Title
Determining the existence of long-range correlations for financial and volcanic time series

Authors
Peter K. Asante,1 Maria C. Mariani,1,2 and Osei K. Tweneboah1
1Computational Science Program and 2Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX

Abstract
In this study, we use the Diffusion Entropy Analysis (DEA) to analyze and detect the scaling properties of time series from both emerging and well-established markets as well as volcanic data, which is known to have high frequencies. The objective is to determine the characterization, i.e., whether they follow a Gaussian or Lévy distribution, as well as establish the existence of long-range correlations in the financial and volcanic time series. The results obtained from the DEA technique are compared with the Hurst R/S analysis and Detrended Fluctuation Analysis (DFA) methodologies. We conclude that given the scaling exponents $\delta$ derived from the DEA and $H$ derived from the Hurst R/S analysis and DFA, if $\delta$ is related to $H$ by the relation $\delta = 1/(3 - 2H)$ then the time series is characterized by a Lévy distribution and if $\delta = H$ then the time series may be characterized by fractional Brownian motion.
Title
Exploiting semantic relationships for prediction of drug-target binding affinity from chemicogenomic data.

Authors
Ricardo Avila, Bioinformatics Program, The University of Texas at El Paso, El Paso, TX

Abstract
The identification of drug-target interactions and their binding affinities from biological data is a big problem in computational drug discovery and development. Existing structure-based methods such as docking have limitations when dealing with proteins that have unknown structures. Conversely, ligand-based prediction methods such as quantitative structure-activity relationship (QSAR) face challenges when the target protein does not have many known binding ligands. In order to improve prediction in these cases, machine learning models have attempted to combine genomic information from the primary sequence of proteins and chemical information from the two-dimensional structure of their binding partners. Several methods for featurization of biological sequences and chemical structures exist, but recently a new method of vector representation for biological n-grams was developed by Asgari and Mofrad (2015), which borrows semantic analysis techniques from the field of natural language processing. Similarly, a method for representing chemical structures as vector embeddings was developed by Jaeger et al. (2018). This project presents a pipeline for building binding affinity prediction models based on the combination of these two featurization methods. Given an input protein sequence, we find homologous proteins, and download their sequences from Uniprot. We also download the SMILES string representations of the ligands associated with the homologous proteins from the ChEMBL database, along with their experimental binding affinity value. Descriptors for protein sequences and chemical structures are generated using vector embedding, and regression models are used to predict binding affinity. Early results demonstrate the feasibility of using these methods for the prediction of drug-target interactions.
Title
Integrated statistical and machine learning algorithms for predicting and classifying G protein-coupled receptors

Authors
Fredrick Ayivor¹ and Ming-Ying Leung¹,²
¹Computational Science Program, ²Bioinformatics Program, ²Border Biomedical Research Center, and ²Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX

Abstract
G Protein-Coupled Receptors (GPCRs) are transmembrane proteins with important functions in signal transduction and often serve as drug targets. With increasing availability of protein sequence information, there is much interest in computationally predicting GPCRs and classifying them according to their biological roles. Such predictions are cost-efficient and can be valuable guides for designing wet lab experiments to help elucidate signalling pathways and expedite drug discovery. There are existing computational tools of GPCR prediction that involve principal component analysis (PCA), intimate sorting (IS), support vector machine, and random forest (RF) techniques using various sequence derived features. While accuracies of over 90% were reported for their own test datasets, the capabilities in distinguishing GPCRs from transmembrane non-GPCRs had not been measured in any of these tools. Furthermore, no direct comparison of the different approaches has been conducted. In this project, we have established two new GPCR prediction algorithms that integrate combinations of PCA, IS, and RF with the univariate feature selection method that has not been used for GPCR predictions before. The same 1355 sequence features are used uniformly with a test dataset with 2776 positive examples of confirmed GPCRs, and 3781 negative examples including transmembrane non-GPCRs. Overall prediction accuracies are over 90%, and the false positive predictions among the transmembrane non-GPCRs are substantially lower than those in existing tools. These results suggest that integrated algorithms performs well with GPCR prediction. We plan to further explore different integrated prediction approaches and apply them to the GPCR classification problem in the future.
Organizational complexity of the insulin-like growth factor 2 gene and locus in vertebrates

Kabita Baral,1 Jonathon Mohl,1,2 Ming-Ying Leung,1,2,3 and Peter Rotwein4
1Bioinformatics Program, 2Computational Science Program, 3Border Biomedical Research Center, and 4Department of Biomedical Sciences, The University of Texas at El Paso, and 4Department of Biomedical Sciences, Texas Tech University Health Sciences Center, El Paso, TX

Abstract:
Insulin-like growth factor 2 (IGF2) protein are essential for fetal and prenatal growth. Overexpression of IGF2 causes asymmetrical organ and tissue overgrowth while under expression causes reduced growth and bodily dysmorphism. Previous studies have demonstrated that the IGF2 gene is generally conserved across groups of fishes and mammals. The goal of this project is to determine the genomic loci of IGF2 genes and subsequently analyze their organizational complexity to identify specific conserved regions across vertebrates with the aim to help better understand IGF2 functions and evolutionary trends. We plan to extract the gene and protein sequences, genomic loci, single nucleotide polymorphisms (SNPs) for IGF2 and its surrounding regions from public databases including ClinVar, Ensembl, NCBI, and UniProt. We will study 16 different organisms including human, mouse, chimpanzee and bats and build phylogenetic trees to see how consistently the IGF2 gene has evolved over time. Multiple sequence alignment will also be performed to find similarity across sequences. Additionally, we will look into SNPs to gain better insights on the biological and clinical significance of the gene. We expect the most important regions for IGF2 expression to be conserved across the species, and other parts of the gene may have evolved slightly more in higher vertebrates as we move towards the tip of the phylogenetic tree. The analysis will help us better understand the importance of IGF2 for normal growth. The results may further shed light on the causes behind abnormal expression and clinical manifestations.
Title
Improvement in auditory comprehension reaction time following a sport-related concussion is linked to visual processing abilities and visual-motor speed

Authors
Adalberto Isaac Garcia,1,2 Anita Bialunska,2 Bess Sirmon-Taylor,2 and Anthony Salvatore2
1Bioinformatics Program, College of Science, and 2Concussion Clinic, Department of Speech Language Pathology, College of Health Sciences, The University of Texas at El Paso, El Paso, TX

Abstract
Studies on mild traumatic brain injury (mTBI, concussion), common in athletic, reveal that this type of injury is heterogeneous, linked to complex and variable neuronal pathophysiology, resulted in different symptoms or recovery patterns. To assess auditory comprehension abilities in athletes we used the computerized version of Revised Token Test (C-RTT), this test showed compromised auditory processing when compared to matched controls. Furthermore, we decided to investigate these individual differences using logistic regression of complex tasks which required auditory comprehension, but also multisensory processing and sensorimotor integration. We tested whether a variable outcome of C-RTT reaction time 20 days after injury across athletes dependent on recovery of an individual’s cognition, and other factors. We explored a variety of potential predictors namely results obtained from 143 controls and 85 athletes with concussion from three consecutive evaluations following injury on VIII Subtest of C-RTT, ImPACT, and Post-Concussion Symptoms Scale, together with additional data obtained on other clinical tools (fluency and balance tests), and demographic variables. A positive change was predicted by visual processing capacities and visual-motor speed on second evaluation, as well as by impulse control and Efficiency Score of C-RTT from first evaluation. More severe deficits in C-RTT performance and impulse control right after the injury, a low visual processing, and a prompt visual-motor speed foster a positive outcome timing of C-RTT performances on third evaluation.
Abstract:
RNA-seq is a next generation sequencing method that can quantify the level of different RNAs in biological samples. It is often used to identify differential gene expression among different samples or treatments. The purpose of this project is to develop an efficient and user-friendly bioinformatics pipeline for processing raw RNA-seq sample data from Illumina sequencers, converting them to a format easily accessible and interpretable by biomedical researchers. A typical computational workflow for RNA-seq is as follows: trimming the sequence, aligning the sequences to reference genome, normalizing transcript levels, merge the assemblies, obtain summary statistics and check for significance using t-test with Benjamini-Hochberg correction. The developed Python script uses the generated results and creates an excel file with various levels of filtering such as significance level, fold changes, etc. Based on the list of genes obtained from these filters, three more steps are performed. 1) Gene ontology terms are retrieved showing the cellular components, molecular functions, and biological processes. 2) A query of up and down regulated genes is performed in LINCS L1000 to analyze the gene expression profile for connectivity to known perturbations. 3) Venn diagrams are generated to display counts of differentially expressed genes in multiple comparative studies among different samples. For illustration, gastric biopsy data from NCBI were processed using this pipeline, demonstrating that the entire workflow could be automated. This RNA-seq pipeline will eliminate manual interactions and errors at intermediate steps, reduce the total completion time of the analysis, and help avoid misinterpretation of results.
Title
Assessment of web-based G protein-coupled receptor prediction and classification bioinformatics tools

Authors
Eder E Perez,1 Khodeza Begum,2 Jonathon E Mohl,2,3 and Ming-Ying Leung1,2,3
1Department of Mathematical Sciences, 2Computational Science Program, 3Bioinformatics Program, and 3Border Biomedical Research Center, The University of Texas at El Paso

Abstract
G Protein-Coupled Receptors (GPCRs) are a large protein family of membrane receptors that take an important role in physiological functions such as vision, neurotransmission, and immunology. Consequently, GPCRs are the target of many pharmaceuticals and are of great interests to researchers. Various prediction algorithms have been developed to identify GPCRs as well as classify the proteins to their respective families using the GRAFS and IUPHAR classification system. Online web-servers are available to predict and classify GPCRs using various machine learning and statistical methods. The goal of the project is to compare the results for prediction and classification for web-servers such as UTEP GPCR Pipeline (TMHMM and GPCRPred), PCAGPCR, PredGPCR, SVMProt, GPCR-GIA, and GPCR-CA. A balanced dataset consisting of 60 confirmed GPCRs and 60 non-GPCRs, including 10 transmembrane non-GPCRs, were used to test the performance of these web-servers. Results showed that web-servers with greater than 90% overall accuracy were TMHMM, GPCRPred, and GPCR-GIA. All other servers had accuracy above 83%. Sensitivity ranged from 68 to 100%, specificity from 73 to 98%, positive predictive value from 79 to 98%, and negative predictive value from 76 to 100%. False positive rate for transmembrane non-GPCR ranged from 10 (PredGPCR) to 100% (PCAGPCR), with average difference between transmembrane non-GPCRs and all non-GPCRs at 31%. Accuracy for GPCR family classification varied from 60 to 81% due to each web-server’s variation of their use of the IUPHAR system. Future work involves an extensive comparison using the sequences in the GPCR-PEn database compiled by our group at gpcr.utep.edu.
Title
A data conversion script for parsing variant calling format file

Authors
Bofei Wang,¹ Jonathon Mohl,¹,² and Ming Ying Leung¹,²,³
¹Computational Science Program, ²Border Biomedical Research Center, ³Bioinformatics Program and ³Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX

Abstract
Rapid advances in next generation sequencing (NGS) technologies provide many opportunities to identify associations between genetic sequence variants (GSV) and diseases, which may lead to better clinical diagnosis and treatments. OncoMiner is a bioinformatics pipeline developed at UTEP (OncoMiner.utep.edu) for mining NGS data. Since the current input for OncoMiner is limited to a specific file format set by the Otogenetics sequencing company, we have developed a Python script to handle the more common variant calling format (VCF) NGS files and convert them to the OncoMiner input (OMI) format. Required OMI data items include genome position, nucleotides involved, and genomic region type for each GSV. A VCF file consists of meta-information lines, a header line, and data lines and contains the genomic locations of the variants found. While most of the required OMI fields can be extracted directly from the VCF file, the genomic region type needs to be determined by comparison with a reference genome containing precise locations for transcripts, exons, and coding regions. To improve efficiency, the script splits the VCF file by chromosomes into smaller files for parallel processing. Eight VCF files, ranging from 430 to 910 KB in size and containing data from leukemia and prostate cancer patients, were downloaded from The Cancer Genome Atlas for testing the script. The average runtimes with one and through four cores were 0.533, 0.409, 0.405, and 0.404 minutes respectively. Future work includes analyzing runtime behavior with larger VCF data files using more cores, and incorporating the script into OncoMiner.
Review Presentation: “Application of a new scaffold concept for computational target deconvolution of chemical cancer cell line screens”

Dristi Adhikari, Bioinformatics Program, The University of Texas at El Paso, El Paso, TX

This poster is a review of the paper “Application of a new scaffold concept for computational target deconvolution of chemical cancer cell line screens” by Kunimoto et al. (2017) published in the journal ACS Omega. The abstract as it appears in the original publication is as follows:

“Target deconvolution of phenotypic assays is a hot topic in chemical biology and drug discovery. The ultimate goal is the identification of targets for compounds that produce interesting phenotypic readouts. A variety of experimental and computational strategies have been devised to aid this process. A widely applied computational approach infers putative targets of new active molecules on the basis of their chemical similarity to compounds with activity against known targets. Herein, we introduce a molecular scaffold-based variant for similarity-based target deconvolution from chemical cancer cell line screens that were used as a model system for phenotypic assays. A new scaffold type was used for substructure-based similarity assessment, termed analog series-based (ASB) scaffold. Compared with conventional scaffolds and compound-based similarity calculations, target assignment centered on ASB scaffolds resulting from screening hits and bioactive reference compounds restricted the number of target hypotheses in a meaningful way and lead to a significant enrichment of known cancer targets among candidates.”
Title
Review Presentation: “Mapping molecular datasets back to the brain regions they are extracted from: remembering the native countries of the hypothalamic expatriates and refugees”

Presenter
Gloria Miranda, Bioinformatics Program, The University of Texas at El Paso, El Paso, TX

Abstract
This poster is a review of the article “Mapping molecular datasets back to the brain regions they are extracted from: remembering the native countries of the hypothalamic expatriates and refugees” by Khan et al. (2018) in the book Systems Neuroscience, Advances in Neurobiology (to appear). The summary as it appears in the original article is as follows:

“In this article, we envision ways in which molecular information extracted from the brain using methods such as transcriptomics, proteomics, and peptidomics can be anchored to locations in standardized atlas maps of the brain in order to preserve the provenance of the datasets and contextualize them with other datasets. We argue that whereas most researchers probe, dissect, mine, or interrogate the living brain and report back with valuable scientific information, such information would be worth more if it included mapped locations of where they traveled and what they found there. Mapping to a standardized reference allows current and future travelers to return to the same landscape with accuracy and precision, generate reproducible data from reproducible experiments, and allows them further to integrate and contextualize new data they gathered in that mapped location with other data gathered in the same space. By carefully documenting the locations, for example, of brain regions from which molecular information is extracted for large-scale analyses, scientists can contribute further to our collective history of the native landscape from which this expatriated molecular information originated.”
Title
Review Presentation: "Transcriptomic analyses and leukocyte telomere length measurement in subjects exposed to severe recent stressful life events"

Presenter
Raymond Anan Otoo, Bioinformatics Program, The University of Texas at El Paso, El Paso, TX

Abstract
This poster is a review of the paper "Transcriptomic analyses and leukocyte telomere length measurement in subjects exposed to severe recent stressful life events" by Lopizzo et al. (2017) published in Translational Psychiatry. The abstract as it appears in the original publication is as follows:

“Stressful life events occurring in adulthood have been found able to affect mood and behavior, thus increasing the vulnerability for several stress-related psychiatric disorders. However, although there is plenty of clinical data supporting an association between stressful life events in adulthood and an enhanced vulnerability for psychopathology, the underlying molecular mechanisms are still poorly investigated. Thus, in this study we performed peripheral/whole-genome transcriptomic analyses in blood samples obtained from 53 adult subjects characterized for recent stressful life events occurred within the previous 6 months. Transcriptomic data were analyzed using Partek Genomics Suite; pathway and network analyses were performed using Ingenuity Pathway Analysis and GeneMANIA Software. We found 207 genes significantly differentially expressed in adult subjects who reported recent stressful life experiences (n = 21) compared with those without such experiences (n = 32). Moreover, the same subjects exposed to such stressful experiences showed a reduction in leukocyte telomere length. A correlation analyses between telomere length and transcriptomic data indicated an association between the exposures to recent stressful life events and the modulation of several pathways, mainly involved in immune-inflammatory-related processes and oxidative stress, such as natural killer cell signaling, interleukin-1 (IL-1) signaling, MIF regulation of innate immunity and IL-6 signaling. Our data suggest an association between exposures to recent stressful life events in adulthood and alterations in the immune, inflammatory and oxidative stress pathways, which could be also involved in the negative effect of stressful life events on leukocyte telomere length. The modulation of these mechanisms may underlie the clinical association between the exposure to recent Stressful life events in adulthood and an enhanced vulnerability to develop psychiatric diseases in adulthood.”