, most people do not have immunity to it, and illness may eventually become more severe and widespread in different demographic and population groups as a result. Along with the actual spread of viral infections, availabilities of 2009 swine H1N1 specific sequences deposited to NCBI’s database GenBank also rapidly proliferated starting early April through July, 2009. Laboratories of worldwide origins using mostly Sanger-
Dideoxy-terminator sequencing method sequenced most of 2009 novel H1N1 sequences. Based on the up to date sequence comparisons, it is clear that not all deposited 2009 swine H1N1 sequences were identical. However, it was uncertain whether the differences of those supposedly identical/similar causative agents were due to various clinical relevancies, i.e., severe or mild infections. Or it was also possible that different sequences were actually derived from different sequencing schemes using various RT-PCR amplification primers and protocols employed by wide-range laboratories all over the world.

Abstract:

Since its initial introduction in 2005, the 454 Roche FLX sequencing platform had been utilized for ultra-depth sequencing projects for various microorganisms. The massively parallel pico-liter scale amplifications and pyrosequencing of individual DNA molecules (Margulies et al. 2005) allow scientists to investigate the heterogeneous populations of microbial words that play important role in determining disease outcome and drug resistance. Here, we systematically investigate the potential of ultra-deep pyrosequencing to determine and assemble full genome sequences of 2009 novel H1N1 viruses from worldwide geographic origins. A robust RT-PCR protocol was established to efficiently amplify across the boards of all 8 2009 novel H1N1 RNAs into sufficient cDNA quantities, i.e., greater than 5 ug to be processed and sequenced by the Roche 454 FLX system using MID bar-coding system. Massive DNA sequences, i.e. >1,000,000 reads with mean >200 base pairs in length derived from de novo sequences of each individual cDNA fragments were readily obtainable from each individual Roche 454 FLX sequencing run containing up to 24 bar-coded full genomic influenza A cDNA of difference origins. In addition to general consensus sequences routinely detected by traditional Sanger sequencing method, rare genetic variants, i.e. 1-2% of total viral population could also be detected and confirmed from pyrosequencing that might play important roles in determining/predicting viral virulence or anti-viral drug resistance. Our readiness to handle the next wave of 2009 H1N1 outbreaks could be greatly enhanced by using Roche 454 as a feasible platform to sequence and analyze large number of 2009 novel H1N1 genomes for the imminent large-scale 2009 winter influenza season in north hemisphere.

Sept., 15 2009
Dr. Jeff Solka (NSWCDD, GMU) A Conditional Entropy Based Approach to Co-clustering for the Analysis of Gene Expression Data

http://bifs.gmu.edu/julka/fall09/bifs704/Fall2009BIFSColloquium_Syllabus_rev1.html

3 of 7
Methods of high-level data exploration capable of robustness in the face of noise found within microarray data are few and far between. Solutions making use of all original features to derive cluster structure can be misleading while those that rely on a trivial feature selection can miss important characteristics. We present a method adopted from previous work in the field of geography (Guo et al, 2003) relying upon conditional entropy between pairs of dimensions to uncover underlying, native cluster structure within a dataset. Applied to an artificially clustered data set, this method performed well though some sensitivity to multiplicative noise was in evidence. When applied to gene expression data, the method produced a clear representation of the underlying data structure.

Sept. 22, 2000
Dr. Stephen Lockett (National Cancer Institute - Frederick / SAIC - Frederick, Fort Detrick)

Title
"Novel Approaches to Cell and Tissue Image Analysis, and its Applications to Genomic Analysis of Breast Cancer"

Abstract
Segmentation of individual cells and cell nuclei from microscopic images continues to be a challenging task, and the appropriate strategy is dependent on the biological application. For understanding communications between cells that drive tissue development and function, as well as disease-related processes such as tumorigenesis, we have developed highly reliable and accurate semi-automatic 3D algorithms. For high throughput screening of nuclei from 2D cell culture images, we have developed model-free segmentation from which the most frequent class of objects in the image is automatically modeled using statistical pattern recognition. This software adapts on the fly to changes between datasets in the characteristics of the imaged nuclei. For automatic analysis of cell nuclei in tissue, we have developed an intelligent framework coupling a hybrid nuclei segmentation algorithm with pattern recognition algorithms to automatically identify well segmented nuclei. Application of this software with spatial statistical analysis of the FISH spots indicates encouraging preliminary results for diagnosing breast cancer based on the positioning of certain genes in cancer cell nuclei versus the nuclei of normal cells.

Sept. 29, 2009
Dr. James Taylor (National Institutes of Health)

Title
Making Rare Diseases More Complex with Genomics: Eye on Childhood Cancer

The initial assembly of human genomic sequence led to anticipation for rapid advances in understanding the genetic basis of human diseases. Efforts are now focused on whole-genome genotype and DNA sequencing to characterize relationships between genotype and phenotype in complex or multifactorial human diseases like diabetes and cancer. A major challenge for these studies is the ability to distinguish between natural and functional genetic variants and to understand how functional mutations contribute to
disease. Our laboratory’s mission is to map and characterize medically important polymorphisms and mutations in rare human diseases using bioinformatic and laboratory based methodologies. In one example, we have examined childhood cancers which are a collection of exceptionally rare diseases where their rarity severely limits suitable numbers of tumor samples available for genomic studies. Rhabdomyosarcoma (RMS) is a childhood cancer originating from skeletal muscle that affects approximately 250 children each year in the US. Patient survival is poor in the presence of metastatic disease, and few determinants that regulate metastasis development have been identified. The receptor tyrosine kinase FGFR4 is highly expressed in RMS tissue suggesting a role in tumorogenesis, although its functional importance has not been elucidated. Examination of existing expression databases demonstrated that higher FGFR4 expression in RMS tumors is associated with advanced-stage cancer and poor survival, while FGFR4 knockdown in a human RMS cell line reduced tumor growth and experimental lung metastases when the cells were transplanted into mice. This led us to hypothesize that FGFR4 activation though either overexpression or mutation might contribute to disease progression and metastasis. Targeted gene sequencing identified FGFR4 tyrosine kinase (TK) domain mutations among 7% of primary human RMS tumors. Mutations occurring at two codons in the FGFR4 TK domain were predicted to promote receptor phosphorylation using bioinformatic algorithms and protein structural modeling. Functionally, FGFR4 mutants K535 and E550 increased autophosphorylation, STAT3 signaling, tumor proliferation, and metastatic potential when expressed in a murine RMS cell line. These mutants also transformed NIH 3T3 cells and led to an enhanced metastatic phenotype. Finally, murine RMS cell lines expressing the K535 and E550 FGFR4 mutants were substantially more susceptible to apoptosis in the presence of a pharmacologic FGFR inhibitor than the control cell lines expressing the empty vector or wild-type FGFR4. Together, these data demonstrate that mutationally activated FGFR4 functions as an oncogene in RMS and these are believed to be the first known mutations in a receptor tyrosine kinase in this tumor. These findings support the potential therapeutic targeting of FGFR4 in RMS. Overall, this study also demonstrates the importance of integrating genetic, bioinformatic and archived databases for the study of rare human diseases.

Oct. 6, 2009
Dr Sharmilla Basu (MindSpec)

Title
AutDB: A disease-driven database model

Abstract
In the post-genomic era, multi-faceted research on complex disorders such as autism has generated diverse types of molecular information related to the disorder. Operationally, the number of articles reporting putative candidate loci, as well as high throughput array-based studies reporting many loci in a single publication is accumulating at a fast pace. To address the genetic complexity of ASD, we have developed AutDB (http://www.mindspec.org/autdb.html), a publicly available web-portal for on-going collection, manual annotation and visualization of genes linked to the disorder. We present a disease-driven database model in AutDB where all genes connected to ASD are collected and classified according to their genetic variation: candidates identified from genetic association studies, rare single gene mutations and genes linked to syndromic autism. Gene entries are richly annotated for their relevance to autism, along with an in-depth view of their molecular functions. The content of AutDB originates entirely from the published scientific literature and is organized to optimize its use by the
research community. The main focus of this resource is to provide an up-to-date, annotated list of ASD candidate genes in the form of reference dataset for interrogating molecular mechanisms underlying the disorder. Our model for consolidated knowledge representation in genetically complex disorders could be replicated to study other such disorders.

Oct. 13, 2009
No Classes Columbus Day Break

Oct. 20, 2009
Avery Bryant (NSWCDD)

Title: Performing Scientometric Analysis through Document Clustering and Dynamic Graph Visualization

Abstract
Scientometrics is performed by the analysis of the open source scientific literature in an attempt to analyze science. Bibliographic databases provide access to this scientific literature in the form of millions of publications from journals and conference proceedings amongst other resources. Document clustering refers to clustering based on free text content-based features such as a publications title or abstract. These features can be used to represent publications in the vector space model by a term-document matrix which clustering methodologies can be applied to. This presentation focuses on the 2-D graph visualization of these clustering solutions using two techniques. The first technique being a specified graph layout obtained by multi-dimensional scaling and the other: a force directed graph layout obtained using distances in the ambient space. Nodes represent clusters while edges represent some relationship between the documents in clusters like overlapping citations or overlapping institution affiliations. Node color or size can also be used to highlight cluster specific features such as the number of documents in a cluster or the average growth rate, by publication year, of the documents belonging to a cluster. Note the focus of this work is not on document clustering or 2-D visualization of high dimensional data but on performing scientometrics analysis at the document cluster level using graph visualization. Using this dynamic graph visualization (node positions being static) scheme we hope to create a system that can be used to take advantage of the feature rich environment provided by the open source scientific literature.1

Oct. 27, 2009
Dr. Françoise Seillier-Moiseiwitsch (Georgetown University Medical Center)

Nov. 3, 2009
TBD

Nov. 10, 2009
TBD

Nov. 17, 2009
Weifan Zheng, Ph.D., NCCS

http://enf.gmu.edu/jsr4k/Drill/Fall09/bin0704/Fall_2009INF_704_colloquium Syllabus_rev1.html
Nov. 24, 2009
Brandon Higgs Ph.D.

Dec. 1, 2009
Jeniffer Barb Ph.D., NIH

Dec. 8, 2008
Please attend student research day