Pacific Symposium on Biocomputing 2016

Abstract Book

Poster Presenters: Poster space is assigned by abstract page number. Please find the page that your abstract is on and put your poster on the poster board with the corresponding number (e.g., if your abstract is on page 50, put your poster on board #50).

Proceedings papers with oral presentations #11-48 are not assigned poster space.

Papers are organized by session and then the last name of the first author.

Presenting authors' names are underlined in the Table of Contents and in the list of authors on the abstracts.

All authors appear in the Author Index.

TABLE OF CONTENTS

DISCOVERY OF MOLECULARLY TARGETED THERAPIES

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

AN INTEGRATED NETWORK APPROACH TO IDENTIFYING BIOLOGICAL PATHWAYS AND ENVIRONMENTAL EXPOSURE INTERACTIONS IN COMPLEX DISEASES

Christian Darabos[1,2], Jingya Qiu[1], Jason H. Moore[2]

[1]Dartmouth College; [2]University of Pennsylvania

Complex diseases are the result of intricate interactions between genetic, epigenetic and environmental factors. In previous studies, we used epidemiological and genetic data linking environmental exposure or genetic variants to phenotypic disease to construct Human Phenotype Networks and separately analyze the effects of both environment and genetic factors on disease interactions. To better capture the intricacies of the interactions between environmental exposure and the biological pathways in complex disorders, we integrate both aspects into a single "tripartite" network. Despite extensive research, the mechanisms by which chemical agents disrupt biological pathways are still poorly understood. In this study, we use our integrated network model to identify specific biological pathway candidates possibly disrupted by environmental agents. We conjecture that a higher number of co-occurrences between an environmental substance and biological pathway pair can be associated with a higher likelihood that the substance is involved in disrupting that pathway. We validate our model by demonstrating its ability to detect known arsenic and signal transduction pathway interactions and speculate on candidate cell-cell junction organization pathways disrupted by cadmium. The validation was supported by distinct publications of cell biology and genetic studies that associated environmental exposure to pathway disruption. The integrated network approach is a novel method for detecting the biological effects of environmental exposures. A better understanding of the molecular processes associated with specific environmental exposures will help in developing targeted molecular therapies for patients who have been exposed to the toxicity of environmental chemicals.

COMPUTING THERAPY FOR PRECISION MEDICINE: COLLABORATIVE FILTERING INTEGRATES AND PREDICTS MULTI-ENTITY INTERACTIONS

Sam Regenbogen, Angela D. Wilkins, Olivier Lichtarge

Baylor College of Medicine

Biomedicine produces copious information it cannot fully exploit. Specifically, there is considerable need to integrate knowledge from disparate studies to discover connections across domains. Here, we used a Collaborative Filtering approach, inspired by online recommendation algorithms, in which non-negative matrix factorization (NMF) predicts interactions among chemicals, genes, and diseases only from pairwise information about their interactions. Our approach, applied to matrices derived from the Comparative Toxicogenomics Database, successfully recovered Chemical-Disease, Chemical-Gene, and Disease-Gene networks in 10-fold cross-validation experiments. Additionally, we could predict each of these interaction matrices from the other two. Integrating all three CTD interaction matrices with NMF led to good predictions of STRING, an independent, external network of protein-protein interactions. Finally, this approach could integrate the CTD and STRING interaction data to improve Chemical-Gene cross-validation performance significantly, and, in a time-stamped study, it predicted information added to CTD after a given date, using only data prior to that date. We conclude that collaborative filtering can integrate information across multiple types of biological entities, and that as a first step towards precision medicine it can compute drug repurposing hypotheses.

INTEGRATING GENETIC AND STRUCTURAL DATA ON HUMAN PROTEIN KINOME IN NETWORK-BASED MODELING OF KINASE SENSITIVITIES AND RESISTANCE TO TARGETED AND PERSONALIZED ANTICANCER DRUGS

Gennady Verkhivker

Chapman University and University of California San Diego

The human protein kinome presents one of the largest protein families that orchestrate functional processes in complex cellular networks, and when perturbed, can cause various cancers. The abundance and diversity of genetic, structural, and biochemical data underlies the complexity of mechanisms by which targeted and personalized drugs can combat mutational profiles in protein kinases. Coupled with the evolution of system biology approaches, genomic and proteomic technologies are rapidly identifying and charactering novel resistance mechanisms with the goal to inform rationale design of personalized kinase drugs. Integration of experimental and computational approaches can help to bring these data into a unified conceptual framework and develop robust models for predicting the clinical drug resistance. In the current study, we employ a battery of synergistic computational approaches that integrate genetic, evolutionary, biochemical, and structural data to characterize the effect of cancer mutations in protein kinases. We provide a detailed structural classification and analysis of genetic signatures associated with oncogenic mutations. By integrating genetic and structural data, we employ network modeling to dissect mechanisms of kinase drug sensitivities to oncogenic EGFR mutations. Using biophysical simulations and analysis of protein structure networks, we show that conformational-specific drug binding of Lapatinib may elicit resistant mutations in the EGFR kinase that are linked with the ligand-mediated changes in the residue interaction networks and global network properties of key residues that are responsible for structural stability of specific functional states. A strong network dependency on high centrality residues in the conformation-specific Lapatinib-EGFR complex may explain vulnerability of drug binding to a broad spectrum of mutations and the emergence of drug resistance. Our study offers a systems-based perspective on drug design by unravelling complex relationships between robustness of targeted kinase genes and binding specificity of targeted kinase drugs. We discuss how these approaches can exploit advances in chemical biology and network science to develop novel strategies for rationally tailored and robust personalized drug therapies.

A FRAMEWORK FOR ATTRIBUTE-BASED COMMUNITY DETECTION WITH APPLICATIONS TO INTEGRATED FUNCTIONAL GENOMICS

Han Yu, Rachael Hageman Blair

Department of Biostatistics, University at Buffalo

Understanding community structure in networks has received considerable attention in recent years. Detecting and leveraging community structure holds promise for understanding and potentially intervening with the spread of influence. Network features of this type have important implications in a number of research areas, including, marketing, social networks, and biology. However, an overwhelming majority of traditional approaches to community detection cannot readily incorporate information of node attributes. Integrating structural and attribute information is a major challenge. We propose a flexible iterative method; inverse regularized Markov Clustering (irMCL), to network clustering via the manipulation of the transition probability matrix (aka stochastic flow) corresponding to a graph. Similar to traditional Markov Clustering, irMCL iterates between "expand" and "inflate" operations, which aim to strengthen the intra-cluster flow, while weakening the inter-cluster flow. Attribute information is directly incorporated into the iterative method through a sigmoid (logistic function) that naturally dampens attribute influence that is contradictory to the stochastic flow through the network. We demonstrate advantages and the flexibility of our approach using simulations and real data. We highlight an application that integrates breast cancer gene expression data set and a functional network defined via KEGG pathways reveal significant modules for survival.

COLLECTIVE PAIRWISE CLASSIFICATION FOR MULTI-WAY ANALYSIS OF DISEASE AND DRUG DATA

Marinka Zitnik, Blaz Zupan

University of Ljubljana

Interactions between drugs, drug targets or diseases can be predicted on the basis of molecular, clinical and genomic features by, for example, exploiting similarity of disease pathways, chemical structures, activities across cell lines or clinical manifestations of diseases. A successful way to better understand complex interactions in biomedical systems is to employ collective relational learning approaches that can jointly model diverse relationships present in multiplex data. We propose a novel collective pairwise classification approach for multi-way data analysis. Our model leverages the superiority of latent factor models and classifies relationships in a large relational data domain using a pairwise ranking loss. In contrast to current approaches, our method estimates probabilities, such that probabilities for existing relationships are higher than for assumed-to-be-negative relationships. Although our method bears correspondence with the maximization of non-differentiable area under the ROC curve, we were able to design a learning algorithm that scales well on multi-relational data encoding interactions between thousands of entities. We use the new method to infer relationships from multiplex drug data and to predict connections between clinical manifestations of diseases and their underlying molecular signatures. Our method achieves promising predictive performance when compared to state-of-the-art alternative approaches and can make "category-jumping" predictions about diseases from genomic and clinical data generated far outside the molecular context.

INNOVATIVE APPROACHES TO COMBINING GENOTYPE, PHENOTYPE, EPIGENETIC, AND EXPOSURE DATA FOR PRECISION DIAGNOSTICS

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

DIAGNOSIS-GUIDED METHOD FOR IDENTIFYING MULTI-MODALITY NEUROIMAGING BIOMARKERS ASSOCIATED WITH GENETIC RISK FACTORS IN ALZHEIMER'S DISEASE

Xiaoke Hao[1], Jingwen Yan[2], Xiaohui Yao[2], Shannon L. Risacher[2], Andrew J. Saykin[2], Daoqiang Zhang[1], Li Shen[1], for the ADNI[3]

[1]Nanjing University of Aeronautics and Astronautics; [2]Indiana University; [3]ADNI

Many recent imaging genetic studies focus on detecting the associations between genetic markers such as single nucleotide polymorphisms (SNPs) and quantitative traits (QTs). Although there exist a large number of generalized multivariate regression analysis methods, few of them have used diagnosis information in subjects to enhance the analysis performance. In addition, few of models have investigated the identification of multi-modality phenotypic patterns associated with interesting genotype groups in traditional methods. To reveal disease-relevant imaging genetic associations, we propose a novel diagnosis-guided multi-modality (DGMM) framework to discover multi-modality imaging QTs that are associated with both Alzheimer's disease (AD) and its top genetic risk factor (i.e., APOE SNP rs429358). The strength of our proposed method is that it explicitly models the priori diagnosis information among subjects in the objective function for selecting the disease-relevant and robust multi-modality QTs associated with the SNP. We evaluate our method on two modalities of imaging phenotypes, i.e., those extracted from structural magnetic resonance imaging (MRI) data and fluorodeoxyglucose positron emission tomography (FDG-PET) data in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The experimental results demonstrate that our proposed method not only achieves better performances under the metrics of root mean squared error and correlation coefficient but also can identify common informative regions of interests (ROIs) across multiple modalities to guide the disease-induced biological interpretation, compared with other reference methods.

METABOLOMICS DIFFERENTIAL CORRELATION NETWORK ANALYSIS OF OSTEOARTHRITIS

Ting Hu, Weidong Zhang, Zhaozhi Fan, Guang Sun, Sergei Likhodi, Edward Randell, Guangju Zhai

Department of Computer Science, Discipline of Genetics, Department of Mathematics and Statistics, Discipline of Medicine, Department of Laboratory Medicine, Department of Laboratory Medicine, Discipline of Genetics Memorial University, St. John's, NL, Canada

Osteoarthritis (OA) significantly compromises the life quality of affected individuals and imposes a substantial economic burden on our society. Unfortunately the pathogenesis of the disease is till poorly understood and no effective medications have been developed. OA is a complex disease that involves both genetic and environmental influences. To elucidate the complex interlinked structure of metabolic processes associated with OA, we developed a differential correlation network approach to detecting the interconnection of metabolite pairs whose relationships are significantly altered due to the diseased process. Through topological analysis of such a differential network, we identified key metabolites that played an important role in governing the connectivity and information flow of the network. Identification of these key metabolites suggests the association of their underlying cellular processes with OA and may help elucidate the pathogenesis of the disease and the development of novel targeted therapies.

INVESTIGATING THE IMPORTANCE OF ANATOMICAL HOMOLOGY FOR CROSS-SPECIES PHENOTYPE COMPARISONS USING SEMANTIC SIMILARITY.

Prashanti Manda[1], Christopher J. Mungall[2], James P. Balhoff[3], Hilmar Lapp[4], Todd J. Vision[1]

[1]University of North Carolina at Chapel Hill; [2]Lawrence Berkeley National Laboratory; [3]RTI International; [4]Duke University

There is growing use of ontologies for the measurement of cross-species phenotype similarity. Such similarity measurements contribute to diverse applications, such as identifying genetic models for human diseases, transferring knowledge among model organisms, and studying the genetic basis of evolutionary innovations. Two organismal features, whether genes, anatomical parts, or any other inherited feature, are considered to be homologous when they are evolutionarily derived from a single feature in a common ancestor. A classic example is the homology between the paired fins of fishes and vertebrate limbs. Anatomical ontologies that model the structural relations among parts may fail to include some known anatomical homologies unless they are deliberately added as separate axioms. The consequences of neglecting known homologies for applications that rely on such ontologies has not been well studied. Here, we examine how semantic similarity is affected when external homology knowledge is included. We measure phenotypic similarity between orthologous and non-orthologous gene pairs between humans and either mouse or zebrafish, and compare the inclusion of real with faux homology axioms. Semantic similarity was preferentially increased for orthologs when using real homology axioms, but only in the more divergent of the two species comparisons (human to zebrafish, not human to mouse), and the relative increase was less than 1% to non-orthologs. By contrast, inclusion of both real and faux random homology axioms preferentially increased similarities between genes that were initially more dissimilar in the other comparisons. Biologically meaningful increases in semantic similarity were seen for a select subset of gene pairs. Overall, the effect of including homology axioms on cross-species semantic similarity was modest at the levels of divergence examined here, but our results hint that it may be greater for more distant species comparisons.

DISCOVERING PATIENT PHENOTYPES USING GENERALIZED LOW RANK MODELS

Alejandro Schuler[1], Vincent Liu[1], Joe Wan[1], Alison Callahan[1], Madeleine Udell[2], David E. Stark[1], Nigam H. Shah[1]

[1]Stanford University; [2]California Institute of Technology

The practice of medicine is predicated on discovering commonalities or distinguishing characteristics among patients to inform corresponding treatment. Given a patient grouping (hereafter referred to as a phenotype), clinicians can implement a treatment pathway accounting for the underlying cause of disease in that phenotype. Traditionally, phenotypes have been discovered by intuition, experience in practice, and advancements in basic science, but these approaches are often heuristic, labor intensive, and can take decades to produce actionable knowledge. Although our understanding of disease has progressed substantially in the past century, there are still important domains in which our phenotypes are murky, such as in behavioral health or in hospital settings. To accelerate phenotype discovery, researchers have used machine learning to find patterns in electronic health records, but have often been thwarted by missing data, sparsity, and data heterogeneity. In this study, we use a flexible framework called Generalized Low Rank Modeling (GLRM) to overcome these barriers and discover phenotypes in two sources of patient data. First, we analyze data from the 2010 Healthcare Cost and Utilization Project National Inpatient Sample (NIS), which contains upwards of 8 million hospitalization records consisting of administrative codes and demographic information. Second, we analyze a small (N=1746), local dataset documenting the clinical progression of autism spectrum disorder patients using granular features from the electronic health record, including text from physician notes. We demonstrate that low rank modeling successfully captures known and putative phenotypes in these vastly different datasets.

INTEGRATING CLINICAL LABORATORY MEASURES AND ICD-9 CODE DIAGNOSES IN PHENOME-WIDE ASSOCIATION STUDIES

Anurag Verma[1,3], Joseph B. Leader[2], Shefali S. Verma[1,3], Alex T. Frase[3], John Wallace[3], Scott Dudek[3], Daniel R. Lavage[2], Cristopher V. Van Hout[4], Frederick E. Dewey[4], John Penn[4], Alex Lopez[4], John D. Overton[4], David J. Carey[5], David H. Ledbetter[1], H. Lester Kirchner[2], Marylyn D. Ritchie[1,3], Sarah A. Pendergrass[1]

[1]Biomedical and Translational Informatics, Geisinger Health System, Danville, PA; [2]Center for Health Research, Geisinger Health System, Danville, PA; [3]Center for Systems Genomics, The Pennsylvania State University, University Park, PA; [4]Regeneron Genetics Center, Tarrytown NY; [5]Weis Center for Research, Geisinger Health System, Danville, PA

Electronic health records (EHR) provide a comprehensive resource for discovery, allowing unprecedented exploration of the impact of genetic architecture on health and disease. The data of EHRs also allow for exploration of the complex interactions between health measures across health and disease. The discoveries arising from EHR based research provide important information for the identification of genetic variation for clinical decision-making. Due to the breadth of information collected within the EHR, a challenge for discovery using EHR based data is the development of high-throughput tools that expose important areas of further research, from genetic variants to phenotypes. Phenome-Wide Association studies (PheWAS) provide a way to explore the association between genetic variants and comprehensive phenotypic measurements, generating new hypotheses and also exposing the complex relationships between genetic architecture and outcomes, including pleiotropy. EHR based PheWAS have mainly evaluated associations with case/control status from International Classification of Disease, Ninth Edition (ICD-9) codes. While these studies have highlighted discovery through PheWAS, the rich resource of clinical lab measures collected within the EHR can be better utilized for high-throughput PheWAS analyses and discovery. To better use these resources and enrich PheWAS association results we have developed a sound methodology for extracting a wide range of clinical lab measures from EHR data. We have extracted a first set of 21 clinical lab measures from the de-identified EHR of participants of the Geisinger MyCodeTM biorepository, and calculated the median of these lab measures for 12,039 subjects. Next we evaluated the association between these 21 clinical lab median values and 635,525 genetic variants, performing a genome-wide association study (GWAS) for each of 21 clinical lab measures. We then calculated the association between SNPs from these GWAS passing our Bonferroni defined p-value cutoff and 165 ICD-9 codes. Through the GWAS we found a series of results replicating known associations, and also some potentially novel associations with less studied clinical lab measures. We found the majority of the PheWAS ICD-9 diagnoses highly related to the clinical lab measures associated with same SNPs. Moving forward, we will be evaluating further phenotypes and expanding the methodology for successful extraction of clinical lab measurements for research and PheWAS use. These developments are important for expanding the PheWAS approach for improved EHR based discovery.

METHODS TO ENHANCE THE REPRODUCIBILITY OF PRECISION MEDICINE

PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

REPRODUCIBLE RESEARCH WORKFLOW IN R FOR THE ANALYSIS OF PERSONALIZED HUMAN MICROBIOME DATA.

Benjamin Callahan, Diana Proctor, David Relman, Julia Fukuyama, Susan Holmes

Stanford University

This article presents a reproducible research workflow for amplicon-based microbiome studies in personalized medicine created using Bioconductor packages and the knitr markdown interface. We show that sometimes a multiplicity of choices and lack of consistent documentation at each stage of the sequential processing pipeline used for the analysis of microbiome data can lead to spurious results. We propose its replacement with reproducible and documented analysis using R packages dada2, knitr and phyloseq. This workflow implements both key stages of amplicon analysis: the initial filtering and denoising steps needed to construct taxonomic feature tables from error-containing sequencing reads (dada2), and the exploratory and inferential analysis of those feature tables and associated sample metadata (phyloseq). This workflow facilitates reproducible interrogation of the full set of choices required in microbiome studies. We present several examples in which we leverage existing packages for analysis in a way that allows easy sharing and modification by others, and give pointers to articles that depend on this reproducible workflow for the study of longitudinal and spatial series analyses of the vaginal microbiome in pregnancy and the oral microbiome in humans with healthy dentition and intra-oral tissues.

DYNAMICALLY EVOLVING CLINICAL PRACTICES AND IMPLICATIONS FOR PREDICTING MEDICAL DECISIONS

Jonathan H. Chen[1,2], Mary K. Goldstein[2,3], Steven M. Asch[1,4], Russ B. Altman[5,6]

[1]Center for Innovation to Implementation (Ci2i), Veteran Affairs Palo Alto Health Care System; [2]Center for Primary Care and Outcomes Research (PCOR), Stanford University; [3]Geriatrics Research Education and Clinical Center, Veteran Affairs Palo Alto Health Care System; [4]Division of General Medical Disciplines, Department of Internal Medicine, Stanford University; [5]Departments of Bioengineering and Genetics, Stanford University; [6]Department of Medicine, Stanford University

Automatically data-mining clinical practice patterns from electronic health records (EHR) can enable prediction of future practices as a form of clinical decision support (CDS). Our objective is to determine the stability of learned clinical practice patterns over time and what implication this has when using varying longitudinal historical data sources towards predicting future decisions. We trained an association rule engine for clinical orders (e.g., labs, imaging, medications) using structured inpatient data from a tertiary academic hospital. Comparing top order associations per admission diagnosis from training data in 2009 vs. 2012, we find practice variability from unstable diagnoses with rank biased overlap (RBO)<0.35 (e.g., pneumonia) to stable admissions for planned procedures (e.g., chemotherapy, surgery) with comparatively high RBO>0.6. Predicting admission orders for future (2013) patients with associations trained on recent (2012) vs. older (2009) data improved accuracy evaluated by area under the receiver operating characteristic curve (ROC-AUC) 0.89 to 0.92, precision at ten (positive predictive value of the top ten predictions against actual orders) 30% to 37%, and weighted recall (sensitivity) at ten 2.4% to 13%, (P<10-10). Training with more longitudinal data (2009-2012) was no better than only using recent (2012) data. Secular trends in practice patterns likely explain why smaller but more recent training data is more accurate at predicting future practices.

REPURPOSING GERMLINE EXOMES OF THE CANCER GENOME ATLAS DEMANDS A CAUTIOUS APPROACH AND SAMPLE-SPECIFIC VARIANT FILTERING

Amanda Koire, Panagiotis Katsonis, Olivier Lichtarge

Baylor College of Medicine

When seeking to reproduce results derived from whole-exome or genome sequencing data that could advance precision medicine, the time and expense required to produce a patient cohort make data repurposing an attractive option. The first step in repurposing is setting some quality baseline for the data so that conclusions are not spurious. This is difficult because there can be variations in quality from center to center, clinic to clinic and even patient to patient. Here, we assessed the quality of the whole-exome germline mutations of TCGA cancer patients using patterns of nucleotide substitution and negative selection against impactful mutations. We estimated the fraction of false positive variant calls for each exome with respect to two gold standard germline exomes, and found large variability in the quality of SNV calls between samples, cancer subtypes, and institutions. We then demonstrated how variant features, such as the average base quality for reads supporting an allele, can be used to identify sample-specific filtering parameters to optimize the removal of false positive calls. We concluded that while these germlines have many potential applications to precision medicine, users should assess the quality of the available exome data prior to use and perform additional filtering steps.

IDENTIFICATION OF QUESTIONABLE EXCLUSION CRITERIA IN MENTAL DISORDER CLINICAL TRIALS USING A MEDICAL ENCYCLOPEDIA

Handong Ma, Chunhua Weng

Department of Biomedical Informatics, Columbia University

Precision medicine requires precise evidence-based practice and precise definition of the patients included in clinical studies for evidence generalization. Clinical research exclusion criteria define confounder patient characteristics for exclusion from a study. However, unnecessary exclusion criteria can weaken patient representativeness of study designs and generalizability of study results. This paper presents a method for identifying questionable exclusion criteria for 38 mental disorders. We extracted common eligibility features (CEFs) from all trials on these disorders from ClinicalTrials.gov. Network Analysis showed scale-free property of the CEF network, indicating uneven usage frequencies among CEFs. By comparing these CEFs' term frequencies in clinical trials' exclusion criteria and in the PubMed Medical Encyclopedia for matching conditions, we identified unjustified potential overuse of exclusion CEFs in mental disorder trials. Then we discussed the limitations in current exclusion criteria designs and made recommendations for achieving more patient-centered exclusion criteria definitions.

REPRODUCIBLE AND SHAREABLE QUANTIFICATIONS OF PATHOGENICITY

Arjun K. Manrai[1], Brice L. Wang[2], Chirag J. Patel[1], Isaac S. Kohane[1]

[1]Harvard Medical School; [2]Illinois Mathematics and Science Academy

There are now hundreds of thousands of pathogenicity assertions that relate genetic variation to disease, but most of this clinically utilized variation has no accepted quantitative disease risk estimate. Recent disease-specific studies have used control sequence data to reclassify large amounts of prior pathogenic variation, but there is a critical need to scale up both the pace and feasibility of such pathogenicity reassessments across human disease. In this manuscript we develop a shareable computational framework to quantify pathogenicity assertions. We release a reproducible "digital notebook" that integrates executable code, text annotations, and mathematical expressions in a freely accessible statistical environment. We extend previous disease-specific pathogenicity assessments to over 6,000 diseases and 160,000 assertions in the ClinVar database. Investigators can use this platform to prioritize variants for reassessment and tailor genetic model parameters (such as prevalence and heterogeneity) to expose the uncertainty underlying pathogenicity-based risk assessments. Finally, we release a website that links users to pathogenic variation for a queried disease, supporting literature, and implied disease risk calculations subject to user-defined and disease-specific genetic risk models in order to facilitate variant reassessments.

PRECISION MEDICINE: DATA AND DISCOVERY FOR IMPROVED HEALTH AND THERAPY

PROCEEDINGS PAPERS WITH ORAL PRESENTATION

KNOWLEDGE DRIVEN BINNING AND PHEWAS ANALYSIS IN MARSHFIELD PERSONALIZED MEDICINE RESEARCH PROJECT USING BIOBIN

Anna O. Basile[1], John R. Wallace[1], Peggy Peissig[2], Catherine A. McCarty[3], Murray Brilliant[2], Marylyn D. Ritchie[1,4]

[1]Department of Biochemistry, Microbiology and Molecular Biology, The Pennsylvania State University University Park, PA,USA; [2]Bioinformatics Research Center, Marshfield Clinic, Marshfield, WI, USA; [3]Essentia Institute of Rural Health; [4]Department of Biomedical and Translational Informatics, Geisinger Health System

Next-generation sequencing technology has presented an opportunity for rare variant discovery and association of these variants with disease. To address the challenges of rare variant analysis, multiple statistical methods have been developed for combining rare variants to increase statistical power for detecting associations. BioBin is an automated tool that expands on collapsing/binning methods by performing multi-level variant aggregation with a flexible, biologically informed binning strategy using an internal biorepository, the Library of Knowledge (LOKI). The databases within LOKI provide variant details, regional annotations and pathway interactions which can be used to generate bins of biologically-related variants, thereby increasing the power of any subsequent statistical test. In this study, we expand the framework of BioBin to incorporate statistical tests, including a dispersion-based test, SKAT, thereby providing the option of performing a unified collapsing and statistical rare variant analysis in one tool. Extensive simulation studies performed on gene-coding regions showed a Bin-KAT analysis to have greater power than BioBin-regression in all simulated conditions, including variants influencing the phenotype in the same direction, a scenario where burden tests often retain greater power. The use of Madsen-Browning variant weighting increased power in the burden analysis to that equitable with Bin-KAT; but overall Bin-KAT retained equivalent or higher power under all conditions. Bin-KAT was applied to a study of 82 pharmacogenes sequenced in the Marshfield Personalized Medicine Research Project (PMRP). We looked for association of these genes with 9 different phenotypes extracted from the electronic health record. This study demonstrates that Bin-KAT is a powerful tool for the identification of genes harboring low frequency variants for complex phenotypes.

MULTITASK FEATURE SELECTION WITH TASK DESCRIPTORS

Victor Bellon, Veronique Stoven, Chloe-Agathe Azencott

MINES ParisTech, PSL-Research University, Institut Curie, INSERM U900

Machine learning applications in precision medicine are severly limited by the scarcity of data to learn from. Indeed, training data often contains many more features than samples. To alleviate the resulting statistical issues, the multitask learning framework proposes to learn different but related tasks joinlty, rather than independently, by sharing information between these tasks. Within this framework, the joint regularization of model parameters results in models with few non-zero coefficients and that share similar sparsity patterns. We propose a new regularized multitask approach that incorporates task descriptors, hence modulating the amount of information shared between tasks according to their similarity. We show on simulated data that this method outperforms other multitask feature selection approaches, particularly in the case of scarce data. In addition, we demonstrate on peptide MHC-I binding data the ability of the proposed approach to make predictions for new tasks for which no training data is available.

PERSONALIZED HYPOTHESIS TESTS FOR DETECTING MEDICATION RESPONSE IN PARKINSON DISEASE PATIENTS USING IPHONE SENSOR DATA

Elias Chaibub Neto, Brian M. Bot, Thanner Perumal, Larsson Omberg, Justin Guinney, Mike Kellen, Arno Klein, Stephen H. Friend, Andrew D. Trister

Sage Bionetworks

We propose hypothesis tests for detecting dopaminergic medication response in Parkinson disease patients, using longitudinal sensor data collected by smartphones. The processed data is composed of multiple features extracted from active tapping tasks performed by the participant on a daily basis, before and after medication, over several months. Each extracted feature corresponds to a time series of measurements annotated according to whether the measurement was taken before or after the patient has taken his/her medication. Even though the data is longitudinal in nature, we show that simple hypothesis tests for detecting medication response, which ignore the serial correlation structure of the data, are still statistically valid, showing type I error rates at the nominal level. We propose two distinct personalized testing approaches. In the first, we combine multiple feature-specific tests into a single union-intersection test. In the second, we construct personalized classifiers of the before/after medication labels using all the extracted features of a given participant, and test the null hypothesis that the area under the receiver operating characteristic curve of the classifier is equal to 1/2. We compare the statistical power of the personalized classifier tests and personalized union-intersection tests in a simulation study, and illustrate the performance of the proposed tests using data from mPower Parkinson's disease study, recently launched as part of Apple's ResearchKit mobile platform. Our results suggest that the personalized tests, which ignore the longitudinal aspect of the data, can perform well in real data analyses, suggesting they might be used as a sound baseline approach, to which more sophisticated methods can be compared to.

KIDNEY DISEASE GENETICS AND THE IMPORTANCE OF DIVERSITY IN PRECISION MEDICINE

Jessica N. Cooke Bailey[1], Sarah Wilson[2], Kristin Brown-Gentry[2], Robert Goodloe[2], Dana C. Crawford[1]

[1]Case Western Reserve University; [2]Vanderbilt University

Kidney disease is a well-known health disparity in the United States where African Americans are affected at higher rates compared with other groups such as European Americans and Mexican Americans. Common genetic variants in the myosin, heavy chain 9, non-muscle (MYH9) gene were initially identified as associated with non-diabetic end-stage renal disease in African Americans, and it is now understood that these variants are in strong linkage disequilibrium with likely causal variants in neighboring APOL1. Subsequent genome-wide and candidate gene studies have suggested that MYH9 common variants among others are also associated with chronic kidney disease and quantitative measures of kidney function in various populations. In a precision medicine setting, it is important to consider genetic effects or genetic associations that differ across racial/ethnic groups in delivering data relevant to disease risk or individual-level patient assessment. Kidney disease and quantitative trait-associated genetic variants have yet to be systematically characterized in multiple racial/ethnic groups. Therefore, to further characterize the prevalence of these genetic variants and their association with kidney related traits, we have genotyped 10 kidney disease or quantitative trait-associated single nucleotide polymorphisms (SNPs) (rs2900976, rs10505955, rs10502868, rs1243400, rs9305354, rs12917707, rs17319721, rs2467853, rs2032487, and rs4821480) in 14,998 participants from the population-based cross-sectional National Health and Nutrition Examination Surveys (NHANES) III and 1999-2002 as part of the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) study. In this general adult population ascertained regardless of health status (6,293 non-Hispanic whites, 3,013 non-Hispanic blacks, and 3,542 Mexican Americans), we observed higher rates of chronic kidney disease among non-Hispanic blacks compared with the other groups as expected. We performed single SNP tests of association using linear regressions assuming an additive genetic model adjusted for age, sex, diastolic blood pressure, systolic blood pressure, and type 2 diabetes status for several outcomes including creatinine (urinary), creatinine (serum), albumin (urinary), eGFR, and albumin-to-urinary creatinine ratio (ACR). We also tested for associations between each SNP and chronic kidney disease and albuminuria using logistic regression. Surprisingly, none of the MYH9 variants tested was associated with kidney diseases or traits in non-Hispanic blacks ($p>0.05$), perhaps attributable to the clinical heterogeneity of kidney disease in this population. Several associations were observed in each racial/ethnic group at $p<0.05$, but none were consistently associated in the same direction in all three groups. The lack of significant and consistent associations is most likely due to power highlighting the importance of the availability of large, diverse populations for genetic association studies of complex diseases and traits to inform precision medicine efforts in diverse patient populations.

PREDICTING SIGNIFICANCE OF UNKNOWN VARIANTS IN GLIAL TUMORS THROUGH SUB-CLASS ENRICHMENT

Alex Fichtenholtz, Nicholas Camarda, Eric Neumann

Foundation Medicine

Glial tumors have been heavily studied and sequenced, leading to scores of findings about altered genes. This explosion in knowledge has not been matched with clinical success, but efforts to understand the synergies between drivers of glial tumors may alleviate the situation. We present a novel molecular classification system that captures the combinatorial nature of relationships between alterations in these diseases. We use this classification to mine for enrichment of variants of unknown significance, and demonstrate a method for segregating unknown variants with functional importance from passengers and SNPs.

PATIENT-SPECIFIC DATA FUSION FOR CANCER STRATIFICATION AND PERSONALIZED TREATMENT

Vladimir Gligorijevic, Noel Malod-Dognin, Natasa Przulj

Imperial College London

According to Cancer Research UK, cancer is a leading cause of death accounting for more than one in four of all deaths in 2011. The recent advances in experimental technologies in cancer research have resulted in the accumulation of large amounts of patient-specific datasets, which provide complementary information on the same cancer type. We introduce a versatile data fusion (integration) framework that can effectively integrate somatic mutation data, molecular interactions and drug chemical data to address three key challenges in cancer research: stratification of patients into groups having different clinical outcomes, prediction of driver genes whose mutations trigger the onset and development of cancers, and repurposing of drugs treating particular cancer patient groups. Our new framework is based on graph-regularised non-negative matrix tri-factorization, a machine learning technique for co-clustering heterogeneous datasets. We apply our framework on ovarian cancer data to simultaneously cluster patients, genes and drugs by utilising all datasets. We demonstrate superior performance of our method over the state-of-the-art method, Network-based Stratification, in identifying three patient subgroups that have significant differences in survival outcomes and that are in good agreement with other clinical data. Also, we identify potential new driver genes that we obtain by analysing the gene clusters enriched in known drivers of ovarian cancer progression. We validated the top scoring genes identified as new drivers through database search and biomedical literature curation. Finally, we identify potential candidate drugs for repurposing that could be used in treatment of the identified patient subgroups by targeting their mutated gene products. We validated a large percentage of our drug-target predictions by using other databases and through literature curation.

PRISM: A DATA-DRIVEN PLATFORM FOR MONITORING MENTAL HEALTH

Maulik Kamdar, Michelle Wu

Stanford University

Neuropsychiatric disorders are the leading cause of disability worldwide and there is no gold standard currently available for the measurement of mental health. This issue is exacerbated by the fact that the information physicians use to diagnose these disorders is episodic and often subjective. Current methods to monitor mental health involve the use of subjective DSM-5 guidelines, and advances in EEG and video monitoring technologies have not been widely adopted due to invasiveness and inconvenience. Wearable technologies have surfaced as a ubiquitous and unobtrusive method for providing continuous, quantitative data about a patient. Here, we introduce PRISM --- Passive, Real-time Information for Sensing Mental Health. This platform integrates motion, light and heart rate data from a smart watch application with user interactions and text entries from a web application. We have demonstrated a proof of concept by collecting preliminary data through a pilot study of 13 subjects. We have engineered appropriate features and applied both unsupervised and supervised learning to develop models that are predictive of user-reported ratings of their emotional state, demonstrating that the data has the potential to be useful for evaluating mental health. This platform could allow patients and clinicians to leverage continuous streams of passive data for early and accurate diagnosis as well as constant monitoring of patients suffering from mental disorders.

BAYESIAN BICLUSTERING FOR PATIENT STRATIFICATION

Sahand Khakabimamaghani, Martin Ester

Simon Fraser University

The move from Empirical Medicine towards Personalized Medicine has attracted attention to Stratified Medicine (SM). Some methods are provided in the literature for patient stratification, which is the central task of SM, however, there are still significant open issues. First, it is still unclear if integrating different datatypes will help in detecting disease subtypes more accurately, and, if not, which datatype(s) are most useful for this task. Second, it is not clear how we can compare different methods of patient stratification. Third, as most of the proposed stratification methods are deterministic, there is a need for investigating the potential benefits of applying probabilistic methods. To address these issues, we introduce a novel integrative Bayesian biclustering method, called B2PS, for patient stratification and propose methods for evaluating the results. Our experimental results demonstrate the superiority of B2PS over a popular state-of-the-art method and the benefits of Bayesian approaches. Our results agree with the intuition that transcriptomic data forms a better basis for patient stratification than genomic data.

BIOFILTER AS A FUNCTIONAL ANNOTATION PIPELINE FOR COMMON AND RARE COPY NUMBER BURDEN

Dokyoon Kim[1], Anastasia Lucas[1], Joseph Glessner[2], Shefali S. Verma[1], Yuki Bradford[1], Ruowang Li[1], Alex T. Frase[1], Hakon Hakonarson[2], Peggy Peissig[3], Murray Brilliant[3], Marylyn D. Ritchie[1]

[1]Penn State; [2]University of Pennsylvania; [3]Marshfield Clinic

Recent studies on copy number variation (CNV) have suggested that an increasing burden of CNVs is associated with susceptibility or resistance to disease. A large number of genes or genomic loci contribute to complex diseases such as autism. Thus, total genomic copy number burden, as an accumulation of copy number change, is a meaningful measure of genomic instability to identify the association between global genetic effects and phenotypes of interest. However, no systematic annotation pipeline has been developed to interpret biological meaning based on the accumulation of copy number change across the genome associated with a phenotype of interest. In this study, we develop a comprehensive and systematic pipeline for annotating copy number variants into genes/genomic regions and subsequently pathways and other gene groups using Biofilter – a bioinformatics tool that aggregates over a dozen publicly available databases of prior biological knowledge. Next we conduct enrichment tests of biologically defined groupings of CNVs including genes, pathways, Gene Ontology, or protein families. We applied the proposed pipeline to a CNV dataset from the Marshfield Clinic Personalized Medicine Research Project (PMRP) in a quantitative trait phenotype derived from the electronic health record – total cholesterol. We identified several significant pathways such as toll-like receptor signaling pathway and hepatitis C pathway, gene ontologies (GOs) of nucleoside triphosphatase activity (NTPase) and response to virus, and protein families such as cell morphogenesis that are associated with the total cholesterol phenotype based on CNV profiles (permutation p-value < 0.01). Based on the copy number burden analysis, it follows that the more and larger the copy number changes, the more likely that one or more target genes that influence disease risk and phenotypic severity will be affected. Thus, our study suggests the proposed enrichment pipeline could improve the interpretability of copy number burden analysis where hundreds of loci or genes contribute toward disease susceptibility via biological knowledge groups such as pathways. This CNV annotation pipeline with Biofilter can be used for CNV data from any genotyping or sequencing platform and to explore CNV enrichment for any traits or phenotypes. Biofilter continues to be a powerful bioinformatics tool for annotating, filtering, and constructing biologically informed models for association analysis – now including copy number variants.

THE CHALLENGES IN USING ELECTRONIC HEALTH RECORDS FOR PHARMACOGENOMICS AND PRECISION MEDICINE RESEARCH

Sarah M. Laper[1], Nicole A. Restrepo[2], Dana C. Crawford[3]

[1]Eastern Virginia Medical School; [2]Vanderbilt University; [3]Case Western Reserve University

Access and utilization of electronic health records with extensive medication lists and genetic profiles is rapidly advancing discoveries in pharmacogenomics. In this study, we analyzed ~116,000 variants on the Illumina Metabochip for response to antihypertensive and lipid lowering medications in African American adults from BioVU, the Vanderbilt University Medical Center's biorepository linked to de-identified electronic health records. Our study population included individuals who were prescribed an antihypertensive or lipid lowering medication, and who had both pre- and post-medication blood pressure or low-density lipoprotein cholesterol (LDL-C) measurements, respectively. Among those with pre- and post-medication systolic and diastolic blood pressure measurements (n=2,268), the average change in systolic and diastolic blood pressure was -0.6 mg Hg and -0.8 mm Hg, respectively. Among those with pre- and post-medication LDL-C measurements (n=1,244), the average change in LDL-C was -26.3 mg/dL. SNPs were tested for an association with change and percent change in blood pressure or blood levels of LDL-C. After adjustment for multiple testing, we did not observe any significant associations, and we were not able to replicate previously reported associations, such as in APOE and LPA, from the literature. The present study illustrates the benefits and challenges with using electronic health records linked to biorepositories for pharmacogenomic studies.

SEPARATING THE CAUSES AND CONSEQUENCES IN DISEASE TRANSCRIPTOME

Yong Fuga Li[1], Fuxiao Xin[2], Russ B. Altman[1]

[1]Stanford University; [2]GE Global Research

The causes of complex diseases are multifactorial and the phenotypes of complex diseases are typically heterogeneous, posting significant challenges for both the experiment design and statistical inference in the study of such diseases. Transcriptome profiling can potentially provide key insights on the pathogenesis of diseases, but the signals from the disease causes and consequences are intertwined, leaving it to speculations what are likely causal. Genome-wide association study on the other hand provides direct evidences on the potential genetic causes of diseases, but it does not provide a comprehensive view of disease pathogenesis, and it has difficulties in detecting the weak signals from individual genes. Here we propose an approach diseaseExPatho that combines transcriptome data, regulome knowledge, and GWAS results if available, for separating the causes and consequences in the disease transcriptome. DiseaseExPatho computationally de-convolutes the expression data into gene expression modules, hierarchically ranks the modules based on regulome using a novel algorithm, and given GWAS data, it directly labels the potential causal gene modules based on their correlations with genome-wide gene-disease associations. Strikingly, we observed that the putative causal modules are not necessarily differentially expressed in disease, while the other modules can show strong differential expression without enrichment of top GWAS variations. On the other hand, we showed that the regulatory network based module ranking prioritized the putative causal modules consistently in 6 diseases, We suggest that the approach is applicable to other common and rare complex diseases to prioritize causal pathways with or without genome-wide association studies.

ONE-CLASS DETECTION OF CELL STATES IN TUMOR SUBTYPES

Artem Sokolov, Evan O. Paull, Joshua M. Stuart

University of California, Santa Cruz

The cellular composition of a tumor greatly influences the growth, spread, immune activity, drug response, and other aspects of the disease. Tumor cells are usually comprised of a heterogeneous mixture of subclones, each of which could contain their own distinct character. The presence of minor subclones poses a serious health risk for patients as any one of them could harbor a fitness advantage with respect to the current treatment regimen, fueling resistance. It is therefore vital to accurately assess the make-up of cell states within a tumor biopsy. Transcriptome-wide assays from RNA sequencing provide key data from which cell state signatures can be detected. However, the challenge is to find them within samples containing mixtures of cell types of unknown proportions. We propose a novel one-class method based on logistic regression and show that its performance is competitive to two established SVM-based methods for this detection task. We demonstrate that one-class models are able to identify specific cell types in heterogeneous cell populations better than their binary predictor counterparts. We derive one-class predictors for the major breast and bladder subtypes and reaffirm the connection between these two tissues. In addition, we use a one-class predictor to quantitatively associate an embryonic stem cell signature with an aggressive breast cancer subtype that reveals shared stemness pathways potentially important for treatment.

SOCIAL MEDIA MINING FOR PUBLIC HEALTH MONITORING AND SURVEILLANCE

PROCEEDINGS PAPERS WITH ORAL PRESENTATION

TEXT CLASSIFICATION FOR AUTOMATIC DETECTION OF E-CIGARETTE USE AND USE FOR SMOKING CESSATION FROM TWITTER: A FEASIBILITY PILOT

Yin Aphinyanaphongs[1], Armine Lulejian[1], Duncan Penfold-Brown[2], Richard Bonneau[3], Paul Krebs[1]

[1]NYU Langone Medical Center; [2]NYU Social Media and Political Participation Lab; [3]Simons Center for Data Analysis

Rapid increases in e-cigarette use and potential exposure to harmful byproducts have shifted public health focus to e-cigarettes as a possible drug of abuse. Effective surveillance of use and prevalence would allow appropriate regulatory responses. An ideal surveillance system would collect usage data in real time, focus on populations of interest, include populations unable to take the survey, allow a breadth of questions to answer, and enable geo-location analysis. Social media streams may provide this ideal system. To realize this use case, a foundational question is whether we can detect ecigarette use at all. This work reports two pilot tasks using text classification to identify automatically Tweets that indicate e-cigarette use and/or e-cigarette use for smoking cessation. We build and define both datasets and compare performance of 4 state of the art classifiers and a keyword search for each task. Our results demonstrate excellent classifier performance of up to 0.90 and 0.94 area under the curve in each category. These promising initial results form the foundation for further studies to realize the ideal surveillance solution.

MONITORING POTENTIAL DRUG INTERACTIONS AND REACTIONS VIA NETWORK ANALYSIS OF INSTAGRAM USER TIMELINES

Rion Brattig Correia[1], Lang Li[2], Luis M. Rocha[1]

[1]Indiana University; [2]Indiana University School of Medicine

Much recent research aims to identify evidence for Drug-Drug Interactions (DDI) and Adverse Drug reactions (ADR) from the biomedical scientific literature. In addition to this "Bibliome", the universe of social media provides a very promising source of large-scale data that can help identify DDI and ADR in ways that have not been hitherto possible. Given the large number of users, analysis of social media data may be useful to identify under-reported, population-level pathology associated with DDI, thus further contributing to improvements in population health. Moreover, tapping into this data allows us to infer drug interactions with natural products---including cannabis---which constitute an array of DDI very poorly explored by biomedical research thus far. Our goal is to determine the potential of Instagram for public health monitoring and surveillance for DDI, ADR, and behavioral pathology at large. Most social media analysis focuses on Twitter and Facebook, but Instagram is an increasingly important platform, especially among teens, with unrestricted access of public posts, high availability of posts with geolocation coordinates, and images to supplement textual analysis. Using drug, symptom, and natural product dictionaries for identification of the various types of DDI and ADR evidence, we have collected close to 7000 user timelines spanning from October 2010 to June 2015.We report on 1) the development of a monitoring tool to easily observe user-level timelines associated with drug and symptom terms of interest, and 2) population-level behavior via the analysis of co-occurrence networks computed from user timelines at three different scales: monthly, weekly, and daily occurrences. Analysis of these networks further reveals 3) drug and symptom direct and indirect associations with greater support in user timelines, as well as 4) clusters of symptoms and drugs revealed by the collective behavior of the observed population. This demonstrates that Instagram contains much drug- and pathology specific data for public health monitoring of DDI and ADR, and that complex network analysis provides an important toolbox to extract health-related associations and their support from large-scale social media data.

TOWARDS EARLY DISCOVERY OF SALIENT HEALTH THREATS: A SOCIAL MEDIA EMOTION CLASSIFICATION TECHNIQUE

Bahadorreza Ofoghi, Meghan Mann, Karin Verspoor

The University of Melbourne

Online social media microblogs may be a valuable resource for timely identification of critical ad hoc health-related incidents or serious epidemic outbreaks. In this paper, we explore emotion classification of Twitter microblogs related to localized public health threats, and study whether the public mood can be effectively utilized in early discovery or alarming of such events. We analyse user tweets around recent incidents of Ebola, finding differences in the expression of emotions in tweets posted prior to and after the incidents have emerged. We also analyse differences in the nature of the tweets in the immediately affected area as compared to areas remote to the events. The results of this analysis suggest that emotions in social media microblogging data (from Twitter in particular) may be utilized effectively as a source of evidence for disease outbreak detection and monitoring.

PREDICTING INDIVIDUAL WELL-BEING THROUGH THE LANGUAGE OF SOCIAL MEDIA

H. Andrew Schwartz[1], Maarten Sap[2], Margaret L. Kern[3], Johannes C. Eichstaedt[2], Adam Kapelner[4], Megha Agrawal[2], Eduardo Blanco[5], Lukasz Dziurzynski[2], Gregory Park[2], David Stillwell[6], Michal Kosinski[6], Martin E.P. Seligman[2], Lyle H. Ungar[2]

[1]Stony Brook University; [2]University of Pennsylvania; [3]University of Melbourne; [4]Queens College; [5]University of North Texas; [6]University of Cambridge

We present the task of predicting individual well-being, as measured by a life satisfaction scale, through the language people use on social media. Well-being, which encompasses much more than emotion and mood, is linked with good mental and physical health. The ability to quickly and accurately assess it can supplement multi-million dollar national surveys as well as promote whole body health. Through crowd-sourced ratings of tweets and Facebook status updates, we create message-level predictive models for multiple components of well-being. However, well-being is ultimately attributed to people, so we perform an additional evaluation at the user-level, finding that a multi-level cascaded model, using both message-level predictions and user-level features, performs best and outperforms popular lexicon-based happiness models. Finally, we suggest that analyses of language go beyond prediction by identifying the language that characterizes well-being.

FINDING POTENTIALLY UNSAFE NUTRITIONAL SUPPLEMENTS FROM USER REVIEWS WITH TOPIC MODELING

Ryan Sullivan, Abeed Sarker, Karen O'Connor, Amanda Goodin, Mark Karlsrud, Graciela Gonzalez

Department of Biomedical Informatics, Arizona State University

Although dietary supplements are widely used and generally are considered safe, some supplements have been identified as causative agents for adverse reactions, some of which may even be fatal. The Food and Drug Administration (FDA) is responsible for monitoring supplements and ensuring that supplements are safe. However, current surveillance protocols are not always effective. Leveraging user-generated textual data, in the form of Amazon.com reviews for nutritional supplements, we use natural language processing techniques to develop a system for the monitoring of dietary supplements. We use topic modeling techniques, specially a variation of Latent Dirichlet Allocation (LDA), and background knowledge in the form of an adverse reaction dictionary to score products based on their potential danger to the public. Our approach generates topics that semantically capture adverse reactions from a document set consisting of reviews posted by users of specific products, and based on these topics, we propose a scoring mechanism to categorize products as "high potential danger", "average potential danger" and "low potential danger." We evaluate our system by comparing the system categorization with human annotators, and we find that the our system agrees with the annotators 69.4% of the time. With these results, we demonstrate that our methods show promise and that our system represents a proof of concept as a viable low-cost, active approach for dietary supplement monitoring.

INSIGHTS FROM MACHINE-LEARNED DIET SUCCESS PREDICTION

Ingmar Weber[1], Palakorn Achananuparp[2]

[1]Qatar Computing Research Institute; [2]Singapore Management University

To support people trying to lose weight and stay healthy, more and more fitness apps have sprung up including the ability to track both calories intake and expenditure. Users of such apps are part of a wider "quantified self" movement and many opt-in to publicly share their logged data. In this paper, we use public food diaries of more than 4,000 long-term active MyFitnessPal users to study the characteristics of a (un-)successful diet. Concretely, we train a machine learning model to predict repeatedly being over or under self-set daily calories goals and then look at which features contribute to the model's prediction. Our findings include both expected results, such as the token "mcdonalds" or the category "dessert" being indicative for being over the calories goal, but also less obvious ones such as the difference between pork and poultry concerning dieting success, or the use of the "quick added calories" functionality being indicative of over-shooting calorie-wise. This study also hints at the feasibility of using such data for more in-depth data mining, e.g., looking at the interaction between consumed foods such as mixing protein- and carbohydrate-rich foods. To the best of our knowledge, this is the first systematic study of public food diaries.

DISCOVERY OF MOLECULARLY TARGETED THERAPIES

PROCEEDINGS PAPER WITH POSTER PRESENTATION

KNOWLEDGE-ASSISTED APPROACH TO IDENTIFY PATHWAYS WITH DIFFERENTIAL DEPENDENCIES

Gil Speyer, Jeff Kiefer, Harshil Dhruv, Michael Berens, Seungchan Kim

Translational Genomics Research Institute

We have previously developed a statistical method to identify gene sets enriched with condition-specific genetic dependencies. The method constructs gene dependency networks from bootstrapped samples in one condition and computes the divergence between distributions of network likelihood scores from different conditions. It was shown to be capable of sensitive and specific identification of pathways with phenotype-specific dysregulation, i.e., rewiring of dependencies between genes in different conditions. We now present an extension of the method by incorporating prior knowledge into the inference of networks. The degree of prior knowledge incorporation has substantial effect on the sensitivity of the method, as the data is the source of condition specificity while prior knowledge incorporation can provide additional support for dependencies that are only partially supported by the data. Use of prior knowledge also significantly improved the interpretability of the results. Further analysis of topological characteristics of gene differential dependency networks provides a new approach to identify genes that could play important roles in biological signaling in a specific condition, hence, promising targets customized to a specific condition. Through analysis of TCGA glioblastoma multiforme data, we demonstrate the method can identify not only potentially promising targets but also underlying biology for new targets.

PHENOME-WIDE INTERACTION STUDY (PHEWIS) IN AIDS CLINICAL TRIALS GROUP DATA (ACTG)

Shefali S. Verma[1], Alex T. Frase[1], Anurag Verma[1], Sarah A. Pendergrass[2], Shaun Mahony[1], David W. Haas[3], Marylyn D. Ritchie[1,2]

[1]Center for System Genomics, The Pennsylvania State University, University Park, PA 16802 USA; [2]Biomedical and Translational Informatics, Geisinger Health System, Danville, PA 17822 USA; [3]Vanderbilt Health, One Hundred Oaks, 719 Thompson Lane, Suite 47183, Nashville, TN 37204 USA

Association studies have shown and continue to show a substantial amount of success in identifying links between multiple single nucleotide polymorphisms (SNPs) and phenotypes. These studies are also believed to provide insights toward identification of new drug targets and therapies. Albeit of all the success, challenges still remain for applying and prioritizing these associations based on available biological knowledge. Along with single variant association analysis, genetic interactions also play an important role in uncovering the etiology and progression of complex traits. For gene-gene interaction analysis, selection of the variants to test for associations still poses a challenge in identifying epistatic interactions among the large list of variants available in high-throughput, genome-wide datasets. Therefore in this study, we propose a pipeline to identify interactions among genetic variants that are associated with multiple phenotypes by prioritizing previously published results from main effect association analysis (genome-wide and phenome-wide association analysis) based on a-priori biological knowledge in AIDS Clinical Trials Group (ACTG) data. We approached the prioritization and filtration of variants by using the results of a previously published single variant PheWAS and then utilizing biological information from the Roadmap Epigenome project. We removed variants in low functional activity regions based on chromatin states annotation and then conducted an exhaustive pairwise interaction search using linear regression analysis. We performed this analysis in two independent pre-treatment clinical trial datasets from ACTG to allow for both discovery and replication. Using a regression framework, we observed 50,798 associations that replicate at p-value 0.01 for 26 phenotypes, among which 2,176 associations for 212 unique SNPs for fasting blood glucose phenotype reach Bonferroni significance and an additional 9,970 interactions for high-density lipoprotein (HDL) phenotype and fasting blood glucose (total of 12,146 associations) reach FDR significance. We conclude that this method of prioritizing variants to look for epistatic interactions can be used extensively for generating hypotheses for genome-wide and phenome-wide interaction analyses. This original Phenome-wide Interaction study (PheWIS) can be applied further to patients enrolled in randomized clinical trials to establish the relationship between patient's response to a particular drug therapy and non-linear combination of variants that might be affecting the outcome.

INNOVATIVE APPROACHES TO COMBINING GENOTYPE, PHENOTYPE, EPIGENETIC, AND EXPOSURE DATA FOR PRECISION DIAGNOSTICS

PROCEEDINGS PAPER WITH POSTER PRESENTATION

TESTING POPULATION-SPECIFIC QUANTITATIVE TRAIT ASSOCIATIONS FOR CLINICAL OUTCOME RELEVANCE IN A BIOREPOSITORY LINKED TO ELECTRONIC HEALTH RECORDS: LPA AND MYOCARDIAL INFARCTION IN AFRICAN AMERICANS

Logan Dumitrescu[1], Kirsten E. Diggins[1], Robert Goodloe[1], Dana C. Crawford[2]

[1]Vanderbilt University; [2]Case Western Reserve University

Previous candidate gene and genome-wide association studies have identified common genetic variants in LPA associated with the quantitative trait Lp(a), an emerging risk factor for cardiovascular disease. These associations are population-specific and many have not yet been tested for association with the clinical outcome of interest. To fill this gap in knowledge, we accessed the epidemiologic Third National Health and Nutrition Examination Surveys (NHANES III) and BioVU, the Vanderbilt University Medical Center biorepository linked to de-identified electronic health records (EHRs), including billing codes (ICD-9-CM) and clinical notes, to test population-specific Lp(a)-associated variants for an association with myocardial infarction (MI) among African Americans. We performed electronic phenotyping among African Americans in BioVU ≥40 years of age using billing codes. At total of 93 cases and 522 controls were identified in NHANES III and 265 cases and 363 controls were identified in BioVU. We tested five known Lp(a)-associated genetic variants (rs1367211, rs41271028, rs6907156, rs10945682, and rs1652507) in both NHANES III and BioVU for association with myocardial infarction. We also tested LPA rs3798220 (I4399M), previously associated with increased levels of Lp(a), MI, and coronary artery disease in European Americans, in BioVU. After meta-analysis, tests of association using logistic regression assuming an additive genetic model revealed no significant associations ($p<0.05$) for any of the five LPA variants previously associated with Lp(a) levels in African Americans. Also, I4399M rs3798220 was not associated with MI in African Americans (odds ratio = 0.51; 95% confidence interval: 0.16 – 1.65; $p=0.26$) despite strong, replicated associations with MI and coronary artery disease in European American genome-wide association studies. These data highlight the challenges in translating quantitative trait associations to clinical outcomes in diverse populations using large epidemiologic and clinic-based collections as envisioned for the Precision Medicine Initiative.

INFERENCE OF PERSONALIZED DRUG TARGETS VIA NETWORK PROPAGATION

Ortal Shnaps, Eyal Perry, Dana Silverbush, Roded Sharan

School of Computer Science, Tel Aviv University

We present a computational strategy to simulate drug treatment in a personalized setting. The method is based on integrating patient mutation and differential expression data with a protein-protein interaction network. We test the impact of in-silico deletions of different proteins on the flow of information in the network and use the results to infer potential drug targets. We apply our method to AML data from TCGA and validate the predicted drug targets using known targets. To benchmark our patient-specific approach, we compare the personalized setting predictions to those of the conventional setting. Our predicted drug targets are highly enriched with known targets from DrugBank and COSMIC ($p < 10^{-5}$), outperforming the non-personalized predictions. Finally, we focus on the largest AML patient subgroup (~30%) which is characterized by an FLT3 mutation, and utilize our prediction score to rank patient sensitivity to inhibition of each predicted target, reproducing previous findings of in-vitro experiments.

PRECISION MEDICINE: DATA AND DISCOVERY FOR IMPROVED HEALTH AND THERAPY

PROCEEDINGS PAPER WITH POSTER PRESENTATION

DO CANCER CLINICAL TRIAL POPULATIONS TRULY REPRESENT CANCER PATIENTS? A COMPARISON OF OPEN CLINICAL TRIALS TO THE CANCER GENOME ATLAS

Nophar Geifman, Atul J. Butte

Institute for Computational Health Sciences, University of California, San Francisco

Open clinical trial data offer many opportunities for the scientific community to independently verify published results, evaluate new hypotheses and conduct meta-analyses. These data provide a springboard for scientific advances in precision medicine but the question arises as to how representative clinical trials data are of cancer patients overall. Here we present the integrative analysis of data from several cancer clinical trials and compare these to patient-level data from The Cancer Genome Atlas (TCGA). Comparison of cancer type-specific survival rates reveals that these are overall lower in trial subjects. This effect, at least to some extent, can be explained by the more advanced stages of cancer of trial subjects. This analysis also reveals that for stage IV cancer, colorectal cancer patients have a better chance of survival than breast cancer patients. On the other hand, for all other stages, breast cancer patients have better survival than colorectal cancer patients. Comparison of survival in different stages of disease between the two datasets reveals that subjects with stage IV cancer from the trials dataset have a lower chance of survival than matching stage IV subjects from TCGA. One likely explanation for this observation is that stage IV trial subjects have lower survival rates since their cancer is less likely to respond to treatment. To conclude, we present here a newly available clinical trials dataset which allowed for the integration of patient-level data from many cancer clinical trials. Our comprehensive analysis reveals that cancer-related clinical trials are not representative of general cancer patient populations, mostly due to their focus on the more advanced stages of the disease. These and other limitations of clinical trials data should, perhaps, be taken into consideration in medical research and in the field of precision medicine.

A BAYESIAN NONPARAMETRIC MODEL FOR RECONSTRUCTING TUMOR SUBCLONES BASED ON MUTATION PAIRS

Subhajit Sengupta[1], Tianjian Zhou[2], Peter Mueller[3], Yuan Ji[1,4]

[1]Program for Computational Genomics and Medicine, NorthShore University HealthSystem; [2]Department of Statistics and Data Sciences, The University of Texas at Austin; [3]Department of Mathematics, The University of Texas at Austin; [4]Department of Public Health Sciences, The University of Chicago

We present a feature allocation model to reconstruct tumor subclones based on mutation pairs. The key innovation lies in the use of a pair of proximal single nucleotide variants (SNVs) for the subclone reconstruction as opposed to a single SNV. Using the categorical extension of the Indian buffet process (cIBP) we define the subclones as a vector of categorical matrices corresponding to a set of mutation pairs. Through Bayesian inference we report posterior probabilities of the number, genotypes and population frequencies of subclones in one or more tumor sample. We demonstrate the proposed methods using simulated and real-world data. A free software package is available at http://www.compgenome.org/pairclone.

RDF SKETCH MAPS - KNOWLEDGE COMPLEXITY REDUCTION FOR PRECISION MEDICINE ANALYTICS

Nattapon Thanintorn, Juexin Wang, Ilker Ersoy, Zainab Al-Taie, Yuexu Jiang, Duolin Wang, Mega Verma, Trupti Joshi, Richard Hammer, Dong Xu, Dmitriy Shin

University of Missouri

Realization of precision medicine ideas requires significant research effort to be able to spot subtle differences in complex diseases at the molecular level to develop personalized therapies. It is especially important in many cases of highly heterogeneous cancers. Precision diagnostics and therapeutics of such diseases demands interrogation of vast amounts of biological knowledge coupled with novel analytic methodologies. For instance, pathway-based approaches can shed light on the way tumorigenesis takes place in individual patient cases and pinpoint to novel drug targets. However, comprehensive analysis of hundreds of pathways and thousands of genes creates a combinatorial explosion, that is challenging for medical practitioners to handle at the point of care. Here we extend our previous work on mapping clinical omics data to curated Resource Description Framework (RDF) knowledge bases to derive influence diagrams of interrelationships of biomarker proteins, diseases and signal transduction pathways for personalized theranostics. We present RDF Sketch Maps – a computational method to reduce knowledge complexity for precision medicine analytics. The method of RDF Sketch Maps is inspired by the way a sketch artist conveys only important visual information and discards other unnecessary details. In our case, we compute and retain only so-called RDF Edges – places with highly important diagnostic and therapeutic information. To do this we utilize 35 maps of human signal transduction pathways by transforming 300 KEGG maps into highly processable RDF knowledge base. We have demonstrated potential clinical utility of RDF Sketch Maps in hematopoietic cancers, including analysis of pathways associated with Hairy Cell Leukemia (HCL) and Chronic Myeloid Leukemia (CML) where we achieved up to 20-fold reduction in the number of biological entities to be analyzed, while retaining most likely important entities. In experiments with pathways associated with HCL a generated RDF Sketch Map of the top 30% paths retained important information about signaling cascades leading to activation of proto-oncogene BRAF, which is usually associated with a different cancer, melanoma. Recent reports of successful treatments of HCL patients by the BRAF-targeted drug vemurafenib support the validity of the RDF Sketch Maps findings. We therefore believe that RDF Sketch Maps will be invaluable for hypothesis generation for precision diagnostics and therapeutics as well as drug repurposing studies.

REGULATORY RNA

PROCEEDINGS PAPER WITH POSTER PRESENTATION

CSEQ-SIMULATOR: A DATA SIMULATOR FOR CLIP-SEQ EXPERIMENTS

Wanja Kassuhn, Uwe Ohler, Philipp Drewe

Max Delbrück Center for Molecular Medicine, Berlin Institute for Medical Systems Biology, 13125 Berlin, Germany

CLIP-Seq protocols such as PAR-CLIP, HITS-CLIP or iCLIP allow a genome-wide analysis of protein-RNA interactions. For the processing of the resulting short read data, various tools are utilized. Some of these tools were specifically developed for CLIP-Seq data, whereas others were designed for the analysis of RNA-Seq data. To this date, however, it has not been assessed which of the available tools are most appropriate for the analysis of CLIP-Seq data. This is because an experimental gold standard dataset on which methods can be accessed and compared, is still not available. To address this lack of a gold-standard dataset, we here present Cseq-Simulator, a simulator for PAR-CLIP, HITS-CLIP and iCLIP-data. This simulator can be applied to generate realistic datasets that can serve as surrogates for experimental gold standard dataset. In this work, we also show how Cseq-Simulator can be used to perform a comparison of steps of typical CLIP-Seq analysis pipelines, such as the read alignment or the peak calling. These comparisons show which tools are useful in different settings and also allow identifying pitfalls in the data analysis.

A MOTIF-BASED METHOD FOR PREDICTING INTERFACIAL RESIDUES IN BOTH THE RNA AND PROTEIN COMPONENTS OF PROTEIN-RNA COMPLEXES

Usha K. Muppirala[1], Benjamin A. Lewis[2], Carla M. Mann[1], Drena L. Dobbs[1]

[1]Iowa State University; [2]Truman State University

Efforts to predict interfacial residues in protein-RNA complexes have largely focused on predicting RNA-binding residues in proteins. Computational methods for predicting protein-binding residues in RNA sequences, however, are a problem that has received relatively little attention to date. Although the value of sequence motifs for classifying and annotating protein sequences is well established, sequence motifs have not been widely applied to predicting interfacial residues in macromolecular complexes. Here, we propose a novel sequence motif-based method for "partner-specific" interfacial residue prediction. Given a specific protein-RNA pair, the goal is to simultaneously predict RNA binding residues in the protein sequence and protein-binding residues in the RNA sequence. In 5-fold cross validation experiments, our method, PS-PRIP, achieved 92% Specificity and 61% Sensitivity, with a Matthews correlation coefficient (MCC) of 0.58 in predicting RNA-binding sites in proteins. The method achieved 69% Specificity and 75% Sensitivity, but with a low MCC of 0.13 in predicting protein binding sites in RNAs. Similar performance results were obtained when PS-PRIP was tested on two independent "blind" datasets of experimentally validated protein-RNA interactions, suggesting the method should be widely applicable and valuable for identifying potential interfacial residues in protein-RNA complexes for which structural information is not available. The PS PRIP webserver and datasets are available at: http://pridb.gdcb.iastate.edu/PSPRIP/.

DETECTION OF BACTERIAL SMALL TRANSCRIPTS FROM RNA-SEQ DATA: A COMPARATIVE ASSESSMENT

Lourdes Pena-Castillo, Marc Gruell, Martin E. Mulligan, Andrew S. Lang

Memorial University of Newfoundland

Small non-coding RNAs (sRNAs) are regulatory RNA molecules that have been identified in a multitude of bacterial species and shown to control numerous cellular processes through various regulatory mechanisms. In the last decade, next generation RNA sequencing (RNA-seq) has been used for the genome-wide detection of bacterial sRNAs. Here we describe sRNA-Detect, a novel approach to identify expressed small transcripts from prokaryotic RNA-seq data. Using RNA-seq data from three bacterial species and two sequencing platforms, we performed a comparative assessment of five computational approaches for the detection of small transcripts. We demonstrate that sRNA-Detect improves upon current standalone computational approaches for identifying novel small transcripts in bacteria.

DISCOVERY OF MOLECULARLY TARGETED THERAPIES

POSTER PRESENTATIONS

MULTI MODELING APPROACH TO EVALUATING RNA SEQUENCE NORMALIZATION METHODS TO IMPROVE TRANSCRIPTOMICS BIOMARKER DISCOVERY EXPERIMENTS

Zachary Abrams, Travis Johnson, Kevin R. Coombes

Department of Biomedical Informatics, The Ohio State University

Since the advent of technologies capable of providing genetic sequence level information in transcriptomics, there has been a need to account for experimental error so that the underlying biology can be measured. This mathematical adjustment, which allows for comparability while preserving biological truth, is known as normalization. Normalization is critical in data sensitive experiments such as biomarker discovery were slight changes in data variability can propagate into large experimental error. There are many different methods for normalization depending on the type of genomic data, the platform used to collect these data, and the planned downstream analyses. Since there are multiple ways of normalizing data, there are an array of comparative studies testing different normalization methods to determine which normalization method best preserves biological reality while reducing experimental noise. Many of these experiments have been conducted on small heterogeneous data sets that were not collected for the specific purpose of systems level evaluation. In this project we propose a gold-standard set of tests to evaluate normalization procedures based on their preservation of linearity across the data, qPCR comparison, and their ability to reduce site-to-site and residual error while preserving biological variability. This is possible because we use the SEQC dataset, a large standard dataset collected from multiple institutions that contains mixture model samples that were designed for systematic evaluation. They also collected qPCR results for to generate a biological gold standard that we can use to help evaluate normalization methods. Since this data has a large number of technical replicates and that it was collected for systematic evaluation this makes it the ideal date is set to tester normalization methods. The dataset is made of four sample types A, B, C and D, where samples C and D are 3:1 and 1:3 mixtures of A and B respectively. This allows us to calculate the internal linearity of the four samples since they are related mixture models. This also allows us to test the qPCR results against the sequencing results using the same mixture linearity model. Our final evaluate of test is to perform a two-way ANOVA across all sites and samples where one variable is the biological mixture models and the other are the different Sites involved in the study. This allows us to calculate the sum-squared error attributable to site, biology and residual error and calculate the proportion of total variability attributable to each sources of error. A good normalization method should reduce site dependent and residual error variability while increasing the proportion of biological variability. This means that the researcher, after normalizing their data, is measuring more biological variability then experimental error in their data. This research helps generate a workflow for testing normalization methods for transcriptomics experiments.

GABP SELECTIVELY BINDS AND ACTIVATES THE MUTANT TERT PROMOTER ACROSS MULTIPLE CANCER TYPES

Robert J.A. Bell[1,2], Tomas Rube[3,4], Alex Kreig[5], Andrew Mancini[1], Shaun D. Fouse[1], Raman P. Nagarajan[1], Serah Choi[6], Chibo Hong[1], Daniel He[1], Melike Pekmezci[7], John K. Wiencke[8,9], Margaret R. Wrensch[8,9], Susan M. Chang[1], Kyle M. Walsh[8], Sua Myong[5], Jun S. Song[2,3,4,5], Joe F. Costello[1]

[1]Deptment of Neurological Surgery, University of California, San Francisco;
[2]Department of Biostatistics and Epidemiology, University of California, San Francisco;
[3]Department of Physics, University of Illinois, Urbana-Champaign;
[4]Institute for Genomic Biology, University of Illinois, Urbana-Champaign;
[5]Department of Bioengineering, University of Illinois, Urbana-Champaign;
[6]Department of Radiation Oncology, University of California, San Francisco;
[7]Department of Anatomic Pathology, University of California, San Francisco Medical School, San Francisco;
[8]Division of Neuroepidemiology, Department of Neurological Surgery, University of California, San Francisco;
[9]Institute for Human Genetics, University of California, San Francisco

Reactivation of telomerase reverse transcriptase (TERT) expression enables cells to overcome replicative senescence and escape apoptosis, fundamental steps in the initiation of human cancer. Multiple cancer types, including up to 83% of glioblastomas (GBM), harbor highly recurrent TERT promoter mutations of unknown function but specific to two nucleotide positions. We identify the functional consequence of these mutations in GBM to be recruitment of the multimeric GABP transcription factor specifically to the mutant promoter. Allelic recruitment of GABP is consistently observed across four cancer types, highlighting a shared mechanism underlying TERT reactivation. Tandem flanking native ETS motifs critically cooperate with these mutations to activate TERT, likely by facilitating GABP heterotetramer binding. GABP thus directly links TERT promoter mutations to aberrant expression and may provide a novel therapeutic target for multiple cancers.

OPTIMIZING GENE EXPRESSION SIGNATURES FOR DISCOVERING SIGNIFICANT DRUG HYPOTHESES IN CONNECTIVITY MAPPING STUDIES

Kelly Regan, Zachary Abrams, Lauren Lin, Tasneem Motiwala, Kevin R. Coombes, Philip R.O. Payne

Department of Biomedical Informatics, The Ohio State University

Drug repurposing, defined as finding new uses for existing therapies, is gaining popularity as a means to shorten the time to approval, limit toxicities, and reduce costs in delivering drugs from the bench to bedside. Connectivity mapping is a systematic method to generate drug-disease hypotheses based on gene expression patterns. In this study, we take a novel approach in order to determine how informative a gene expression signature is for connectivity mapping analyses. We conducted connectivity mapping analyses using the Library of Integrated Network-based Cellular Signatures (LINCS) database. We used the LINCS L1000 landmark set of genes (n=978) in order to generate distributions of connectivity scores for random genes for different gene list subsets (GLSs). We developed a novel normalization procedure for connectivity scores based on the size of the gene list and the ratio of up-to-down regulated genes. We applied our normalization procedure to pathway-related gene sets from MSigDB, three hepatocellular carcinoma (HCC) patient gene expression datasets including 91 tumor and 62 normal samples, and several established meta-analysis techniques for the HCC gene expression data. We showed that normalizing to gene list size and the proportion of up-to-down regulated genes can increase the overall utility of a gene signature for connectivity mapping. We provide preliminary evidence for a data-driven approach to define gene list subsets (GLSs) in order to maximize the number of significant connectivity mapping hypotheses. We observed that biological pathway content within gene lists has important implications in increasing how informative the gene signatures are in connectivity mapping analyses. Finally, we showed that different methods for meta-analysis, including vote counting, combined p-values and combined effect sizes, can increase the signal density and therefore how informative the resulting gene expression signatures in connectivity mapping analyses. We provide the first evidence suggesting that distributions of significant connectivity scores are dependent upon several gene list properties and that increasing the signal density of a query gene list can improve connectivity mapping hypotheses.

STRUCTURAL CHARACTERIZATION OF HIV-2 PROTEASE: ANALYSIS OF THE PROTEASE DEFORMATION INVOLVED BY THE INHIBITOR INTERACTION

Dhoha Triki[1], Mario Cano[1], Anne Badel[1], Colette Geneix[1], Delphine Flatters[1], Benoît Visseaux[2], Diane Descamps[2], Anne-Claude Camproux[1], Leslie Regad[1]

[1]MTI INSERM UMR-S 973 Université Paris Diderot - Sorbonne Paris Cité Paris France;
[2]Laboratoire de Virologie IAME UMR 1137 Inserm GH Bichat-Claude Bernard HUPNVS Paris France

Mario, Dhoha, Delphine, Colette, Anne, Benoit, Diane, Anne-Claude, Leslie HIV-2 is a retrovirus discovered a few years after HIV-1. HIV-2 infections are restricted mainly to West Africa, including Guinea-Bissau, Gambia, Senegal, and Guinea. Some European countries are also concerned with HIV-2 infection, which represents 5% of HIV infection in a series of patients in Portugal (Valadas et al., 2009) and 2% of the new HIV infections in France (Brunet S, et al. 2008). The HIV-1 and HIV-2 genomes differ by about 50 to 60% at the nucleotide level. Such differences may be correlated with differential responses to some antiretrovirals, as observed with the natural resistance of HIV-2 to some protease inhibitors (PIs) currently used as the atazanavir (Poveda E et al., 2005, Ren J, et al., 2002). Moreover, some resistance mutations are selected during treatment of the HIV-2 disease and are associated with HIV-2 drug resistance. With a much more limited therapeutic arsenal than HIV-1, it is necessary to develop new therapeutic molecules specific to HIV-2. One approach is based on the identification of new molecules inhibiting the HIV-2 protease (PR2). To improve this research it is important to understand, which features are contributing to the PI selectivity and efficiency for the HIV-1 protease (PR1) and absent in PR2 This means it is necessary to understand the antiretroviral drugs interaction mode in PR1 and PR2 and the structural deformation of the PR, implied by the inhibitor binding: the PR changes from an open to closed form. Currently there are 19 X-ray structures of the PR2 in the Protein Data Bank (PDB): one apo form and 18 holo forms complexed with different inhibitors, which 3 antiretroviral drugs. These 18 pdb structures reflect several conformations of the PR2 resulting from binding of different inhibitors. These conformations associated to the PR2 can translate its three dimensional plasticity. Thus a careful analysis and comparison of these conformations can help to identify structural variable and conserved regions of PR2. In this study, we will perform a comparison of the set of 19 X-ray structures of PR2. First we will focus on the structural analysis of PR2 structures using a tool SA-conf. This tool is in development and allows the identify the structural variable and conserved positions from a set of several conformations of a same protein. SA-conf is based on the structural alphabet HMM-SA (Camproux et al., 2004 ; Regad et al., 2008), which is a 4-residue fragments of protein classification based on their geometry. Then, we focus on the comparison of the inhibitor-binding pockets and the interaction between PR2 and inhibitors extracted from the 18 PR2 holo form. This study will allow detecting residues important for the inhibitor binding in PR2 and to understand the PR2 deformation implies by the inhibitor-binding.

PUTATIVE EFFECTORS FOR PROGNOSIS IN LUNG ADENOCARCINOMA ARE ETHNIC AND GENDER SPECIFIC

Andrew Woolston[1], Nardnisa Sintupisut[1], Tzu-Pin Lu[2], Liang-Chuan Lai[2], Mong-Hsun Tsai[2], Eric Y. Chuang[2], Chen-Hsiang Yeang[1]

[1]Academia Sinica, [2]National Taiwan University

Background: Lung adenocarcinoma possesses distinct patterns of EGFR/KRAS mutations between East Asian and Western, male and female patients. However, beyond the well-known EGFR/KRAS distinction, gender and ethnic specific molecular aberrations and their effects on prognosis remain largely unexplored. Method: Association modules capture the dependency of an effector molecular aberration and target gene expressions. We established association modules from the copy number variation (CNV), DNA methylation and mRNA expression data of a Taiwanese female cohort. The inferred modules were validated in four external datasets of East Asian and Caucasian patients by examining the coherence of the target gene expressions and their associations with prognostic outcomes. Results: Modules 1 (cis-acting effects with chromosome 7 CNV) and 3 (DNA methylations of UBIAD1 and VAV1) possessed significantly negative associations with survival times among two East Asian patient cohorts. Module 2 (cis-acting effects with chromosome 18 CNV) possessed significantly negative associations with survival times among the East Asian female subpopulation alone. By examining the genomic locations and functions of the target genes, we identified several putative effectors of the two cis-acting CNV modules: RAC1, EGFR, CDK5 and RALBP1. Furthermore, module 3 targets were enriched with genes involved in cell proliferation and division and hence were consistent with the negative associations with survival times. Conclusion: We demonstrated that association modules in lung adenocarcinoma with significant links of prognostic outcomes were ethnic and/or gender specific. This discovery has profound implications in diagnosis and treatment of lung adenocarcinoma and echoes the fundamental principles of the personalized medicine paradigm.

CLINICOPATHOLOGICAL CHARACTERISTICS OF MET PROTO-ONCOGENE IN GASTRIC CARCINOMAS

Kiwook Yang, Hyunsu Lee, Jae-ho Lee, In-jang Choi

Department of Anatomy, Keimyung University School of Medicine, Daegu, Republic of Korea

Gastric cancer is still leading cause of cancer death in world wild. Although overall survival of gastric cancer has been enhanced owing to the application of national fiber optic esophagogastroduodenos¬copy screening program in adults aged over 40 years in Korea, a large proportion of patients are still diagnosed at metastatic stage. Its development has been shown as a multi-step process, ranging from chronic gastritis to atrophy, intestinal metaplasia, dysplasia, and finally, invasive cancer. To understanding of the molecular mechanisms of this progression, lots of molecularly targeted pathways need to be discovered. A better understanding of the molecular mechanisms of this progression may provide excellent survival outcome by suggesting new potential new novel treatment strategies. However there was still not enough to make new novel treatment strategies in gastric cancer. In gastric cancer, one of the most important pathways was recently identified as MET proto-oncogene. The Met oncogene encodes for a receptor tyrosine kinase that binds to, and is activated by, the growth and motility factor so called hepatocyte growth factor. And it controls genetic programs leading to cell growth, invasion and protec¬tion from apoptosis. Although the role of MET oncogene is still to be determined in carcinogenesis of gastric cancer, overexpression and amplification of c-Met has been demonstrated in gastric cancer cell lines. In addition, earlier reports described MET gene amplification was founded approximately 10-20% of gastric cancer tissues. In the present article, we examined the frequency of amplification of MET gene in GC. And then, their clinicopathological characteristics and prognostic value were analyzed.

INNOVATIVE APPROACHES TO COMBINING GENOTYPE, PHENOTYPE, EPIGENETIC, AND EXPOSURE DATA FOR PRECISION DIAGNOSTICS

POSTER PRESENTATIONS

INTERPRETABLE UNSUPERVISED LEARNING OF THE ELECTRONIC HEALTH RECORD FOR PHENOTYPE STRATIFICATION

Brett K. Beaulieu-Jones[1], Casey S. Greene[2]

[1]Genomics and Computational Biology Graduate Group, Institute of Biomedical Informatics, University of Pennsylvania; [2]Systems Pharmacology and Translational Therapeutics, University of Pennsylvania

Extracting phenotypes from the electronic health record requires the use of either potentially biased billing codes or substantial expert knowledge provided by clinicians. Complex diseases are caused by the combination of several underlying conditions acting in parallel. We provide a realistic simulation of complex human disease, in which hidden states combine to influence observed clinical variables. We then demonstrate unsupervised learning techniques and their ability to enhance stratification classification accuracy. Denoising autoencoders provide a powerful platform to learn the structure of the underlying distribution of a disease in an unsupervised fashion. Through this process denoising autoencoders also effectively reduce the dimensionality of the data. In a series of simulations we show the efficacy of semi-supervised learning by using the attributes of a trained denoising autoencoder to achieve higher classification accuracy than the input data alone, in addition to maintaining interpretability due to the clear distribution of contribution to each constructed feature. These methods allow for the unsupervised construction of phenotypes, removing potential biases inherent in ICD billing codes while allowing for training on unsupervised data and testing on a smaller sample of clinician reviewed supervised examples.

GENOMIC DETERMINANTS OF MITF BINDING

Miroslav Hejna, Jun S. Song

Department of Physics, University of Illinois at Urbana-Champaign

Microphthalmia-associated transcription factor (MITF) is the master regulator of melanocyte differentiation and is also an oncogene amplified in 15-20% of melanomas. MITF has recently emerged as a proliferation-promoting lineage-restricted oncogene in melanomas and is implicated in conferring drug resistance. It is thus critical to understand the regulatory functions of MITF by comprehensively characterizing its genome-wide binding pattern and discovering its oncogenic transcriptional targets. MITF belongs to a family of basic helix-loop-helix leucine zipper (bHLH-Zip) transcription factors that bind a consensus hexamer CANNTG E-box motif. E-boxes are also targets of MYC, another potent oncogene that drives cellular proliferation during development as well as in many types of cancers. MITF and MYC not only share the core binding motif, but are also the two most highly expressed bHLH-Zip transcription factors in melanocytes, raising the possibility that MITF and MYC may compete for the same binding sites in select oncogenic targets. We built a novel computational predictive model that uses genetic sequence features flanking E-boxes to distinguish MITF vs. MYC binding sites with 80% accuracy genome wide. This finding demonstrates that specific combinatorial sequence features that interact with E-boxes play an important role in differentially recruiting MITF vs. MYC to target genes. We found that select MITF binding sites that can be bound by both MITF and MYC form a distinct class within the MITF binding sites in melanocytes; this MITF-MYC overlapping subclass of E-boxes is characterized by differential sequence content in the flanking region, diminished interaction with SOX10, higher conservation, and less tissue-specific chromatin organization and function. Our work suggests that the overlapping MITF-MYC subclass may be disproportionally responsible for the oncogenic activity of MITF, while the genomic loci bound by MITF but not by MYC regulate melanocyte-specific genes, and that predictive local sequence features determine these binding modes.

DEVELOPMENT OF COMPUTATIONALLY PREDICTED ADVERSE OUTCOME PATHWAY (AOP) NETWORKS THROUGH DATA MINING AND INTEGRATION OF PUBLICLY AVAILABLE IN VIVO, IN VITRO, PHENOTYPE, AND BIOLOGICAL PATHWAY DATA

Noffisat Oki[1,2], Shannon Bell[1,3], Rong-Lin Wang[2], Mark Nelms[1,2], Stephen Edwards[2]

[1]Oak Ridge Institute for Science and Education; [2]U.S. Environmental Protection Agency; [3]Integrated Laboratory Systems Contractor Supporting the NTP Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM)

The Adverse Outcome Pathway (AOP) framework is increasingly being adopted as a tool for organizing and summarizing the mechanistic information connecting molecular perturbations by environmental stressors with adverse outcomes relevant for ecological and human health outcomes. However, the conventional process for assembly of these AOPs is time and resource intensive, and has been a rate limiting step for AOP use and development. Therefore computational approaches to accelerate the process need to be developed. We previously developed a method for generating computationally predicted AOPs (cpAOPs) by association mining and integration of data from publicly available databases. In this work, a cpAOP network of ~21,000 associations was established between 105 phenotypes from TG-GATEs rat liver data from different time points (including microarray, pathological effects and clinical chemistry data), 994 REACTOME pathways, 688 High-throughput assays from ToxCast and 194 chemicals. A second network of 128,536 associations was generated by connecting 255 biological target genes from ToxCast to 4,980 diseases from CTD using either HT screening activity from ToxCast for 286 chemicals or CTD gene expression changes in response to 2,330 chemicals. Both networks were separately evaluated through manual extraction of disease-specific cpAOPs and comparison with expert curation of the relevant literature. By employing data integration strategies that involve the weighting of network edges for prioritization, the two networks can be merged into a global network of cpAOPs. Automated extraction techniques can then be used to identify individual cpAOPs that connect molecular perturbations with adverse outcomes by traversing the network for the most probable paths based on the weighted edges. This will result in a more comprehensive hypothetical AOP list than is possible by expert evaluation alone. Our workflow allows for additional datasets to be integrated into the global network regardless of the methods used to generate the individual networks, which means that the value of this resource will continue to grow as additional datasets are added. These methods highlight the value and utility of data mining and integration strategies in the development and assembly of AOPs. When prioritized, a rapid review of the resulting list of putative AOPs found can be performed by domain experts given an adverse outcome of interest.

METHODS TO ENHANCE THE REPRODUCIBILITY OF PRECISION MEDICINE

POSTER PRESENTATION

IMPLEMENTING AUTOMATED WORKFLOWS FOR CANCER IMMUNOTHERAPY MONITORING

Christopher Dubay, Valerie Conrad, Yoshinobu Koguchi

Earle A. Chiles Research Institute, Providence Cancer Center, Portland, Oregon, USA

The Immune Monitoring Laboratory (IML) in the Earle A. Chiles Research Institute at the Providence Cancer Center in Portland Oregon provides services which include the application of flow cytometry (FACS) to the monitoring of immune cell populations from blood samples and tissues. Immune cell monitoring is critical for physicians and researchers interpreting subject's responses to immunotherapies. We have developed a set of assays that use 8 color FACS to quantitate cell counts for hundreds of distinct immune cell populations. Current workflows which require same-day sample receipt and processing, and a growing number of open clinical trials using the IML, have created a requirement to maximize the automation of these workflows. This is currently being pursued at the workflow stages of: 1) sample and assay management with a Laboratory Information Management System (RURO Limsfinity LIMS), 2) sample handling and assay preparation/run automation (BD FACS Lyse/Wash, Bio-Rad Bio-Plex HTF robot, BD FACSCalibur w/ HTS loader), and 3) assay result data analysis (FlowJo, openCyto, flowDensity, etc.). For assay result data analysis automation we are moving away from a completely manual cell population gating strategy using the FlowJo software, to a supervised automated analysis leveraging FlowJo X scripting capabilities to use a Bioconductor tool set that includes openCyto modules such as flowDensity. Briefly, samples are tracked in the LIMS and associated with assay results (e.g. total blood lymphocyte counts, FCS files from FACS, etc.). When all ordered assay results for a sample are acknowledged as present in the LIMS with pointers to files stored on a networked file system, automated scripts pass appropriate files to an R server, where an openCyto analysis workflow is initiated. Results from these analysis workflows and transferred back to the LIMS and associated with the sample. The LIMS then formats the numerical results and plots into quality and results reports suitable for human interpretation, and manages report review, sign-off, distribution, archiving, and addition of the processed results to a central data repository that can be queried by clinical trial management and analytics tools. With the described workflow and support systems in place, IML personnel have time for evaluation and implementation of current and emerging new tools for analysis of multi-parameter high-dimensional flow cytometry data (e.g. SPADE, viSNE, etc.).

PRECISION MEDICINE: DATA AND DISCOVERY FOR IMPROVED HEALTH AND THERAPY

POSTER PRESENTATIONS

PATIENT-SPECIFIC DATA FUSION FOR CANCER STRATIFICATION AND PERSONALIZED TREATMENT

Vladimir Gligorijevic, Noel Malod-Dognin, Natasa Przulj

Imperial College London

According to Cancer Research UK, cancer is a leading cause of death accounting for more than one in four of all deaths in 2011. The recent advances in experimental technologies in cancer research have resulted in the accumulation of large amounts of patient-specific datasets, which provide complementary information on the same cancer type. We introduce a versatile data fusion (integration) framework that can effectively integrate somatic mutation data, molecular interactions and drug chemical data to address three key challenges in cancer research: stratification of patients into groups having different clinical outcomes, prediction of driver genes whose mutations trigger the onset and development of cancers, and repurposing of drugs treating particular cancer patient groups. Our new framework is based on graph-regularised non-negative matrix tri-factorization, a machine learning technique for co-clustering heterogeneous datasets. We apply our framework on ovarian cancer data to simultaneously cluster patients, genes and drugs by utilising all datasets. We demonstrate superior performance of our method over the state-of-the-art method, Network-based Stratification, in identifying three patient subgroups that have significant differences in survival outcomes and that are in good agreement with other clinical data. Also, we identify potential new driver genes that we obtain by analysing the gene clusters enriched in known drivers of ovarian cancer progression. We validated the top scoring genes identified as new drivers through database search and biomedical literature curation. Finally, we identify potential candidate drugs for repurposing that could be used in treatment of the identified patient subgroups by targeting their mutated gene products. We validated a large percentage of our drug-target predictions by using other databases and through literature curation.

PRECISION METRIC FOR CLINICAL GENOME SEQUENCING REVEALS MISSING COVERAGE OF KEY DISEASE GENES

Rachel L. Goldfeder, Euan A. Ashley

Stanford University

A common challenge in the application of next-generation DNA sequencing to clinical medicine is the achievement of sufficient depth of coverage to obtain accurate and complete genotype calls across the genome. This is particularly important when bodies such as the American College of Medical Genetics and Genomics (ACMG) recommend routine search of certain genes for potentially pathogenic variants. We developed a metric to quantify the extent to which current sequencing technologies reach a "clinical grade" reporting standard of every coding base including splice dinucleotides called for every gene of interest. As an exemplar, we examined the read depth at each locus in each ACMG "actionable" gene for 41 clinical genomes sequenced on Illumina HiSeq platforms, and 12 clinical exomes (Personalis ACE Exome, Agilent Clinical Research Exome, and Baylor Clinical Exome). We observe that a large number of coding and other exonic bases are not covered at a depth that would be consistent with confident calling. Exome sequencing methods adequately cover more coding bases than whole genome sequencing. Poor coverage of any base in a disease-related gene can result in false negative variant calls that are problematic for clinical decision-making. Our coverage metric provides transparency about the limitations of next-generation sequencing and will inform genotype interpretation, technology improvement, and sequencing platform choices for clinical genome sequencing.

BAYESIAN BICLUSTERING FOR PATIENT STRATIFICATION

Sahand Khakabimamaghani, Martin Ester

Simon Fraser University

The move from Empirical Medicine towards Personalized Medicine has attracted attention to Stratified Medicine (SM). Some methods are provided in the literature for patient stratification, which is the central task of SM, however, there are still significant open issues. First, it is still unclear if integrating different datatypes will help in detecting disease subtypes more accurately, and, if not, which datatype(s) are most useful for this task. Second, it is not clear how we can compare different methods of patient stratification. Third, as most of the proposed stratification methods are deterministic, there is a need for investigating the potential benefits of applying probabilistic methods. To address these issues, we introduce a novel integrative Bayesian biclustering method, called B2PS, for patient stratification and propose methods for evaluating the results. Our experimental results demonstrate the superiority of B2PS over a popular state-of-the-art method and the benefits of Bayesian approaches. Our results agree with the intuition that transcriptomic data forms a better basis for patient stratification than genomic data.

EMERGING RESOURCES FOR PHARMACOGENOMIC CLINICAL DECISION SUPPORT

Richard C. Kiefer[1], Deepak K. Sharma[1], Matthias Samwald[2], Margaret K. Linan[3], Davide Sottara[3], Kelly K. Wix[1], Robert R. Freimuth[1]

[1]Mayo Clinic; [2]Medical University of Vienna; [3]Arizona State University

The clinical implementation of pharmacogenomics (PGx) guidelines is becoming widespread, but the technical effort required to operationalize these guidelines remains significant. The PGRN Pharmacogenomics Ontology (PHONT) network resource is developing resources to reduce barriers and facilitate the scalable implementation of PGx clinical decision support (CDS) programs within existing clinical information systems. These emerging resources are being developed collaboratively with groups associated with research networks including the PGRN (e.g., CPIC, TPP, eMERGE, ClinGen), standards development organizations (e.g., HL7, W3C), and professional organizations (e.g., Institute of Medicine). We present an overview of four of these projects: - Extending the RxNorm Standard Drug Terminology - Developing Sharable CDS Rules Based on PGx Guidelines - Creating a Library for PGx CDS Artifacts - Enabling Mobile PGx CDS: The Medicine Safety Code Initiative.

AUTOMATED 3D TISSUE IMAGE SEGMENTATION AND QUANTIFICATION OF CELLULAR EXPRESSION FOR PRECISION HISTO-CYTOMETRY

Nikolay Samusik, Yury Goltsev, Garry P. Nolan

Stanford Medical School

Flow cytometry allows highly quantitative analysis of complex cell populations, but a typical instrument setup only allows analysis of single cell suspensions. Cells from solid tissues can also be analyzed, but only after tissue dissociation and cell extraction procedures that, among other problems, ablate key tissue architecture information. In contrast, optical microscopy methods provide spatial information-- however visualization and quantification of cellular subsets defined by complex phenotypic marker combinations from microscopy images is difficult and to date requires specially designed cell labeling and image analysis approaches. There is an unmet need for segmentation algorithms that identify individual cells and quantify marker expression in the complex and crowded 3D environment of the tissue. Lymphoid tissues present particular challenge because of tight packing of cells one next to another and high nucleus-to-cytoplasm ratio. We developed a 3D segmentation scheme for confocal z-stacks of tissue sections that relies on a combination of two fluorescent stains: DNA staining (Hoechst 34580) that localizes to the nuclei and actin (phalloidin-FITC) that stains cytoplasm and is absent from the nuclei. First, our algorithm identifies all cavities in the actin staining by inverting the phalloidin image, convolving it with a truncated Gaussian kernel and segmenting it around the local maxima of intensity with a custom watershed algorithm. Next, regions with low DNA signal intensity are filtered out because they mostly tend to represent spurious signals, such as gaps in the tissue section and lumena of blood vessels. Additional filtering is performed based on region size and geometry. Finally, fluorescent signals of protein markers of interest are overlaid on the region map and the intensities are quantified using summary statistics. We compared mean, median and upper quartile statistics and found that median produces the best result in terms of minimizing the artifacts of signal carryover between neighboring cells. Having identified the single-cell patterns of protein expression, we use this information to identify the major cell types and furthermore automatically define cell neighborhoods based on the local patterns of cell type frequencies. The algorithm developed here can be applied to extract single-cells expression profiles transform fluorescent images as well as the information about the cell localization. We believe this approach may be of interest wherein by converting histopathology images of healthy and disease tissue samples into information-rich representations of spatial cell type distributions, this may better enable automated diagnosis, complex disease stratification, and prediction of clinical outcomes.

PHARMGKB: TOOLS FOR ANNOTATING AND VISUALIZING GENOTYPE-BASED DRUG DOSING GUIDELINES

Michelle Whirl-Carrillo[1], Ryan Whaley[1], Maria Avarellos[1], Julia Barbarino[1], Lester Carter[1], Alison Fohner[1], Li Gong[1], Katrin Sangkuhl[1], Caroline Thorn[1], Russ B. Altman[1,2], Teri Klein[1]

[1]Department of Genetics, Stanford University; [2]Department of Bioengineering, Stanford University

The Pharmacogenomics Knowledge Base (PharmGKB) is a publically available online resource devoted to pharmacogenomic knowledge acquisition and clinical applications of this knowledge, including genotype-based drug dosing guidelines. PharmGKB contains guidelines from (1) the Clinical Pharmacogenomic Implementation Consortium (CPIC), (2) the Dutch Pharmacogenetics Working Group, and (3) several professional societies (eg. American College of Rheumatology), and plans to include genotype-based dosing recommendations from FDA-approved drug labels soon. Many dosing recommendations include phenotype groupings based on complex diplotypes of genes such as those in the cytochrome P450 family. The mapping from individual genetic variants to gene haplotypes, to diplotypes, to predicted phenotypes, to dosing recommendations can quickly become cumbersome due to the large number of potential combinations. PharmGKB has developed tools to aid curators in annotating the mapping required to move from individual variants to dosing recommendations based on guidelines published by the previously mentioned organizations. The tools include a combination of excel files and web-based interfaces that allow curators to define haplotypes, diplotypes and phenotypes appropriately for each guideline. After a guideline is annotated in PharmGKB, a web-page display is created that allows users to select genotypes/diplotypes of interest, and then presents the corresponding recommendation for that selection.

(Work supported by NIH grant R24 GM061374.)

UNDERSTANDING NON-SMALL CELL LUNG CANCER MORPHOLOGY AND PROGNOSIS BY INTEGRATING OMICS AND HISTOPATHOLOGY

Kun-Hsing Yu[1], Ce Zhang[2], Gerald J. Berry[3], Russ B. Altman[1], Christopher Ré[2], Daniel Rubin[1], Michael Snyder[4]

[1]Biomedical Informatics Program, Stanford University; [2]Department of Computer Science, Stanford University; [3]Department of Pathology, Stanford University; [4]Department of Genetics, Stanford University

Non-small cell lung cancer accounts for more than 85% of lung malignancy, and microscopic pathology is indispensable to its diagnosis. However, human evaluation of pathology slides cannot accurately predict patients' prognoses, and how omics abnormalities contribute to histopathology findings remains largely unknown. In this study, we obtained hematoxylin and eosin stained whole-slide histopathology images, pathology reports, RNA-sequencing, and proteomics data of lung adenocarcinoma (n=515) and squamous cell carcinoma patients (n=502) from The Cancer Genome Atlas (TCGA) as well as histopathology images from 294 additional patients from Stanford Tissue Microarray (TMA) Database. We extracted 9879 quantitative image features and used machine learning methods to distinguish shorter-term survivors from longer-term survivors with stage I adenocarcinoma ($P<0.003$) or squamous cell carcinoma ($P=0.023$) in the TCGA dataset. We validated the survival prediction framework with the TMA cohort ($P<0.036$ for both tumor types). In addition, we successfully predicted histology grade of adenocarcinoma with transcriptomics and proteomics signatures (area under receiver operating characteristic curve > 0.75), and identified important biological pathways associated with the most predictive features, such as apoptosis and proteolysis. These results suggest that our automatically derived image features can predict the prognosis of lung cancer patients and that the integration of histopathology and omics studies can reveal molecular mechanisms of pathology findings. Our methods are extensible to other types of malignancy.

REGULATORY RNA

POSTER PRESENTATIONS

APPLYING ARTIFICIAL NEURAL NETWORK SURVIVAL PACKAGE COX-NNET TO IDENTIFY PANCANCER PROGNOSTIC LINCRNAS

Travers Ching[1], Lana Garmire[2]

[1]Graduate Program of Molecular Biology and Bioengineering, University of Hawaii at Manoa, Honolulu, HI 96822; [2]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI 96813

Artificial neural network (ANN) is the computing architecture with massively parallel interconnections of simple neurons, and it has been applied to biomedical fields such as imaging analysis and diagnosis. However, packages that use ANN to predict disease prognosis are severely lacking. Here we have developed a Python package called Cox-nnet, which extends Cox Regression to the non-linear ANN framework. We tested cox-nnet among four sets of high-throughput gene expression data and three sets of clinical data, and found that Cox-nnet generally performs better compared to other methods, including Cox Proportional Hazards (Cox-PH), andom Forests Survival and CoxBoost. We then applied Cox-nnet to identify a panel of prognostic lincRNA pan-cancer biomarkers, among over 30,000 lincRNAs obtained from approximately 7000 TCGA RNA-Seq samples composed of 14 tumor types. In comparison to the Cox-PH model, Cox-nnet distinguished patient overall survival better in 13 of the 14 tumor types. Finally, we determined the relative importance of each lincRNA in the prognosis-biomarker panel, using the drop-one-out approach.

RNAcompete-S: COMPLEX RNA SEQUENCE/STRUCTURE MODELS DERIVED FROM A SINGLE-STEP IN VITRO SELECTION

Kate B. Cook, Shankar Vembu, Debashish Ray, Hong Zheng, Quaid D. Morris, Timothy R. Hughes

University of Toronto

RNA-binding proteins recognize RNA sequences and structures, but there is currently no systematic and accurate method to derive both types of preferences that are large, complex, and reflect direct binding. To address this absence, we introduce RNAcompete-S, which couples a single-step competitive binding reaction with an excess of random RNA 40-mers to a custom computational pipeline for interrogation of the bound RNA sequences and derivation of SSMs (Sequence and Structure Models). RNAcompete-S confirms that HuR, QKI, and SRSF1 prefer binding sites that are single stranded, and recapitulates known 8-10 bp sequence and structure preferences for Vts1, RBMY, and Drosophila SLBP. The SSM derived for SLBP is 18 bases long, and it is more accurate than previous SLBP motifs at discriminating replication-dependent histone genes. Thus, RNAcompete-S enables accurate identification of large, complex, and intrinsic specificities with a uniform assay.

COMPREHENSIVE IDENTIFICATION OF LONG NON-CODING RNAS IN PURIFIED CELL TYPES FROM THE BRAIN REVEALS FUNCTIONAL LNCRNA IN OPC FATE DETERMINATION

Xiaomin Dong[1,2], Kenian Chen[1,2], Raquel Cuevas-Diaz Duran[1,2], Yanan You[1,2], Steven A. Sloan[3], Ye Zhang[3], Shan Zong[1,2], Qilin Cao[1,2], Ben A. Barres[3], Jia Qian Wu[1,2]

[1]The Vivian L. Smith Department of Neurosurgery, The University of Texas Health Science Center at Houston (UTHealth) Medical School , Houston, Texas, USA; [2]Center for Stem Cell and Regenerative Medicine, The Brown Foundation Institute of Molecular Medicine for the Prevention of Human Diseases , Houston, Texas, USA; [3]Department of Neurobiology, Stanford University School of Medicine, Stanford, California, USA

70% to 90% of the mammalian genome is transcribed at some point during development; however, only <2% of the genome is associated with protein-coding genes. Emerging evidences suggest that long noncoding RNAs (lncRNA; >200bp) play important roles in cell fate determination. However, it is challenging to identify lncRNA comprehensively, since lncRNAs are often expressed at lower levels and are more cell type-specific than protein-coding genes. In this study, we performed ab initio transcriptome reconstruction using nine purified cell populations from mouse cortex and detected more than 5,000 lncRNAs. Predicting lncRNAs' function using cell type specific data revealed their potential functional roles in central nervous system (CNS) development. ENCODE DNase I digital footprint data and Mouse ENCODE promoters were utilized to infer transcription factor (TF) occupancy. By integrating TF binding and cell-type specific transcriptomic data, we constructed a novel framework useful for systematically identifying lncRNAs that are potentially essential in brain cell fate determination. Based on this integrative analysis, we identified lncRNAs that are regulated during Oligodendrocyte Precursor Cell (OPC) differentiation from Neural Stem Cells (NSCs) and likely to be involved in oligodendrogenesis. The top candidate lnc-OPC shows highly specific expression in OPCs and remarkable sequence conservation among placental mammals. Interestingly, lnc-OPC is significantly up-regulated in glial progenitors of experimental autoimmune encephalomyelitis (EAE) mouse models for multiple sclerosis. OLIG2-binding sites in the upstream regulatory region of lnc-OPC were identified by ChIP (chromatin immunoprecipitation)-Sequencing and validated by luciferase assays. Furthermore, loss-of-function experiments confirmed that lnc-OPC plays a functional role in OPC genesis. Overall, our results substantiated the role of lncRNA in OPC fate determination and provided an unprecedented data source for future functional investigations in CNS cell types. We present our datasets and analysis results via the interactive genome browser at our laboratory website freely accessible to the research community (http://jiaqianwulab.org/braincell/lncRNA.html) (Username: lncRNA; Password: rnaseq). This is the first lncRNA expression database of collective populations of glia, vascular cells and neurons. We anticipate that these studies will advance the knowledge of this major class of non-coding genes and their potential roles in neurological development and diseases.

SOCIAL MEDIA MINING FOR PUBLIC HEALTH MONITORING AND SURVEILLANCE

POSTER PRESENTATIONS

CITY LEVEL EXPLORATION OF DRUG-DRUG-INTERACTIONS: THE CASE OF BLUMENAU

Rion Brattig Correia[1,2], Mauro M. Mattos[3], David Wild[1], Luis M. Rocha[1,4] *

[1]School of Informatics & Computing, Indiana University, Bloomington, IN 47408; [2]CAPEs Foundation, Ministry of Education of Brazil, Brasília, DF 70040-020, Brazil; [3]Regional University of Blumenau (FURB), Blumenau, SC 89030-903, Brazil; [4]Instituto Gulbenkian de Ciência, Oeiras 2780-156, Portugal
* rocha@indiana.edu

From a public health care perspective, drug-drug-interactions (DDI) from multiple drug prescriptions is a serious problem, specially in the elderly population. Both for individuals and the system itself since patients with complications due DDI will re-enter the system at a costlier level. In spite of the issue data to measure such a problem are still scarce and hospitalizations may only account for part of the story. Thus, here we present an 18-month longitudinal study on DDI on a medium size city in southern Brazil – Blumenau, SC. Our aim is to explore the hypothesis of multiple drug prescriptions as a source of adverse drug reactions (ADR). Using city-wide drug dispensing data from the city's Health Information System (HIS) we were able to find a long list of severe interactions, most of which among the elderly population. Viewed by neighborhood we also found that interactions are directly related to the number of dispensing drugs and small but inversely related to average household income level. We also discuss how public health would benefit from city wide data in an effort to minimize DDI and enhance the quality of family health through personalized care.

SOCIAL MEDIA IMAGE ANALYSIS FOR PUBLIC HEALTH

Venkata Rama Kiran Garimella[1], Abdulrahman Alfayad[2], Ingmar Weber[3]

[1]Department of Computer Science, Aalto University, Helsinki, Finland; [2]Carnegie Mellon University in Qatar, Doha, Qatar; [3]Qatar Computing Research Institute, Doha, Qatar

Several projects have shown the feasibility to use textual social media data to track public health concerns, such as temporal influenza patterns or geographical obesity patterns. In this paper, we look at whether geo-tagged images from Instagram also provide a viable data source. Especially for "lifestyle" diseases, such as obesity, drinking or smoking, images of social gatherings could provide information that is not necessarily shared in, say, tweets. In this study, we explore whether (i) tags provided by the users and (ii) annotations obtained via automatic image tagging are indeed valuable for studying public health. We find that both user-provided and machine-generated tags provide information that can be used to infer a county's health statistics. Whereas for most statistics user-provided tags are better features, for predicting excessive drinking machine-generated tags such as "liquid" and "glass" yield better models. This hints at the potential of using machine-generated tags to study substance abuse.

GENERAL

POSTER PRESENTATIONS

NETWORK DIFFUSION PREDICTS NEW DISEASE GENES

Christie M. Buchovecky, Benjamin J. Bachman, Angela D. Wilkins, Olivier Lichtarge

Baylor College of Medicine

Most diseases have a complex genetic basis, and a fundamental limit to therapy is that often only a fraction of the genes contributing to pathogenesis is known. Since prior studies show that genes with similar phenotypes tend to be connected in protein interaction networks, we asked whether the global connectivity of protein networks could identify new genes that cause disease. Specifically, we applied Graph Information Diffusion (GID) to transfer information, in the form of labels, from known disease-causing genes to other genes. The protein interaction network was taken from the human STRING database, and the disease labels were defined by either the Online Mendelian Inheritance in Man (OMIM) database or the manually-curated DisGeNET database. Initially, to establish the connectivity of each existing OMIM disease-gene set in our network, we performed leave-one-gene-out GID analyses, achieving an average area under the Receiver Operating Characteristic curve (AU-ROC) of 0.81. To assess predictive performance, we then completed a series of time-stamped experiments. That is, we used labels and networks available in 2013 to predict genes added since then. Strikingly, we were able to predict, from the 2013 input data, the genes identified later with an average AU-ROC of 0.76. Moreover, we could also use GID analyses to identify instances in which disease-gene associations were removed, potentially due to correction of a previous error. A positive correlation (R-squared = 0.5) between leave-one-gene-out and time-stamped analyses suggests that leave-one-gene-out performance is an indicator of reliability of future predictions. With this in mind, and to ensure our results were not limited to the OMIM dataset, we applied GID to make novel predictions for nearly 700 diseases in DisGeNET. On average, leave-one-gene-out analysis of these label sets resulted in an AU-ROC of 0.83. Although this average AU-ROC significantly outperforms random chance, results varied greatly by disease. For example, we expect novel GID-based predictions to be of higher quality for Long-QT Syndrome (leave-one-gene-out AU-ROC = 0.96), than for Coronary Artery Disease (leave-one-gene-out AU-ROC = 0.71). In summary, these data show that GID, when applied to protein interaction networks and disease phenotype labels, can identify new disease genes and suggest others that are currently thought to be disease-associated, but likely are not. In the future, these techniques should be combined with GWAS and sequencing studies to help prioritize disease gene candidates and focus resources. Additionally, these same techniques could be applied to more diverse interaction networks to guide therapeutic discovery.

ENRICHMENT OF VUSS IN MOLECULAR SUBCLASSES SUGGESTS ROLE FOR CHROMATIN SIGNALING NETWORKS AND CELLULAR PROLIFERATION MECHANISMS IN GLIAL CANCER

Nicholas Camarda, Alex Fichtenholtz, Husnain Bohkari, Alyna Kahn, Eric Neumann

Foundation Medicine

Genomic alterations in tumors show both dominant driver patterns as well less conspicuous “long-tail” structures. Using an unsupervised learning method called Latent Class Analysis (LCA), we were able to obtain a classification scheme for glial tumors based on gene alterations that are either known or likely to be somatic drivers. We hypothesized that we could utilize this classification scheme to infer the role of the less frequently occurring variants of unknown significance (VUSs) by identifying which molecular subclasses are enriched for interesting VUSs, thus allowing us to determine synergistic relationships between known and likely genes and genes that have an as-of-yet unknown relationship with cancer. The ‘associative alignment’ of VUSs with known and likely variants would allow us to uncover interesting interactions and mechanisms that are part of the different oncogenic processes indicated by the various molecular subclasses. Specifically, the class driven primarily by known and likely mutations in IDH1 and frequent mutations in CIC defines the molecular subclass for glial cancers in which perturbations to chromatin signaling networks appear to play a role. We found that not only are VUSs in NOTCH1 enriched in this molecular class but also NOTCH4, SMARCA4 and ARID1A. The role of chromatin signaling networks in this molecular subclass of glial tumors has not been previously described, which makes it an open area for investigation and discussion. Additionally, the class driven primarily by CDKN2A/B co-deletions and well-characterized EGFR alterations defines the molecular subclass in which perturbations to cell proliferation pathways appear to be involved. We found that this class shows an affinity for PTEN VUSs. The majority of PTEN VUSs tended to have alterations on the phosphatase tensin-type domain, specifically between the 14th and 185th residues; many of these VUSs overlap with COSMIC-verified somatic events in this interval. These findings illuminate how some of the VUS genes may interact with known driver alterations in tumors, and may suggest new modes for therapeutically targeting glial cancers.

EXTRACTING A PROGNOSTIC MESENCHYMAL TRANSITION SIGNATURE IN GLIOMAS

Orieta Celiku, Anita Tandle, Kevin Camphausen, Uma Shankavaram

National Cancer Institute, National Institutes of Health

Gliomas are the most common malignant brain tumors. Grade IV glioma -- glioblastoma (GBM) -- arises primarily de novo and has particularly poor prognosis with less than 5% of patients surviving 5 years after diagnosis. Processes that mimic Epithelial to Mesenchymal Transition -- a common process in organ development, wound healing, and tissue remodeling -- are responsible for disorganization of the extracellular matrix, tumor cell motility, the highly invasive nature of GBM, and are thought to be responsible for resistance to radiotherapy and various therapeutic agents. By comparison, lower grade gliomas (LGGs) show fewer mesenchymal characteristics and prolonged survival. Targeting or reversing mesenchymal transition is a promising avenue in the development of molecular therapies for GBM. We use data from The Cancer Genome Atlas to study Mesenchymal Transition (MT) in gliomas and report on differences between LGGs and GBM, molecular markers associated with increased survival, potential therapeutic targets, and drug candidates for these targets. We use the average expression of 64-genes identified by Kim et al [1] in a multi-cancer setting, and shown by Cheng [2] to be associated with prolonged time to recurrence in GBM to stratify a cohort of primary untreated GBMs and grade II LGGs (astrocytoma and oligodendroma) into samples overexpressing the MT signature (High MT cohort), and those with (relatively) low expression of the signature (Low MT cohort). We use PANDA with leave-1-out Jacknifing approach following Glass et al. [3] to construct transcription factor-gene regulatory networks for the Low MT and High MT phenotypes, and extract genes that are differentially regulated between the two. The involved genes and transcription factors are analyzed for enrichment of biological pathways, and prioritized using Lasso and Elastic Net Models to extract a 25-gene signature with prognostic significance. 1. Kim H, Watkinson J, Varadan V, Anastassiou D. Multi-cancer computational analysis reveals invasion-associated variant of desmoplastic reaction involving INHBA, THBS2 and COL11A1. BMC Med Genomics. 2010;3:51. doi: 10.1186/1755-8794-3-51. PubMed PMID: 21047417; PubMed Central PMCID: PMCPMC2988703. 2. Cheng WY, Kandel JJ, Yamashiro DJ, Canoll P, Anastassiou D. A multi-cancer mesenchymal transition gene expression signature is associated with prolonged time to recurrence in glioblastoma. PLoS One. 2012;7(4):e34705. doi: 10.1371/journal.pone.0034705. PubMed PMID: 22493711; PubMed Central PMCID: PMCPMC3321034. 3. Glass K, Quackenbush J, Silverman EK, Celli B, Rennard SI, Yuan GC, et al. Sexually-dimorphic targeting of functionally-related genes in COPD. BMC Syst Biol. 2014;8(1):118. doi: 10.1186/s12918-014-0118-y. PubMed PMID: 25431000; PubMed Central PMCID: PMCPMC4269917.

COMPARATIVE STUDY OF THE SEQUENCES AND STRUCTURES AROUND O-GLYCOSYLATION SITES BETWEEN EACH SUGAR TYPE IN MAMMALIAN PROTEINS

Kenji Etchuya, Yuri Mukai

Graduate School of Science & Technology, Meiji University, Japan

Glycosylation, one of the protein post-translational modifications, is known to contribute to protein folding, functions and enzyme activities. In O-glycosylation, various sugar chains are attached to motif residues (usually Ser or Thr) by glycosyltransferases in the Golgi body. O-glycosylations were thought to occur in the Golgi body, however some of them were confirmed in nucleus, the endoplasmic reticulum and cytosol. These sugar types promote each biological function, and play different roles in living cells. However, the characteristics of protein primary sequences around the O-glycosylation sites were weak and lacked consistency. Consensus sequences except the motif residues for O-glycosylation have not been clarified completely. Many prediction tools for protein O-glycosylation have been published on the web. These methods can predict major types of O-glycosylation sites using protein primary sequences and secondary structures. Sugar type discrimination which is applied to more sugar types will be useful to clarify the correlation between each sugar type and its biological function. In our previous study, a method for the sugar type discrimination using protein primary sequences was developed and could newly predict two more sugar types in addition to the major sugar types. However, the accuracy of the prediction of GlcNAc modification was low and it was difficult to use protein primary sequences. Therefore, the propensities of amino acids, the secondary and tertiary structures around sugar binding sites were analyzed. In this study, to find the three-dimensional recognition motifs based on each sugar type, the three-dimensional coordinate data of atoms in the amino acids around the O-glycosylation sites were analyzed. Mammalian protein data with O-glycosylation was extracted from the Uniprot KB/Swiss-Prot 2015_03 and the Protein Data Bank release 2015_03. The propensities of the amino acids and secondary structures depending on each sugar type around O-glycosylation site were compared with each sugar type and applied to the sugar type discrimination method. The tertiary structure characteristics based on each sugar type around O-glycosylation sites were shown in this study.

EVOLUTIONARY ACTION INTERPRETS CODING MUTATIONS IN CANCER AND IN GENETIC DISEASES

P. Katsonis, T.K. Hsu, A. Koire, O. Lichtarge

Baylor College of Medicine

The relationship between genotype variations and phenotype changes determines health in the short term and evolution over the long term. A fundamental difficulty in determining the action of mutations on fitness is that it depends on the unique context of each mutation, which is complex and cryptic. As a result, the effect of most genome variations on molecular function and overall fitness remains unknown and this is the current bottleneck in genome interpretation. Here, we hypothesize that evolution is a continuous and differentiable physical process that couples genotype to phenotype, which leads to a formal equation for the action of coding mutations on fitness as the product of the evolutionary importance of the mutated site with the magnitude of the amino acid substitution. Approximations for these terms are computable from phylogenetic sequence analysis, and we show mutational, clinical, and population genetic evidence that this action equation predicts the effect of point mutations in vivo and in vitro in diverse proteins. This approach was ranked top amongst current state-of-the-art methods by independent judges of the Critical Assessment of Genome Interpretation community in two consecutive contests for predicting the experimental and clinical impact of protein mutants. In a Mendelian disorder (MMIHS), the EA analysis identified a strong bias to high EA ($p=10\text{-}10$) in the germline mutations of the ACTG2 gene in a cohort of only 17 patients. When the Evolutionary Action (EA) approach was applied to stratify head and neck cancer patients by the impact of their somatic TP53 mutations, the two patient groups differed significantly in overall and disease-free survival ($p<0.05$), while the EA score of the TP53 mutations was found to be predictive of platinum response. Thus, this work explicitly bridges molecular evolution with genome interpretation and population genetics. In practice, it has applications from protein function prediction to the assessment of genetic variations in both Mendelian and complex diseases.

ASSESSING THE IMPACT OF VARIANT CALLING METHODS FROM WHOLE-EXOME SEQUENCING DATA ON GENE BASED RARE-VARIANT ASSOCIATION TESTS OF CD4 RECOVERY DURING SUPPRESSIVE cART: WIHS

Kord M. Kober[1], Ruth M. Greenblatt[1], Peter Bacchetti[1], Ross Boylan[1], Kathyrn Anastos[2], Mardge Cohen[3], Mary Young[4], Deborah Gustafson[5], Bradley E. Aouizerat[6]

[1]University of California San Francisco; [2]Montefiore Medical Center; [3]Chicago CORE Center; [4]Georgetown University; [5]State University of New York Downstate Medical Center; [6]New York University

Whole-exome sequencing (WES) allows for genotyping of both common and rare variants in a population on a large scale. New methods of rare variant analysis are being developed concurrently for both variant calling and association testing. Recently our group used 50X coverage WES and gene-wise burden testing using the kernel-based adaptive cluster (KBAC) regression to identify rare-variant associations between HIV patients and CD4 cell recovery. The original variant set was generated from a now-outdated pipeline (UnifiedGenotyper from GATK v2.4.49, GATK Bundle v2.4). Understanding that variant calling from WES data is still maturing, the purpose of this study is to quantify how concordant the gene associations are from KBAC regression with input from variants generated from more recent pipelines. For comparison we generated two additional variant calls sets from same WES alignments using the most recent version of GATK (HaplotypeCaller from GATK v3.4-46 and Bundle v2.8) and another popular variant calling tool Freebayes (v0.9.21-26). The GATK v2.4, GATK v3.4, and Freebayes v0.9 pipelines called 1,930,167, 1,026,661, and 3,256,710 single-nucleotide variation (SNV) sites, respectively. Indels were excluded. 908,588 sites were found in common between the GATK v2.4 and GATK v3.4 variant sets with an overall genotype concordance of 0.989. 1,522,650 sites were found in common between the GATK v2.4 and Freebayes v0.9 variant sets. Depending on how immunologic recovery is defined, 17-75% of recipients of virologically suppressive cART (<80 copies) experience very slow or minimal recovery of circulating CD4 T cells. Worse CD4 cell recovery has been linked to: greater age, more chronic inflammation, co-infections, longer time from infection to treatment, fibrosis of lymph nodes, persistent perturbation of mucosal lymphoid tissue, AIDS-defining and non-defining cancer, and other complications. Several host genetic mechanisms have been implicated in CD4 restoration during cART. We investigated factors that influenced CD4 recovery in a group of 96 Women's Interagency HIV Study (WIHS) participants with well-characterized treatment response phenotypes, including WES. This study affirmed the strong, inverse relationship between age and CD4 cell recovery during virologically suppressive cART. KBAC regression and the GATK v2.4, GATK v3.4, and Freebayes v0.9 variant sets identified 30, 29, and 22 genes associated with CD4 cell recovery, respectively. 20 of the 29 (69.0%) genes identified from the GATK v3.4 variant sets were concordant with the GATK v2.4 results. 18 of 22 (81.8%) genes identified from the Freebayes v0.9 variant sets were concordant with the GATK v2.4 results. The variant call sets differed in the number of variants called, but were highly concordant at shared sites. KBAC regression produced a surprisingly similar set of genes associated with CD4 cell recovery. A consensus approach to gene selection may limit the variation introduced by genotypes generated from different variant calling methods.

EFFECTIVE BOOLEAN DYNAMICS ANALYSIS TO IDENTIFY FUNCTIONALLY IMPORTANT GENES IN LARGE-SCALE SIGNALING NETWORKS

Hung-Cuong Trinh; Yung-Keun Kwon

University of Ulsan

Efficiently identifying functionally important genes in order to understand the minimal requirements of normal cellular development is challenging. To this end, a variety of structural measures have been proposed and their effectiveness has been investigated in recent literature; however, few studies have shown the effectiveness of dynamics-based measures. This led us to investigate a dynamic measure to identify functionally important genes, and the effectiveness of which was verified through application on two large-scale human signaling networks. We specifically consider Boolean sensitivity-based dynamics against an update-rule perturbation (BSU) as a dynamic measure. Through investigations on two large-scale human signaling networks, we found that genes with relatively high BSU values show slower evolutionary rate and higher proportions of essential genes and drug targets than other genes. Gene-ontology analysis showed clear differences between the former and latter groups of genes. Furthermore, we compare the identification accuracies of essential genes and drug targets via BSU and five well-known structural measures. Although BSU did not always show the best performance, it effectively identified the putative set of genes, which is significantly different from the results obtained via the structural measures. Most interestingly, BSU showed the highest synergy effect in identifying the functionally important genes in conjunction with other measures. Our results imply that Boolean-sensitive dynamics can be used as a measure to effectively identify functionally important genes in signaling networks.

NOVEL APPLICATION OF BETA-BINOMIAL MODELS TO ASSESS X CHROMOSOME INACTIVATION PATTERNS IN RNA-SEQ EXPRESSION OF OVARIAN TUMORS

Nicholas B. Larson[1], Stacey Winham[1], Zach Fogarty[1], Melissa Larson[1], Brooke Fridley[2], Ellen L. Goode[1]

[1]Mayo Clinic; [2]University of Kansas Medical Center

In females, X-chromosome inactivation (XCI) epigenetically silences transcription of one of the two homologous X chromosomes to achieve similar expression levels to males. Some genes are known to escape XCI under normal conditions, and aberrant XCI patterns may occur in female-specific cancers, such as ovarian cancer (OVCA). Which homolog is silenced in a given cell is randomly selected in early development and transmitted mitotically, and tissues that are skewed toward a specific homolog can inform the XCI status of individual genes. We conducted a two-stage analysis to estimate XCI in 453 X genes in 114 OVCA tumor samples using allele-specific expression read counts derived from genome-wide SNP and RNA-Seq expression data. We first applied a composite likelihood-ratio test using a single parameter beta-binomial model, identifying 89 skewed XCI samples to use for gene-level XCI evaluation. We then assessed genic XCI via a two-component beta-binomial mixture model fit using a Bayesian MCMC approach, accommodating extra-binomial variation and sample-specific skewness. Posterior samples of XCI mixture component variables were used to estimate the posterior probability of XCI escape per gene. Overall, our estimates of genic escapee patterns conformed well to previous LCL studies. However, a mean 5.4% of genes per sample thought to escape silencing showed evidence of XCI, while 8.5% indicated the opposite pattern. Moreover, 22% of genes demonstrated heterogeneity of escape status across samples. These results may indicate inter-tissue XCI differences or cancer-related aberrant XCI, and further research on paired tumor-normal tissues is necessary to evaluate somatic XCI alteration in cancer.

DISCOVER EQTL WITH FLEXIBLE LD STRUCTURE AND TREE-GUIDED GROUP LASSO

Li Liu[1], Sudhir Kumar[2], Gregory Gibson[3], Biao Zeng[3]

[1]Arizona State University; [2]Temple University; [3]Georgia Institute of Technology

Expression quantitative trait loci (eQTL) are genomic loci that contribute to variation in expression levels of mRNAs. Traditional genetic mapping of eQTL relies on univariate analysis that naively assumes independence of genomic loci and overlooks their inter-relationships. Recently, sparse group lasso has been applied to eQTL analysis, in which linked SNPs are modeled as a group as defined by ad hoc linkage disequilibrium (LD) cut-off. This is problematic because an optimal LD cut-off is not known a priori, and its choice significantly affects the between-group sparsity (the number of blocks selected) and the within-group sparsity (the number of SNPs selected in each block). Furthermore, the same LD cut-off is not appropriate for linked SNPs in all regions of the genome. To address these problems, we developed a new method that considers flexible LD structure as a hierarchical tree and uses it to guide sparse linear regression. Specifically, we computed correlation coefficient (r2) between SNP frequencies which served as proxies for LD. Based on r2, SNPs were organized in a hierarchical structure such that SNPs in high and low degrees of linkages are clustered at different levels of the hierarchy (from leaves of the tree towards the interior nodes). Then, tree-guided group lasso was performed to find the optimal solution of a sparse linear regression. We evaluated the performance of this new approach using computational simulations. In these simulations, an artificial population was generated that shared a similar LD structure with the CEU population in the 1000 Genome Project. Continuous response variables were simulated to represent gene expression level. Multiple SNPs with various levels of linkages and effect sizes were added into these artificial populations as positive controls. With these simulated datasets, our new approach achieved an average precision of 77% (fraction of selected SNPs that are true eQTL SNPs) and recall of 75% (fraction of true eQTL SNPs that are selected). The missed eQTL SNPs contributed to only 10% of the residual variance, suggesting that the most significant eQTL SNPs have been successfully recovered. We then applied this new method to identify eQTL SNPs for inflammatory bowel diseases (IBD). Our new method successfully identified blocks of SNPs that were significantly correlated with the gene expression level. The linkages (r2) within each block varied from 0.50 to 0.95, demonstrating the ability to capture flexible LD structure. To detect both univariate and multivariate effects, we further narrowed the lead SNPs by performing simple linear regression within a block and multiple linear regression across blocks. The final list of SNPs consists of many known eQTL loci for IBD.

A MOTIF-BASED METHOD FOR PREDICTING INTERFACIAL RESIDUES IN BOTH THE RNA AND PROTEIN COMPONENTS OF PROTEIN-RNA COMPLEXES

Carla M. Mann[1], Usha K. Muppirala[2], Benjamin A. Lewis[3], Drena L. Dobbs[1]

[1]Iowa State University, Department of Genetics, Development, and Cell Biology; [2]Iowa State University, Genome Informatics Facility; [3]Truman State University

Interactions between RNA and proteins are now known to be critical in a wide variety of biological processes, ranging from well-studied roles in translation and RNA processing, to more recently discovered roles in cellular differentiation, epigenetic regulation, and protein localization. Disruptions in these interactions can lead to a host of disorders, including various cancers and neurodegenerative disorders such as Alzheimer's disease. Experiments aimed at determining the partners and interfaces in ribonucleoprotein particles (RNPs) can be time-consuming and expensive; thus computational methods for predicting whether a given RNA and protein will interact (partner prediction), and identifying the amino acids and ribonucleotides involved in the interface (interface prediction) are valuable. Most existing RPN interface prediction methods have two weaknesses: (i) they predict only those amino acids in a protein that are likely to bind RNA in general, and cannot predict an interface for a specific RNA partner; (ii) with two exceptions, they cannot predict the protein-binding nucleotides in the RNA. Here we present PS-PRIP (Partner-Specific Protein-RNA Interface Prediction), a new motif-based method for predicting the interfacial residues in RNP complexes in both the protein and RNA components of an interface, in a partner-specific manner. The reliability of PS-PRIP was significantly greater for predicting RNA-binding amino acids in proteins (specificity 0.92, sensitivity 0.61, MCC of 0.58) than for predicting protein-binding ribonucleotides in RNA (specificity 0.69, sensitivity 0.75, MCC of 0.13), when evaluated in 5-fold cross-validation experiments on a non-redundant dataset of 1,130 RNP complexes. Performance was comparable on an independent test set of 327 RNP complexes, with better performance in predicting RNA-binding residue (specificity 0.92, sensitivity 0.64,MCC of 0.59) than protein-binding ribonucleotides (specificity 0.67, sensitivity 0.79, MCC of 0.13). These results suggest that PS-PRIP can be a valuable tool for experimentalists who wish to target interfaces in specific protein-RNA complexes or to perturb specific interactions in protein-RNA interaction networks. Current work is focused on improving interface predictions for the RNA component of RNPs. PS-PRIP is freely available at http://pridb.gdcb.iastate.edu/PSPRIP/index.html

PREDICTING DRUG RESPONSE IN HUMAN PROSTATE CANCER FROM PRECLINICAL ANALYSIS OF IN VIVO MOUSE MODELS

Antonina Mitrofanova[1,2], Alvaro Aytes[2], Min Zou[2], Michael M. Shen[2], Cory Abate-Shen[2], Andrea Califano[2]

[1]Rutgers University, Health Informatics Department; [2]Columbia University Medical Center

Recent large-scale genomic analyses of cancer have led to the identification of "actionable" driver genes that represent therapeutically accessible targets, including oncogene and non-oncogene dependencies. However, the accurate and efficient identification of drugs and drug combinations that target such drivers represents a major challenge, particularly for transcriptional regulators, which are generally considered pharmacologically inaccessible for therapeutic targeting. Here we introduce an approach that uses in vivo drug perturbation data from Genetically Engineered Mouse models of aggressive prostate cancer to predict drug efficacy in human patients. Transcriptional regulatory network-based analysis of expression profiles from these pharmacological perturbations identified drugs and drug combinations that inhibit the transcriptional activity of FOXM1 and CENPF, which have been identified as key regulatory genes that drive prostate cancer malignancy. Validation of our computational predictions in mouse and human prostate cancer experimental essays confirmed the specificity and synergy of predicted drugs to abrogate activity of FOXM1 and CENPF and inhibit tumorigenicity. Furthermore, computational analysis of transcriptional regulatory alterations after the drug administration identified treatment-responsive genes, which are potential biomarkers of patient response to the therapy. We propose that this approach may allow systematic identification of drugs targeting specific tumor dependencies, and thus providing therapeutic benefits for patients with transcriptional dysregulations.

BOOSTING MCMC SAMPLING FOR MULTIPLE SEQUENCE ALIGNMENT WITH DIVIDE AND CONQUER

Michael Nute[1], Nam Nguyen[2], Tandy Warnow[2,3]

[1]University of Illinois at Urbana-Champaign, Department of Statistics; [2]University of Illinois at Urbana-Champaign, Department of Computer Science; [3]University of Illinois at Urbana-Champaign, Department of Bioenginnering

Multiple sequence alignment is the first step in the pipeline of many bioinformatic analyses, including phylogeny estimation and metagenomics. However, insertions and deletions (joinly, "indels") cause difficulty in that their implications depend on the interpretation of the data generation process. For indels that occur as part of the evolutionary process, a phylogeny aware alignment method can increase accuracy of multiple sequence alignments, particularly with respect to reducing false positives. We show that boosting an MCMC method (BAli-Phy) for phylogeny aware reconstruction with a previously published divide-and-conquer strategy (PASTA) gives superior alignments for alignments of 500 taxa, and can be potentially applied in parallel to many more taxa than that.

AUTOMATING BIOMEDICAL DATA SCIENCE THROUGH TREE-BASED PIPELINE OPTIMIZATION

Randal S. Olson[1], Ryan J. Urbanowicz[1], Peter C. Andrews[1], Nicole A. Lavender[2], La Creis Kidd[2], Jason H. Moore[1]

[1]University of Pennsylvania Institute for Biomedical Informatics; [2]University of Louisville

Over the past decade, data science and machine learning has grown from a mysterious art form to a staple tool across a variety of fields in academia, business, and government. In this poster, we introduce the concept of tree-based pipeline optimization for automating one of the most tedious parts of machine learning -- pipeline design. We implement a Tree-based Pipeline Optimization Tool (TPOT) and demonstrate its effectiveness on a series of simulated and real-world genetic data sets. In particular, we show that TPOT can build machine learning pipelines that achieve competitive classification accuracy and discover novel pipeline operators -- such as synthetic feature constructors -- that significantly improve classification accuracy on these data sets. We also highlight the current challenges to pipeline optimization, such as the tendency to produce pipelines that overfit the data, and suggest future research paths to overcome these challenges. In addition to these findings, we release TPOT as an open source Python package for all practitioners to use. As such, this work represents an early step toward fully automating machine learning pipeline design.

DETECTION OF BACTERIAL SMALL TRANSCRIPTS FROM RNA-SEQ DATA: A COMPARATIVE ASSESSMENT

Lourdes Pena-Castillo, Marc Gruell, Martin E. Mulligan, Andrew S. Lang

Memorial University of Newfoundland

Small non-coding RNAs (sRNAs) are regulatory RNA molecules that have been identified in a multitude of bacterial species and shown to control numerous cellular processes through various regulatory mechanisms. In the last decade, next generation RNA sequencing (RNA-seq) has been used for the genome-wide detection of bacterial sRNAs. Here we describe sRNA-Detect, a novel approach to identify expressed small transcripts from prokaryotic RNA-seq data. Using RNA-seq data from three bacterial species and two sequencing platforms, we performed a comparative assessment of five computational approaches for the detection of small transcripts. We demonstrate that sRNA-Detect improves upon current standalone computational approaches for identifying novel small transcripts in bacteria.

THE SPREAD OF THE 2014 EBOLA ZAIRE VIRUS IN WEST AFRICA

Matthew Scotch[1], Rachel Beard[1], Robert Pahle[1], Anuj Mubayi[1], Sirish Namilae[2], Ashok Srinivasan[3]

[1]Arizona State University; [2]Embry-Riddle Aeronautical University; [3]Florida State University

Air travel has been identified as a leading factor in the spread of several infectious diseases. This has motivated calls for limitations on air travel during epidemics, such as during the 2014 Ebola outbreak in West Africa. However, such limitations carry considerable economic and human costs. We are developing a computational framework for public health decision support during infectious disease outbreaks that will enable individuals to make informed real-time decisions that mitigate the risk of spread without eliminating air travel. This multi-component platform models elements of human movement with global diffusion and molecular evolution of the virus. Our human movement model will estimate the number of infected individuals at each airport. We will then use these data in our phylogeography models of global spread to empirically evaluate the impact of air travel policy options such as changes to enplaning, in-flight movement, and deplaning. Here we report the initial results of our phylogeography models. We collected full genome 2014-15 West African Ebola Zaire virus sequences from the Virus Pathogen Resource database (ViPR) [1]. We recorded temporal and geographical metadata of each sequence including the decimal day of year and the location of the virus. We aligned 810 sequences using Muscle v3.8.31 [2] and selected an HKY + Γ [3] DNA substitution model based on prior phylogeography studies of the 2014 strains [4, 5]. We used BEAUti v1.8.2 [6] to specify different assumptions including codon partitions, strict or relaxed molecular clocks, and different tree priors for the Bayesian inference. For each of these scenarios, we created a separate data partition that considered each location as a discrete character trait and specified a non-reversible transmission rate matrix under a Bayesian Stochastic Search Variable Selection (BSSVS) as defined by Lemey et al [7]. We used BEAST v.1.8.2 [6] to initiate Markov chain Monte Carlo simulations and used Tracer v1.6. [8] to visualize the effective sample sizes of model parameters and produced log marginal likelihood estimates to identify the model with the best fit for the data. In this presentation, we will report the preliminary phylogeographic results including the Kullback-Leibler divergence test [9] for the most likely origin, the Bayes factors of key pairwise migration routes, and the association index for the influence of geography on the genomic evolution of the virus. In addition to our phylogeographic results, we will discuss our approach to integrate human movement models with these models of global diffusion.

GENE EXPRESSION PREDICTION OF INFECTIOUS DISEASE RESILIENCE

Solveig K. Sieberts[1], Ricardo Henao[2], Ephraim Tsalik[3], Lara M. Mangravite[1]

[1]Sage Bionetworks, Seattle, WA; [2]Department of Electrical and Computer Engineering, Duke University, Durham, NC; [3]Department of Medicine, Duke University, Durham, NC

Studying disease resilience can be a powerful approach to identify protective biological mechanisms, which may ultimately aid in the development of prevention and treatment therapies. In the area of infectious disease, infection rates are variable following exposure to respiratory viruses. Transcriptional profiles derived from blood samples collected 40 hours or longer post-exposure are able to detect signatures of active infection that distinguish infected from uninfected individuals but it is unknown whether infectious status can be predicted prior to symptom onset. Data: The data was collected from 134 individuals in seven viral challenge trials across four different viruses: Influenza H1N1, Influenza H3N2, RSV, and Rhinovirus. Whole blood gene expression profiling was available for 3-10 time points ranging between 30 hours pre-exposure and 36 hours post-exposure. Resilience was defined in two ways: (1) patients who do not become symptomatic post-exposure and (2) patients who do not exhibit viral shedding post-exposure. Results: We generated predictive models using sparse logistic regression based on pre-exposure and early post-exposure time points. These models demonstrated increasing ability to predict resilience in held out data as time progressed from t=0 (just prior to viral exposure) to later post-exposure time points. On average these models are an improvement over those focused on prediction using solely demographic predictors (in this case age and gender). Naïve approaches incorporating data from multiple time points to build predictors do not result in an improvement over models built using only data from the single most recent time point. We explore the potential for more sophisticated modeling of these data.

DETECTION AND EXPLORATION OF HCMV IN HEALTHY BLOOD DONORS

Vivian Young, Carolyn Tu, Sneha Krishna, Patricia Francis-Lyon, Juliet Spencer

University of San Francisco

Human cytomegalovirus (HCMV), which is widespread in the general population, has recently been linked to breast cancer and neuroblastoma. As a known threat to newborn and immune-compromised individuals, a test for HCMV serostatus is currently administered by blood banks. The objectives of this research are 1. to improve detection of the HCMV virus in healthy blood donors and 2. to explore the relation of HCMV biomarkers with a variety of immunological factors that are detectable in human blood. HCMV establishes lifelong latency with expression of a limited subset of viral genes. During lytic infection cmvIL-10, a potent immunosuppressive homolog of human IL-10 (hIL-10) is produced. During both latent and lytic infection alternative splicing of the same gene yields a truncated protein, LAcmvIL-10, having a limited range of immune suppressive functions. The two viral cytokines, are collectively termed viral IL-10 (vIL-10). We address our first objective with the development of a sensitive ELISA to detect vIL-10 in human blood samples. The ELISA has been found to be specific for vIL-10, and no cross-reactivity with hIL-10 or ebvIL-10 is observed. In preliminary studies vIL-10 was detected among seronegative donors. These donors lacked any measurable IgG or IgM response to HCMV, but viral DNA was detected by PCR for exon 4 of IE1 in genomic DNA isolated from the whole blood. These findings demonstrate that vIL-10 can be found at measurable levels in healthy adults regardless of HCMV serostatus, suggesting that vIL-10 testing may offer improved screening for HCMV in the blood supply. For our second objective we explore the relation of the two isoforms of vIL-10 with a variety of biomarkers in human blood such as TNF-alpha, IL-10 and other cytokines. We utilize approaches in both R and Python (sci-kit-learn) to analyze data from 2 data sets: 61 male/female subjects and 150 female subjects. We produce visualizations to allow domain experts to obtain insight into the relations of HCMV with other immune factors. This study is part of a project to develop models for risk assessment and early detection of breast cancer utilizing blood/serum assays and patient medical information.

COMPARATIVE ANALYSIS OF GPI MODIFICATION MECHANISMS BETWEEN HUMAN AND PLANT PROTEINS FOCUSING ON SIGNAL-PEPTIDES

Hiromu Sugita, Naoyuki Takachio, Noritaka Kato, Hanae Kaku, Yuri Mukai

Meiji University

Glycosylphosphatidylinositol (GPI) is a kind of glycolipid which can anchor proteins to the cell membrane. GPI-anchored proteins (GPI-AP) are known to localize themselves to the microdomain on the cell membrane, called raft, via the Endoplasmic Reticulum (ER). GPI-APs have two signal sequences, signal-peptides (SP) and GPI-attachment signals (GPI-AS). These are N-terminal ER localization sequences and C-terminal signal sequences for GPI modification, respectively. Human GPI-APs are closely related to incurable human disorders including cancer, Parkinson's disease and bovine spongiform encephalopathy. Plant proteins which have similar GPI modification systems as mammal GPI-APs, can translocate to the cell wall or the raft region on the cell membrane. However, because few plant proteins have been isolated as GPI-APs, the GPI modification mechanism has not been clarified. CEBiP (Chitin Elicitor Binging Protein), one example of cell membrane localized plant GPI-AP, is known to bind chitin which is a major ingredient of fungus and then activate a defense reaction called elicitor. In this study, GPI modification mechanisms were compared between human and plant proteins using bioinformatics and experimental methods. The SP and GPI-AS sequences of both human and plant proteins were extracted from the Uniprot Knowledgebase/Swiss-prot release 2015_03. The hydrophobicity and position-specific scoring matrix (PSSM) were calculated using the extracted sequence dataset of SPs and GPI-ASs. Some differences were found in the physicochemical characteristics, especially in SPs. The hydrophobicity analysis showed that the average hydrophobicity of plant SPs were higher, and the hydrophobic regions were longer than those of mammal SPs. The PSSM also indicated that the propensities of the amino acids with high hydrophobicities were higher in plant SPs than in human SPs. The differences of the position-specific amino acid propensities between human and plant proteins regarding the GPI modification mechanisms were discussed based on the bioinformatics analysis. Furthermore, wild-type CEBiP protein was expressed in the HeLa cells which are typical mammalian cells, and the subcellular localization of CEBiP was observed by the immunostaining method. As a result, wild-type CEBiP was localized to the cytoplasm and nucleus. Chimera CEBiP, of which the SP region had been replaced with a human Prion SP sequence, was also expressed in the HeLa cells and localized to the cell membrane. These results indicated that one of the differences of the GPI modification mechanisms between human and plant proteins is the protein transport system which depends on SPs.

WISHBONE IDENTIFIES BIFURCATING DEVELOPMENTAL TRAJECTORIES IN SINGLE CELL DATA

Manu Setty[1], Michelle Tadmor[1], Shlomit Reich-Zeliger[2], Nir Friedman[2], Dana Pe'er[1]

[1]Columbia University; [2]Tel Aviv University

Wishbone is an algorithm to order high dimensional single cell data of a system where a cell's developmental trajectory bifurcates to one of two fates. Wishbone first builds a similarity graph between cells and derives an initial ordering relative to an input early cell. Then, a triangulation approach is employed to map cells to branches. The ordering and branch mapping is refined by a converging iterative process. Wishbone recovers the chronological ordering of cells as well as a branch association score for each cell to one of two developmental branches. With the wishbone output we trace the expression of lineage markers along the branching trajectory and characterize the decision making process for lineage commitment in single cells. T cell development in the mouse thymus is an ideal system to study bifurcating developmental trajectories. In this system, CD4+ helper T cells and CD8+ cytotoxic T cells develop from lymphoid progenitors seeded from the bone marrow. We applied Wishbone to 42 channel mass cytometry data and accurately recovered the various known stages in T cell development including the bifurcation point. Further investigation revealed a preference for IL7 signaling activity along with Helios and Notch3 mediated signaling specifically in the CD8 lineage prior to negative selection. Thus Wishbone also allows characterization of marker dynamics at specific stages of development at novel granularity. Wishbone is robust to the free parameters used and significantly outperforms existing methods such as Monocle and Scuba in identification of both ordering and branching of cells.

INTERACTION BETWEEN GPI-ANCHORED PROTEINS AND GPI TRANSMAMIDASE

Daiki Takahashi, Hiromu Sugita, Tsubasa Ogawa, Kota Hamada, Kenji Etchuya, Yuri Mukai

Meiji University

Glycosyl-phosphatidyl-inositol (GPI) is composed of sugars and lipids and is known as one of the major post-translational modification molecules. Many proteins anchored to the plasma membrane by GPI exist on the eukaryotic cell surface. GPI-anchored proteins play essential roles in living cells including immunity, signaling regulation and cell adhesion. The malfunction of GPI modification is well known to have correlations to serious disorders. Thus, the understanding of the GPI modification mechanism is believed to be crucial for the medical treatment of those disorders. GPI-anchored proteins have a single signal-peptide at the N-terminus essential for protein localization to the endoplasmic reticulum (ER) and also have a GPI-attachment signal at the C-terminus for GPI modification. After the signal-peptide is inserted into the ER, the signal-peptide is separated by the signal-peptidase. GPI transamidase enzymes recognize the GPI-attachment signal sequences and then modify GPI to the protein's ω-site in the ER. GPI transamidase is known as a protein complex consisting of PIG-K, GPAA1, PIG-S, PIG-T and PIG-U. PIG-K is the catalytic subunit which has cysteine and histidine in the active site. PIG-U and PIG-T are also known to be related to GPI recognition and the scaffold protein for these subunits. Though these subunits are indispensable to enzyme reactions, their functions have not been clarified in detail. The GPI-attachment signals have no consensus in their primary sequences, therefore the molecular mechanisms of the GPI-attachment signal recognition and separation by the GPI transamidase are also unsolved. Through the sequence and structural analysis of GPI anchored proteins in this study, many alanine, lysine and tyrosine residues were confirmed in the space that surrounds the ω-sites. The domain which has an interaction with GPI-anchored proteins was specified by the alignment analysis of GPI8 and GPAA1 which are PIG-K orthologs. In addition, the interactions between the substrates and enzymes based on physicochemical factors were considered by the positions of the preservation domains on their tertiary structures. Moreover, through the secondary structural observation of GPI-attachment signals by the circular dichroism measurement, GPI recognition mechanisms are discussed in this study.

PIPELINES FOR NEXT GENERATION DATA ANALYSIS AND GENOME ANNOTATION

Victor Solovyev, Igor Seledtsov, Vladimir Molodsov, Oleg Fokin

Softberry Inc.

Advances in technologies generate unprecedented stream of biological data. The complexity and amount of these data require developing multi-steps automatic pipelines that allow users to rapidly interrogate vast data sets to make biological and medical discoveries. We present here bioinformatics tools for efficient analysis of large-scale NGS data. OligoZip - de novo NGS reads assembler pipeline. The assembler provides effective solutions to the following tasks: 1) de novo reconstruction of genomic sequence; 2) reconstruction of sequence using a reference genome from the same or close organism. The pipeline starts with Adapter_trim and Quality_trim modules that remove adapter sequences from PE and MP reads and low quality read ends. Next step includes ReadsClean module that can correct 99% of errors in Illumina generated reads. The assembling algorithm greatly benefits from using the cleaned reads. It also incorporates the following modules: Contig_extender, iterative Contig joining (scaffolds building) module and the final "holes" patching module. Using PE and MP reads it can assemble bacterial genomes in one or a few contigs. OligoZip pipeline can build eukaryotic chromosome sequences in just a dozen contigs using sets of PE and MP simulated reads with less then 1% of errors. ReadsMap pipeline is designed for NGS reads mapping to the genome and splice sites and SNP identification. The pipeline has been significantly improved in speed of data processing and its accuracy to align spliced and un-spliced RNASeq reads to the reference genome reached Sensitivity 0.99 and Specificity 0.96. We incorporate this pipeline into our Fgenesh++ gene prediction pipeline that can use RNASeq supported splicing sites to improve the accuracy of ab initio gene prediction. TransSeq pipeline is developed for de novo assembling alternative transcripts from short reads and gene expression quantification. It assembles the RNA-Seq data into unique sequences of transcripts and generates full-length transcripts for a set of alternatively spliced isoforms. The program demonstrates Sensitivity 0.97 and Specificity 0.96 in identifying known RNA transcripts on C. elegans test data. Another GenomeMatch pipeline serves to compare long genome sequences. Fgenesb - automatic annotation pipeline includes self-training gene-finding parameters, prediction of CDS and identification of closest protein homolog from COG, KEGG and NR databases, mapping rRNA and tRNA genes, prediction of operons, promoters and terminators. All these pipelines can use multiprocessor architecture of modern computers that significantly speed up their running time. To scale up the data analysis several pipelines were implemented to be run with multiple virtual machines in Amazon cloud Web Services. Many modules of these pipelines can be run at the Softberry public Web server (www.softberry.com) as well as using MolQuest – an easy-to-use desktop application for sequence analysis and molecular biology data management that can be downloaded at www.molquest.com.

EXPERIMENTAL CHARACTERIZATION OF COMPUTATIONALLY PREDICTED "METAMORPHIC" PROTEINS

James O. Wrabl[1], Jordan Hoffmann[2], Mark Sowers[1], Vincent J. Hilser[1]

[1]Department of Biology and T. C. Jenkins Department of Biophysics, Johns Hopkins University, Baltimore, MD 21218; [2]Paulson School of Engineering and Applied Sciences, Harvard University, Boston, MA 02115

The emerging biological phenomenon of "metamorphic" proteins, single amino acid sequences that adopt two physiologically distinct structures and functions, has been associated with key cellular functions such as transcription/translation, circadian rhythm, mitosis, and chemotaxis. Unfortunately, the frequency of occurrence of metamorphic proteins in a given proteome is completely unknown, because annotation of such proteins challenges current prediction methods largely reliant on sequence similarity. Thus, the very existence of metamorphic proteins may potentially confuse development of molecularly targeted therapies or hinder interpretation of results from precision medicine diagnostic tools. To address this problem, we develop a novel metric for sequence-structure compatibility, using energetic information derived from an experimentally validated ensemble-based description of protein thermodynamics. This approach naturally permits a single amino acid sequence to be compatible with more than one fold, if the energetics of adopting each fold are favorable. The compatibility metric takes both native and denatured state energetic information into account, with separate calibration of Gaussian probability models for background sequence-structure scores in each state. High-identity sequences, previously demonstrated (Alexander, et al. PNAS 106:21149) to adopt either Streptococcus protein GA or GB folds, were correctly recapitulated, suggesting that the simple compatibility metric indeed reflected the energetic determinants of fold. To test this hypothesis, ten arbitrarily chosen uncharacterized members of the high-identity sequence space were expressed and purified; nine were found to be consistent with their predicted folds as assessed by circular dichroism and fluorescence spectroscopy. Because our ensemble-based framework is general and can be applied to any desired fold, it may be useful for large-scale proteomic detection, or future targeted design, of novel metamorphic proteins.

NAME ENTITY RECOGNITION FOR DRUG METABOLITE BY USING TEXT MINING METHOD

Heng-Yi Wu[1], Deshun Lu[1], Mustafa Hyder[2], Lang Li[3]

[1]Center for Computational Biology and Bioinformatics, School of Medicine, Indiana University; [2]Division of Clinical Pharmacology, School of Medicine, Indiana University; [3]Department of Medical and Molecular Genetics, School of Medicine, Indiana University

Background: According to the recent draft DDI guidance from FDA and European Medicines Agency (EMA) Guideline, the characteristics of drug metabolite were recognized as an important feature to the investigation of drug-drug interaction studies, adverse effects of chemical compounds or the extraction of pathway and metabolic reaction relations. However, there is no NER system focusing on the identification for drug metabolite and reaction. Unlike protein, gene, or drug NERs, drug metabolite annotation confronts different obstacles, including the limited resource of terminology and the issue of multi-token entities. In this work, we propose a system to annotate drug metabolites and reactions in scientific text, utilizing an integrated dictionary and machine learning algorithms Materials: The proposed method requires two resources: (a) metabolite-rich corpora (210 MEDLINE abstracts) and, (b) a comprehensive lexical repository, including a drug name dictionary (DrugBank and MeSH Term) and the lexicon of general nomenclatures (prefix/suffix) for drug metabolite (Collected from books). Method: To annotate drug metabolites, the system can be divided into three parts: (a) The first part, dictionary-based tagging, finds the entities using drug name and the metabolite's pre/suffix lexicon with partial-match method. Both drug names and their abbreviations can be also detected in this phase; (b) The second part is to create a window based on the annotated drug name. The window sequentially labels the entities surrounding the recognized drug name, which means the entities within the window can be the candidate entities for drug metabolite. (c) Third part is the supervised method with the results of dictionary-based NER and POS tagging. This module takes advantage of the information of POS, drug index and pre/suffix feature to determine if the entities within window belong to the drug metabolite or not. Finally, we utilize the information from both (b) and (c) to annotate the drug metabolite. The whole picture of annotation procedure is shown in Figure. Result: To evaluate performance, a golden-standard corpus is created by three annotators. SVM outperformance J48 and LMT with the precision (0.8897), recall (0.7820), F-score (0.8324). In this work, we compare our performance with an existing NER system, whatizitChemical, which is designed for recognizing small molecules or chemical entities. Our NER system with SVM algorithm outperforms whatizitChemical by 13% in recall. Conclusion and discussion: The contribution of this work is the construction of gold-standard drug metabolite corpus. Another is the proposed NER system for annotating drug metabolites and reactions, which can achieve the competitive F-Score and outperform the existing Whatizit pipeline.

CHARACTERIZATION, CLASSIFICATION AND EVOLUTIONARY ANALYSIS OF LNCRNAS COMBINING MACRO- AND MICRO-HOMOLOGY

Tuanlin Xiong, Ge Han, Qiangfeng Zhang

MOE Key Laboratory of Bioinformatics, Center for Synthetic and Systems Biology, Center for Tsinghua-Peking Joint Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing 100084, China

More and more studies have found that long non-coding RNAs (lncRNAs) are involved in many biological processes including epigenetic regulations. In particular, in recent years, some research found that genetic variations and abnormal expression of lncRNAs are correlated with human diseases such as cancer. The characterization, classification and evolutionary analysis of lncRNAs are vital to understand and trace the birth and death and also the dynamic changes of these non-coding genes and their functions. However, due to their low levels of sequence conservation and the poor annotations in many species, it is very difficult to comparatively study lncRNAs in different species. Our previous work based on analysis of genomic co-linearity and sequence motif distribution pattern successfully identified Drosophila roX lncRNA orthologs in more than 30 species. We further developed a framework that combines the macro- and micro-homology analysis with more features and machine learning techniques, to systematically identify and study a potential group of lncRNA orthologous in a number of species across big evolutionary distance, providing new understandings to lncRNA evolution, and more importantly the characterization and classification of their functional roles.

CORRELATION BETWEEN THE CODON USAGES OF THE TRANSMEMBRANE REGIONS AND SUBCELLULAR LOCATIONS OF MEMBRANE PROTEINS

Takuya Yamaguchi, Daiki Takahashi, Kenji Etchuya, Yuri Mukai

Meiji University

Each organellar biomembrane has its own particular thickness and is composed of proteins and lipids with unique proportions. Therefore, it can be thought that the physico-chemical characteristics of the transmembrane regions in the membrane proteins depend on each organelle. Additionally, depending on the variations in biomembranes, the functions necessary to transport proteins to the subcellular organelles which can carry out their biological function accurately vary. Many proteins have organelle targeting signals which determine protein destinations, and their subcellular localization can be predicted based on the signal sequences. However, membrane protein localizations are difficult to predict because many of them have no targeting signals. Therefore, the correlation of the characteristics of the transmembrane region and its subcellular locations were analyzed in this study. First, the human single-pass type II membrane protein data with the experimental evidence in the subcellular location was extracted from the uniprotKB/Swiss-Prot Release2015_03. The usages and positions of codon were analyzed around the transmembrane region of the type II membrane proteins. The Golgi, endoplasmic reticulum and plasma membrane localized proteins were the main focus of this study, because their transport pathways are thought to relate to their subcellular locations. As a result, the unique characteristics of the synonymous codon usages of the arginine and proline residues were found in each organellar membrane. Moreover, the fact that the positions in which the particular codons show high propensities depend on each organelle was also found through the analysis in this study. From these results, the synonymous codon usages and their positions in the transmembrane regions are considered to regulate the subcellular localization of the membrane proteins to the appropriate destination.

A DATA-DRIVEN METHOD FOR IMPUTATION IN CROSS-PLATFORM GENE EXPRESSION STUDIES

Weizhuang Zhou[1], Lichy Han[2], Russ B. Altman[2]

[1]Stanford University Bioengineering Department; [2]Stanford University Biomedical Informatics

Publicly available data from the Gene Expression Omnibus (GEO) are often incorporated into gene expression studies to improve statistical power, many of which are based on Affymetrix microarrays. In particular, the Affymetrix HGU133A and HGU133Plus2 microarray platforms are widely used in gene expression studies, with approximately 1000 GEO Series using the former platform and 4000 using the latter. When different platforms are involved, it is common for researchers to focus only on the subset of common probes. While this is a reasonable approach, discarding data that was already measured is wasteful, and may also affect correlation-based methods in downstream analysis. We hypothesized that the expression values for dissimilar probes can be inferred by using the data from the common probes, and that using this larger feature space can improve downstream analysis. In this work, we used machine learning models to build a series of gene expression inference models for each of the dissimilar probes between the Affymetrix HGU133A and HGU133Plus2 microarray. We show that we can predict the gene expression values of the dissimilar probes to high accuracy, and also demonstrate its utility in downstream analysis in previously published studies.

JUMPING ACROSS BIOMEDICAL CONTEXTS USING COMPRESSIVE DATA FUSION

Marinka Zitnik[1,2], Blaz Zupan[2,3]

[1]Department of Computer Science, Stanford University, USA; [2]Faculty of Computer and Information Science, University of Ljubljana, Slovenia; [3]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA

BACKGROUND. With the rapid growth of diverse biological data, there are more and more entities in the interaction scenarios. With entities we mean genes, drugs, compounds, diseases, publications, ontology terms, pathways and others. Collective learning that can take into consideration tens or even hundreds of data sets describing tens of distinct entity types is becoming a significant challenge. Current, even state-of-the art computational approaches in predictive modeling often disregard the subtlety of different entity and interaction types. However, since entities of different types carry different meanings it does not make sense to mix them without distinguishing their semantics. In recent years, collective latent factor models (e.g., Zitnik and Zupan, IEEE TPAMI 2015 and Zitnik et al., PLoS Comp Biol 2015) have been developed that can handle both multi-relational and multi-entity type data in a principled manner. Methods of this kind have showed promising performance for many predictive tasks including gene function prediction, association mining and prioritization. RESULTS. Assuming a collective latent factor framework applied to a multi-typed biomedical data domain, we focus on the study of relevance of a given entity to other entities that can be of the same or of different type. We address questions of the kind: "Given entities of type A, provide the most relevant entities of type B", where "A" and "B" can be distinct entity types (e.g., genes, pathways, diseases, variants, phenotypes) in the domain. Using compressive data fusion, we first infer a latent factor model that we use to generate composite relations (chains) between any two given entity types. Intuitively, different chains correspond to different semantics and imply different similarities between the entities. We then combine multiple chains together by learning their weights based on the training examples of known relevant entity pairs, which are provided by the user. The approach turns out to be more meaningful in many scenarios compared with the random walk based techniques. The approach is attractive because it can infer new relations between seemingly unrelatable entity types, it can estimate relevance of any entity with regard to any other entity in the domain, and it can model different semantic meanings described by data relations. CONCLUSION. Compressive data fusion and collective matrix factorization provide a powerful mechanism for inference within and across biomedical contexts.

WORKSHOP: COMPUTATIONAL APPROACHES TO STUDY MICROBES AND MICROBIOMES

POSTER PRESENTATIONS

COMBINING BACTERIAL FINGERPRINTS: A NEW ALGORITHM

James A. Foster

University of Idaho, Department of Biological Sciences, Institute for Bioinformatics and Evolutionary Studies

High-throughput sequencing technologies allow researchers to characterize microbial community composition and structure without cultivation. This has made it possible to decipher the complexity of the human microbiome more accurately than ever before. When asking ecological questions, the most common current approach is to use high throughput sequencing to acquire DNA sequences from a deeply conserved housekeeping gene, such as regions in the 16S small ribosomal subunit. Given these sequences, known as amplicons, one reassembles the structure of the microbial community by clustering sequences by similarity, associating clusters with species when possible, and measuring the ecological diversity reflected by these clusters. However, current algorithms for clustering analysis of amplicon sequences presume that the investigator select a single 16S region up front. This biases the downstream analyses, since different regions are more or less appropriate for different bacterial populations. We present a new algorithm and software to combine data from multiple 16S genomic markers, making it possible to overcome the unknown biases and limitations of choosing a single marker. We developed this algorithm to analyze data from a world-wide survey of human breast milk, in order to search for correlations between the microbiome, human milk oligosaccharide (HMO) profiles, and cultural practices.

EXAMINING LOST READS TO SURVEY THE MICROBIOME COMPONENT OF THE HUMAN BODY ACROSS MULTPLE TISSUES

Serghei Mangul[1], Nicolas Strauli[2], Harry Yang[1], Franziska Gruhl[3], Ryan Hernandez[2], Roel Ophoff[1], Eleazar Eskin[1], Noah Zaitlen[2]

[1]University of California Los Angeles; [2]University of California San Francisco; [3]Université de Lausanne

In this work we aim to profile the lost reads to study immune system function across tissues directly from RNA-seq data. We use 1641 RNASeq samples corresponding to 175 individuals and 43 sites from GTEx project : 29 solid organ tissues, 11 brain subregions, whole blood, and two cell lines, LCL and cultured fibroblasts from skin. . We use the unmapped reads to obtain a detailed profile of the microbial and immune components of the human body. We filtered out 54. 68%±7. 28% of the unmapped reads, which were low-quality and/or low complexity as well as 33. 18%±4. 82% of the unmapped reads that mapped to the human reference sequence (MegaBLAST, edit distance 10). The remaining high-quality unique reads are used to perform a survey of the microbiome and immune components. To profile the microbiome we used phylogenetic marker genes to assign candidate microbial reads to the bacterial and archeal taxa. A total of 713 taxa were assigned with Phylosift, with 8 taxa on the phylum level. No microbial organisms were observed in heart, pituitary and adrenal gland. All other tissues contain at least one bacterial or archeal phyla (0. 79+-0. 55 phyla per sample). We observe two viruses harbored in multiple tissues. EBV virus is present in 20% of the skin samples and 66% of the liver samples but it is not present in any of the brain samples. Enterobacteria phage phiX174 virus is present in 20% of the skin samples but is not present in liver and brain tissues. To profile B and T cells across tissues candidate immune reads were mapped to the V(D)J regions of the Ig loci. We compared normalized read counts of V(D)J mapped reads between tissues in order to determine the related abundances of immune cells. As a positive control we confirm that abundances of V(D)J reads in LCLs are significantly greater than thyroid and brain. Consistent with its role as an interface to environmental exposure, lung also contained and increased abundance of V(D) J reads relative to thyroid and brain (p-value < 10-6). We next measured the genetic diversity of V(D)J variants by examine number of gene alleles expressed per sample We observed a lower diversity of gene alleles in the lung and thyroid compared to LCLs (p-value < 10-16). Examining immune and microbial genes in GTEx can help define typical profiles for a healthy tissue. It is essential to monitor microbial and immune diversity, and this work may eventually help diagnose immune and microbiome imbalance in a tissue specific manner.

TIPP: TAXONOMIC IDENTIFICATION AND PHYLOGENETIC PROFILING

Nam-phuong Nguyen[1], Siavash Mirarab[2], Bo Liu[3], Mihai Pop[3], Tandy Warnow[1]

[1]University of Illinois at Urbana-Champaign; [2]University of California at San Diego;
[3]University of Maryland at College Park

Abundance profiling (also called 'phylogenetic profiling') is a crucial step in understanding the diversity of a metagenomic sample, and one of the basic techniques used for this is taxonomic identification of the metagenomic reads. We present taxon identification and phylogenetic profiling (TIPP), a new marker-based taxon identification and abundance profiling method. TIPP combines SEPP (a phylogenetic placement method), with statistical techniques to control the classification precision and recall, and results in improved abundance profiles. TIPP is highly accurate even in the presence of high indel errors and novel genomes, and matches or improves on previous approaches, including NBC, mOTU, PhymmBL, MetaPhyler and MetaPhlAn.

ENTROPY, STATISTICAL DEPENDENCE, AND THE NETWORK STRUCTURE OF THE INFANT MICROBIOME

Weston Viles[1], Hilary G. Morrison[2], Mitchell L. Sogin[2], Jason H. Moore[3], Julitte C. Madan[4], Margaret R. Karagas[1], Anne G. Hoen[1]

[1]Geisel School of Medicine, Dartmouth College; [2]Josephine Bay Paul Center Marine Biological Laboratory; [3]Perelman School of Medicine, University of Pennsylvania; [4]Children's Hospital at Dartmouth

The microbes that inhabit the human gut are known to play critical roles in digestion and vitamin synthesis, conferring resistance to colonization by pathogens, and immune programming. These functions are particularly important during infant and child development and, it is believed, that the infant gut-associated microbiota may impact long-term health. The infant gut microbiome contains numerous bacterial taxa that coevolved in association with their hosts. While it is known that these organisms interact with one another, the organizing principals of this complex network have not been established. Here we examine the interactions among taxa in the infant gut microbiome using a maximum entropy approach applied to 255 bacterial operational taxonomic units (OTUs) identified in 111 6 week old infants followed as part of the New Hampshire Birth Cohort Study. We compare the entropy of the distribution of OTUs among infants with that of the theoretical maximum entropy assuming up to K-way interactions among OTUs, from K=1, in which OTUs are independently distributed among infants, to K=M, M the number of considered OTUs, which allows for interactions of arbitrary complexity. We identify the minimum order of interactions among OTUs that provides an effective description of the system, i.e. the order for which a sufficient reduction of entropy is observed. This is determined via comparison of the entropy difference IM=S1-SM, the total amount of correlation in the network, with I2=S1-S2 (the entropy difference for pairwise correlations), I3=S1-S3 (for 3-way correlations), and, so on. Our results demonstrate that complex, higher order (K>2) interactions among microbial taxa govern the assembly and organization of the infant microbiome. The successful development of therapies targeting the human microbiota, e.g. probiotics, requires that careful attention be paid to understanding the nature and structure of these complex, interacting sub-systems.

WORKSHOP: USE OF GENOME DATA IN NEWBORNS AS A STARTING POINT FOR LIFE-LONG PRECISION MEDICINE

POSTER PRESENTATION

DIAGNOSTIC ROLE OF EXOME SEQUENCING IN IMMUNE DEFICIENCY DISORDERS

Aashish N. Adhikari[1], Jay P. Patel[2], Alice Y. Chan[3], Divya Punwani[3], Haopeng Wang[3], Antonia Kwan[3], Theresa A. Kadlecek[3], Morton J. Cowan[3], Marianne Mollenauer[3], John Kuriyan[1], Shu Man Fu[4], Uma Sunderam[5], Sadhna Rana[5], Ajithavalli Chellappan[5], Kunal Kundu[5], Arend Mulder[6], Frans H. J. Claas[6], Joseph A. Church[7], Arthur Weiss[3], Richard A. Gatti[8], Jennifer M. Puck[3], Rajgopal Srinivasan[5] , Steven E. Brenner[1]

[1]University of California, Berkeley, CA, USA; [2]Children's Hospital of Los Angeles, Los Angeles, CA, USA; [3]University of California, San Francisco, CA, USA; [4]University of Virginia School of Medicine, Charlottesville, VA, USA; [5]Innovation Labs, Tata Consultancy Services Hyderabad, AP, India; [6]Leiden University Medical Centre, Leiden, The Netherlands; [7]University of Southern California, Los Angeles, CA, USA; [8]University of California, Los Angeles, CA, USA;

We developed an analysis protocol for individual genome interpretation and used its distinctive features to diagnose numerous clinical cases. We applied the protocol to exomes from newborn patients with undiagnosed primary immune disorders. To yield high quality sets of possible causative variants, we used multiple callers with multisample calling and integrated variant annotation, variant filtering, and gene prioritization. In two unrelated infant immunodeficient girls with no diagnoses, we discovered compound heterozygous variants in the ATM gene for both the infants offering a very early diagnosis of Ataxia Telangiectasia (AT) which allowed for avoidance of undue irradiation and live vaccinations. In another case, the affected siblings had early onset bullous phempigoid, a chronic autoimmune disorder. Our analysis revealed compound heterozygous mutations in ZAP70, a gene associated with profound primary immunodeficiency, the opposite phenotype. Cellular immunological studies indicated that one variant was hypomorphic and the other was hyperactive. These combined to yield a novel presentation, adding to the existing phenotype repertoire of ZAP70 in humans. We also discovered variants in PRKDC occurring after the genomically-encoded stop codon, since we correctly identified a premature stop codon in the stop codon resulting from a single base deletion. Our protocol has been similarly revealing in other SCID and CID cases including Nijmegen Breakage Syndrome, which highlight unique features of the analysis framework that facilitate genetic discovery. These help provide crucial information to offer prompt appropriate treatment, family genetic counseling, and avoidance of diagnostic odyssey.

AUTHOR INDEX

E

F

G

H

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z