Pacific Symposium on Biocomputing 2013

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 #10-41 are not assigned poster space.

Papers are organized by session then the last name of the first author. Presenting authors' names are underlined.

ABERRANT PATHWAYS
ACCEPTED PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

IDENTIFYING MASTER REGULATORS OF CANCER AND THEIR DOWNSTREAM TARGETS BY INTEGRATING GENOMIC AND EPIGENOMIC ABERRANT FEATURES

Gevaert O, Plevritis S
Stanford University

Vast amounts of molecular data characterizing the genome, epigenome and transcriptome are becoming available for a variety of cancers. The current challenge is to integrate these diverse layers of molecular biology information to create a more comprehensive view of key biological processes underlying cancer. We developed a biocomputational algorithm that integrates copy number, DNA methylation, mutation and gene expression data to study master regulators of cancer and identify their targets. Our algorithm starts by generating a list of candidate driver genes based on the rationale that genes that are driven by multiple genomic events in a subset of samples are unlikely to be randomly deregulated. We then select the master regulators from the candidate driver and identify their targets by inferring the underlying regulatory network of gene expression. We applied our biocomputational algorithm to identify master regulators and their targets in glioblastoma multiforme (GBM) and serous ovarian cancer. Our results suggest that the expression of candidate drivers is more likely to be influenced by copy number variations than DNA methylation. Next, we selected the master regulators and identified their downstream targets using module networks analysis. As a proof-of-concept, we show that the GBM and ovarian cancer module networks recapitulate known processes in these cancers. In addition, we identify master regulators that have not been previously reported and suggest their likely role. In summary, focusing on genes whose expression can be explained by their genomic and epigenomic aberrations is a promising strategy to identify master regulators of cancer.

MODULE COVER - A NEW APPROACH TO GENOTYPE-PHENOTYPE STUDIES

Yoo-Ah Kim, Raheleh Salari, Stefan Wuchty, Teresa M. Przytycka

Uncovering and interpreting phenotype/genotype relationships are among the most challenging open questions in disease studies. Set cover approaches are explicitly designed to provide a representative set for diverse disease cases and thus are valuable in studies of heterogeneous datasets. At the same time pathway-centric methods have emerged as key approaches that significantly empower studies of genotype-phenotype relationships. Combining the utility of set cover techniques with the power of network-centric approaches, we designed a novel approach that extends the concept of set cover to network modules cover. We developed two alternative methods to solve the module cover problem: (i) an integrated method that simultaneously determines network modules and optimizes the coverage of disease cases. (ii) a two-step method where we first determined a candidate set of network modules and subsequently selected modules that provided the best coverage of the disease cases. The integrated method showed superior performance in the context of our application. We demonstrated the utility of the module cover approach for the identification of groups of related genes whose activity is perturbed in a coherent way by specific genomic alterations, allowing the interpretation of the heterogeneity of cancer cases.

INTERPRETING PERSONAL TRANSCRIPTOMES: PERSONALIZED MECHANISM-SCALE PROFILING OF RNA-SEQ DATA

Alan Perez-Rathke
Department of Medicine, University of Illinois at Chicago

Haiquan Li
Department of Medicine, University of Illinois at Chicago

Yves A. Lussier
Departments of Medicine & Bioengineering, University of Illinois at Chicago

Despite thousands of reported studies unveiling gene-level signatures for complex diseases, few of these techniques work at the single-sample level with explicit underpinning of biological mechanisms. This presents both a critical dilemma in the field of personalized medicine as well as a plethora of opportunities for analysis of RNA-seq data. In this study, we hypothesize that the "Functional Analysis of Individual Microarray Expression" (FAIME) method we developed could be smoothly extended to RNA-seq data and unveil intrinsic underlying mechanism signatures across different scales of biological data for the same complex disease. Using publicly available RNA-seq data for gastric cancer, we confirmed the effectiveness of this method (i) to translate each sample transcriptome to pathway-scale scores, (ii) to predict deregulated pathways in gastric cancer against gold standards (FDR<5%, Precision=75%, Recall =92%), and (iii) to predict phenotypes in an independent dataset and expression platform (RNA-seq vs microarrays, Fisher Exact Test p<10-6). Measuring at a single-sample level, FAIME could differentiate cancer samples from normal ones; furthermore, it achieved comparative performance in identifying differentially ex-pressed pathways as compared to state-of-the-art cross-sample methods. These results motivate future work on mechanism-level biomarker discovery predictive of diagnoses, treatment, and therapy.

COMPUTATIONAL DRUG REPOSITIONING
ACCEPTED PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

EVALUATION OF ANALYTICAL METHODS FOR CONNECTIVITY MAP DATA

Jie Cheng 1, Qing Xie 2, Vinod Kumar 2, Mark Hurle 2, Johannes M. Freudenberg 3, Lun Yang 2, Pankaj Agarwal 2

1 Statistical and Platform Technologies, GlaxoSmithKline R&D
UP4335, 1250 S Collegeville Rd, Collegeville, PA 19426 USA
2 Computational Biology, GlaxoSmithKline R&D
UW2230, 709 Swedeland Road, King of Prussia, PA 19406 USA
3 Computational Biology, GlaxoSmithKline R&D, 3.2085A, 5 Moore Drive, Durham, NC, 27709, USA

Connectivity map data and associated methodologies have become a valuable tool in understanding drug mechanism of action (MOA) and discovering new indications for drugs. However, few systematic evaluations have been done to assess the accuracy of these methodologies. One of the difficulties has been the lack of benchmarking data sets. Iskar et al. (PLoS. Comput. Biol. 6, 2010) predicted the Anatomical Therapeutic Chemical (ATC) drug classification based on drug-induced gene expression profile similarity (DIPS), and quantified the accuracy of their method by computing the area under the curve (AUC) of the Receiver Operating Characteristic (ROC) curve. We adopt the same data and extend the methodology, by using a simpler eXtreme cosine (XCos) method, and find it does better in this limited setting than the Kolmogorov-Smirnov (KS) statistic. In fact, for partial AUC (a more relevant statistic for actual application to repositioning) XCos does 17% better than the DIPS method ($p=1.2e-7$). We also observe that smaller gene signatures (with 100 probes) do better than larger ones (with 500 probes), and that DMSO controls from within the same batch obviate the need for mean centering. As expected there is heterogeneity in the prediction accuracy amongst the various ATC codes. We find that good transcriptional response to drug treatment appears necessary but not sufficient to achieve high AUCs. Certain ATC codes, such as those corresponding to corticosteroids, had much higher AUCs possibly due to strong transcriptional responses and consistency in MOA.

A NOVEL MULTI-MODAL DRUG REPURPOSING APPROACH FOR IDENTIFICATION OF POTENT ACK1 INHIBITORS

Sharangdhar S. Phatak, Shuxing Zhang

Exploiting drug polypharmacology to identify novel modes of actions for drug repurposing has gained significant attentions in the current era of weak drug pipelines. From a serendipitous to systematic or rational ways, a variety of unimodal computational approaches have been developed but the complexity of the problem clearly needs multi-modal approaches for better solutions. In this study, we propose an integrative computational framework based on classical structure-based drug design and chemical-genomic similarity methods, combined with molecular graph theories for this task. Briefly, a pharmacophore modeling method was employed to guide the selection of docked poses resulting from our high-throughput virtual screening. We then evaluated if complementary results (hits missed by docking) can be obtained by using a novel chemo-genomic similarity approach based on chemical/sequence information. Finally, we developed a bipartite-graph based on the extensive data curation of DrugBank, PDB, and UniProt. This drug-target bipartite graph was used to assess similarity of different inhibitors based on their connections to other compounds and targets. The approaches were applied to the repurposing of existing drugs against ACK1, a novel cancer target significantly overexpressed in breast and prostate cancers during their progression. Upon screening of ~1,447 marketed drugs, a final set of 10 hits were selected for experimental testing. Among them, four drugs were identified as potent ACK1 inhibitors. Especially the inhibition of ACK1 by Dasatinib was as strong as IC50=1nM. We anticipate that our novel, integrative strategy can be easily extended to other biological targets with a more comprehensive coverage of known bio-chemical space for repurposing studies.

PROTEIN-CHEMICAL INTERACTION PREDICTION VIA KERNELIZED SPARSE LEARNING SVM

Yi Shi 1, Xinhua Zhang 1, Xiaoping Liao 2, Guohui Lin 1, Dale Schuurmans 1

1 Department of Computing Science, University of Alberta, Edmonton, Alberta T6G 2E8, Canada
2 Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, Alberta T6G 2P5, Canada

Given the difficulty of experimental determination of drug-protein interactions, there is a significant motivation to develop effective in silico prediction methods that can provide both new predictions for experimental verification and supporting evidence for experimental results. Most recently, classification methods such as support vector machines (SVMs) have been applied to drug-target prediction. Unfortunately, these methods generally rely on measures of the maximum "local similarity" between two protein sequences, which could mask important drug-protein interaction information since drugs are much smaller molecules than proteins and drug-target binding regions must comprise only small local regions of the proteins. We therefore develop a novel sparse learning method that considers sets of short peptides. Our method integrates feature selection, multi-instance learning, and Gaussian kernelization into an L1 norm support vector machine classifier. Experimental results show that it not only outperformed the previous methods but also pointed to an optimal subset of potential binding regions.

DRUG TARGET PREDICTIONS BASED ON HETEROGENEOUS GRAPH INFERENCE

Wenhui Wang†, Sen Yang† and, Jing Li*

Department of Electrical Engineering and Computer Science
Case Western Reserve University Cleveland, Ohio, 44106, USA
Emails:{wxw134@case.edu,sxy221@case.edu, jingli@case.edu}

A key issue in drug development is to understand the hidden relationships among drugs and targets. Computational methods for novel drug target predictions can greatly reduce time and costs compared with experimental methods. In this paper, we propose a network based computational approach for novel drug and target association predictions. More specifically, a heterogeneous drug-target graph, which incorporates known drug-target interactions as well as drug-drug and target-target similarities, is first constructed. Based on this graph, a novel graph-based inference method is introduced. Compared with two state-of-the-art methods, large-scale cross-validation results indicate that the proposed method can greatly improve novel target predictions.

EPIGENOMICS
ACCEPTED PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

USING DNASE DIGESTION DATA TO ACCURATELY IDENTIFY TRANSCRIPTION FACTOR BINDING SITES

Kaixuan Luo and Alexander J. Hartemink

Program in Computational Biology and Bioinformatics, and Department of Computer Science, Duke University, Durham, NC 27708, USA

Identifying binding sites of transcription factors (TFs) is a key task in deciphering transcriptional regulation. ChIP-based methods are used to survey the genomic locations of a single TF in each experiment. But methods combining DNase digestion data with TF binding specificity information could potentially be used to survey the locations of many TFs in the same experiment, provided such methods permit reasonable levels of sensitivity and specificity. Here, we present a simple such method that outperforms a leading recent method, CENTIPEDE, marginally in human but dramatically in yeast (average auROC across 20 TFs increases from 74% to 94%). Our method is based on logistic regression and thus benefits from supervision, but we show that partially and completely unsupervised variants perform nearly as well. Because the number of parameters in our method is at least an order of magnitude smaller than CENTIPEDE, we dub it MILLIPEDE.

EPIGENOMIC MODEL OF CARDIAC ENHANCERS WITH APPLICATION TO GENOME WIDE ASSOCIATION STUDIES

Avinash Das Sahu, Radhouane Aniba, Yen-Pei Christy Chang, Sridhar Hannenhall

Mammalian gene regulation is often mediated by distal enhancer elements, in particular, for tissue specific and developmental genes. Computational identification of enhancers is difficult because they do not exhibit clear location preference relative to their target gene and also because they lack clearly distinguishing genomic features. This represents a major challenge in deciphering transcriptional regulation. Recent ChIP-seq based genome-wide investigation of epigenomic modifications have revealed that enhancers are often enriched for certain epigenomic marks. Here we utilize the epigenomic data in human heart tissue along with validated human heart enhancers to develop a Support Vector Machine (SVM) model of cardiac enhancers. Cross-validation classification accuracy of our model was 84\% and 92\% on positive and negative sets respectively with ROC AUC = 0.92. More importantly, while P300 binding has been used as gold standard for enhancers, our model can distinguish P300-bound validated enhancers from other P300-bound regions that failed to exhibit enhancer activity in transgenic mouse. While GWAS studies reveal polymorphic regions associated with certain phenotypes, they do not immediately provide causality. Next, we hypothesized that genomic regions containing a GWAS SNP associated with a cardiac phenotype might contain another SNP in a cardiac enhancer, which presumably mediates the phenotype. Starting with a comprehensive set of SNPs associated with cardiac phenotypes in GWAS studies, we scored other SNPs in LD with the GWAS SNP according to its probability of being an enhancer and choose one with best score in the LD as enhancer. We found that our predicted enhancers are enriched for known cardiac transcriptional regulator motifs and are likely to regulate the nearby gene. Importantly, these tendencies are more favorable for the predicted enhancers compared with an approach that uses P300 binding as a marker of enhancer activity.

PERSONALIZED MEDICINE: FROM GENOTYPES AND MOLECULAR PHENOTYPES TOWARDS THERAPY
ACCEPTED PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

ASSEMBLY MATCHING PURSUIT

Surojit Biswas and Vladimir Jojic

Metagenomics, the study of the total genetic material isolated from a biological host, promises to reveal host-microbe or microbe-microbe interactions that may help to personalize medicine or improve agronomic practice. We introduce a method that discovers metagenomic units (MGUs) relevant for phenotype prediction through sequence-based dictionary learning. The method aggregates patient-specific dictionaries and estimates MGU abundances in order to summarize a whole population and yield universally predictive biomarkers. We analyze the impact of Gaussian, Poisson, and Negative Binomial read count models in guiding dictionary construction by examining classification efficiency on a number of synthetic datasets and a real dataset from Ref. 1. Each outperforms standard methods of dictionary composition, such as random projection and orthogonal matching pursuit. Additionally, the predictive MGUs they recover are biologically relevant.

CHARACTERIZATION OF THE METABOCHIP IN DIVERSE POPULATIONS FROM THE INTERNATIONAL HAPMAP PROJECT IN THE EPIDEMIOLOGIC ARCHITECTURE FOR GENES LINKED TO ENVIRONMENT (EAGLE) PROJECT

Dana C. Crawford
Department of Molecular Physiology and Biophysics, Center for Human Genetics Research, Vanderbilt University, 2215 Garland Avenue, 519 Light Hall Nashville, TN 37232, USA Email: crawford@chgr.mc.vanderbilt.edu

Robert Goodloe
Center for Human Genetics Research, Vanderbilt University, 2215 Garland Avenue, 519 Light Hall Nashville, TN 37232, USA Email: robert.j.goodloe@vanderbilt.edu

Kristin Brown-Gentry
Center for Human Genetics Research, Vanderbilt University, 1207 17th Avenue, Suite 300 Nashville, TN 37232, USA Email: kristin.brown@chgr.mc.vanderbilt.edu

Sarah Wilson
Center for Human Genetics Research, Vanderbilt University, 1207 17th Avenue, Suite 300 Nashville, TN 37232, USA Email: sarah.wilson@chgr.mc.vanderbilt.edu

Jamie Roberson
Center for Human Genetics Research, Vanderbilt University, 2215 Garland Avenue, 519 Light Hall Nashville, TN 37232, USA Email: jamie.l.roberson@vanderbilt.edu

Niloufar B. Gillani
Center for Human Genetics Research, Vanderbilt University, 2215 Garland Avenue, 519 Light Hall Nashville, TN 37232, USA Email: nila.gillani@vanderbilt.edu

Marylyn D. Ritchie
Department of Biochemistry and Molecular Biology, Center for System Genomics, Pennsylvania State University, 512 Wartik Lab University Park, PA 16802, USA Email: marylyn.ritchie@psu.edu

Holli H. Dilks
Department of Molecular Physiology and Biophysics, Center for Human Genetics Research, Vanderbilt University, 2215 Garland Avenue, 519 Light Hall Nashville, TN 37232, USA Email: holli.dilks@chgr.mc.vanderbilt.edu

William S. Bush
Department of Biomedical Informatics, Center for Human Genetics Research, Vanderbilt University, 2215 Garland Avenue, 519 Light Hall Nashville, TN 37232, USA Email: william.s.bush@vanderbilt.edu

Genome-wide association studies (GWAS) have identified hundreds of genomic regions associated with common human disease and quantitative traits. A major research avenue for mature genotype-phenotype associations is the identification of the true risk or functional variant for downstream molecular studies or personalized medicine applications. As part of the Population Architecture using Genomics and Epidemiology (PAGE) study, we as Epidemiologic Architecture for Genes Linked to Environment (EAGLE) are fine-mapping GWAS-identified genomic regions for common diseases and quantitative traits. We are currently genotyping the Metabochip, a custom content BeadChip designed for fine-mapping metabolic diseases and traits, in~15,000 DNA samples from patients of African, Hispanic, and Asian ancestry linked to de-identified electronic medical records from the Vanderbilt University biorepository (BioVU). As an initial study of quality control, we report here the genotyping data for 360 samples of European, African, Asian, and Mexican descent from the International HapMap Project. In addition to quality control metrics, we report the overall allele frequency distribution, overall population differentiation (as measured by FST), and linkage disequilibrium patterns for a select GWAS-identified region associated with low-density lipoprotein cholesterol levels to illustrate the utility of the Metabochip for fine-mapping studies in the diverse populations expected in EAGLE, the PAGE study, and other efforts underway designed to characterize the complex genetic architecture underlying common human disease and quantitative traits.

INSIGHTS INTO DISEASES OF HUMAN TELOMERASE FROM DYNAMICAL MODELING

Samuel Coulbourn Flores
Cell and Molecular Biology Department, Uppsala University, Biomedical Center, Box 596, 75124 Uppsala, Sweden Email: samuel.flores@icm.uu.se

Georgeta Zemora
Max F. Perutz Laboratories, University of Vienna, Dr. Bohrgasse 9/5, 1030 Vienna, Austria

Christina Waldsich
Max F. Perutz Laboratories, University of Vienna, Dr. Bohrgasse 9/5, 1030 Vienna, Austria

Mutations in the telomerase complex disrupt either nucleic acid binding or catalysis, and are the cause of numerous human diseases. Despite its importance, the structure of the human telomerase complex has not been observed crystallographically, nor are its dynamics understood in detail. Fragments of this complex from Tetrahymena thermophila and Tribolium castaneum have been crystallized. Biochemical probes provide important insight into dynamics. In this work we summarize evidence that the T. castaneum structure is Telomerase Reverse Transcriptase. We use this structure to build a partial model of the human Telomerase complex. The model suggests an explanation for the structural role of several disease-associated mutations. We then generate a 3D kinematic trajectory of telomere elongation to illustrate a "typewriter" mechanism: the RNA template moves to keep the end of the growing telomeric primer in the active site, disengaging after every 6-residue extension to execute a "carriage return" and go back to its starting position. A hairpin can easily form in the primer, from DNA residues leaving the primer-template duplex. The trajectory is consistent with available experimental evidence. The methodology is extensible to many problems in structural biology in general and personalized medicine in particular.

SPECTRAL CLUSTERING STRATEGIES FOR HETEROGENEOUS DISEASE EXPRESSION DATA

Grace T. Huang (1,2,3), Kathryn I. Cunningham (4), Panayiotis V. Benos (1,3), and Chakra S. Chennubhotla (1)

1 Department of Computational and Systems Biology
2 Joint CMU-Pitt PhD Program in Computational Biology
3 Clinical and Translational Science Institute University of Pittsburgh, Pittsburgh, Pennsylvania, USA
4 Department of Computer Science, University of Arizona, Tucson, Arizona, USA

Clustering of gene expression data simplifies subsequent data analyses and forms the basis of numerous approaches for biomarker identification, prediction of clinical outcome, and personalized therapeutic strategies. The most popular clustering methods such as K-means and hierarchical clustering are intuitive and easy to use, but they require arbitrary choices on their various parameters (number of clusters for K-means, and a threshold to cut the tree for hierarchical clustering). Human disease gene expression data are in general more difficult to cluster efficiently due to background (genotype) heterogeneity, disease stage and progression differences and disease subtyping; all of which cause gene expression datasets to be more heterogeneous. Spectral clustering has been recently introduced in many fields as a promising alternative to standard clustering methods. The idea is that pairwise comparisons can help reveal global features through the eigen techniques. In this paper, we developed a new recursive K-means spectral clustering method (ReKS) for disease gene expression data. We benchmarked ReKS on three large-scale cancer datasets and we compared it to different clustering methods with respect to execution time, background models and external biological knowledge. We found ReKS to be superior to the hierarchical methods and equally good to K-means, but much faster than them and without the requirement for a priori knowledge of K. Overall, ReKS offers an attractive alternative for efficient clustering of human disease data.

SYSTEMATIC IDENTIFICATION OF RISK FACTORS FOR ALZHEIMER'S DISEASE THROUGH SHARED GENETIC ARCHITECTURE AND ELECTRONIC MEDICAL RECORDS

Li Li
Stanford University School of Medicine, Pediatric Department/Systems Medicine Division

David Ruau
Stanford University School of Medicine, Pediatric Department/Systems Medicine Division

Rong Chen
Personalis Inc.

Susan Weber
Stanford Center for Clinical Informatics, Stanford University School of Medicine

Atul Butte
Stanford University School of Medicine, Pediatric Department/Systems Medicine Division

Alzheimer's disease (AD) is one of the leading causes of death for older people in US with rapidly increasing incidence. AD irreversibly and progressively damages the brain, but there are treatments in clinical trials to potentially slow the development of AD. We hypothesize that the presence of clinical traits, sharing common genetic variants with AD, could be used as a non-invasive means to predict AD or trigger for administration of preventative therapeutics. We developed a method to compare the genetic architecture between AD and traits from prior GWAS studies. Six clinical traits were significantly associated with AD, capturing 5 known risk factors and 1 novel association: erythrocyte sedimentation rate (ESR). The association of ESR with AD was then validated using Electronic Medical Records (EMR) collected from Stanford Hospital and Clinics. We found that female patients and with abnormally elevated ESR were significantly associated with higher risk of AD diagnosis (OR: 1.85 [1.32-2.61], p=0.003), within 1 year prior to AD diagnosis (OR: 2.31 [1.06-5.01], p=0.032), and within 1 year after AD diagnosis (OR: 3.49 [1.93-6.31], p<0.0001). Additionally, significantly higher ESR values persist for all time courses analyzed. Our results suggest that ESR should be tested in a specific longitudinal study for association with AD diagnosis, and if positive, could be used as a prognostic marker.

PHYLOGENOMICS AND POPULATION GENOMICS: MODELS, ALGORITHMS, AND ANALYTICAL TOOLS
ACCEPTED PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

EVALUATING VARIATIONS ON THE STAR ALGORITHM FOR RELATIVE EFFICIENCY AND SAMPLE SIZES NEEDED TO RECONSTRUCT SPECIES TREES

James H Degnan

University of Canterbury

Many methods for inferring species trees from gene trees have been developed when incongruence among gene trees is due to incomplete lineage sorting. A method called STAR (Liu et al, 2009), assigns values to nodes in gene trees based only on topological information and uses the average value of the most recent common ancestor node for each pair of taxa to construct a distance matrix which is then used for clustering taxa into a tree. This method is very efficient computationally, scaling linearly in the number of loci and quadratically in the number of taxa, and in simulations has shown to be highly accurate for moderate to large numbers of loci as well as robust to molecular clock violations and misestimation of gene trees from sequence data. The method is based on a particular choice of numbering nodes in the gene trees; however, other choices for numbering nodes in gene trees can also lead to consistent inference of the species tree. Here, expected values and variances for average pairwise distances and differences between average pairwise distances in the distance matrix constructed by the STAR algorithm are used to analytically evaluate efficiency of different numbering schemes that are variations on the original STAR numbering for small trees.

THE BEHAVIOR OF ADMIXED POPULATIONS IN NEIGHBOR-JOINING INFERENCE OF POPULATION TREES

Naama M. Kopelman and Lewi Stone
Porter School of Environmental Studies, Department of Zoology, Tel Aviv University, Ramat Aviv, Israel

Olivier Gascuel
Methodes et Algorithmes pour la Bioinformatique, LIRMM-CNRS, Montpellier, France

Noah A. Rosenberg
Department of Biology, Stanford University, Stanford, California, USA

Neighbor-joining is one of the most widely used methods for constructing evolutionary trees. This approach from phylogenetics is often employed in population genetics, where distance matrices obtained from allele frequencies are used to produce a representation of population relationships in the form of a tree. In phylogenetics, the utility of neighbor-joining derives partly from a result that for a class of distance matrices including those that are additive or tree-like---generated by summing weights over the edges connecting pairs of taxa in a tree to obtain pairwise distances---application of neighbor-joining recovers exactly the underlying tree. For populations within a species, however, migration and admixture can produce distance matrices that reflect more complex processes than those obtained from the bifurcating trees typical in the multispecies context. Admixed populations---populations descended from recent mixture of groups that have long been separated---have been observed to be located centrally in inferred neighbor-joining trees, with short external branches incident to the path connecting their source populations. Here, using a simple model, we explore mathematically the behavior of an admixed population under neighbor-joining. We show that with an additive distance matrix, a population admixed among two source populations necessarily lies on the path between the sources. Relaxing the additivity requirement, we examine the smallest nontrivial case---four populations, one of which is admixed between two of the other three---showing that the two source populations never merge with each other before one of them merges with the admixed population. Furthermore, the distance on the constructed tree between the admixed population and either source population is always smaller than the distance between the source populations, and the external branch for the admixed population is always incident to the path connecting the sources. We define three properties that hold for four taxa and that we hypothesize are satisfied under more general conditions: antecedence of clustering, intermediacy of distances, and intermediacy of path lengths. Our findings can inform interpretations of neighbor-joining trees with admixed groups, and they provide an explanation for patterns observed in trees of human populations.

MAXIMUM LIKELIHOOD PHYLOGENETIC RECONSTRUCTION FROM HIGH-RESOLUTION WHOLE-GENOME DATA AND A TREE OF 68 EUKARYOTES

Yu Lin
Laboratory for Computational Biology and Bioinformatics, EPFL, Lausanne VD, CH-1015, Switzerland, (E-mail: yu.lin@epfl.ch)

Fei Hu
Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA, (E-mail: hu5@cse.sc.edu)

Jijun Tang
Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA, (E-mail: jtang@cse.sc.edu)

Bernard M.E. Moret
Laboratory for Computational Biology and Bioinformatics, EPFL, Lausanne VD, CH-1015, Switzerland, (E-mail: bernard.moret@epfl.ch)

The rapid accumulation of whole-genome data has renewed interest in the study of the evolution of genomic architecture, under such events as rearrangements, duplications, losses. Comparative genomics, evolutionary biology, and cancer research all require tools to elucidate the mechanisms, history, and consequences of those evolutionary events, while phylogenetics could use whole-genome data to enhance its picture of the Tree of Life. Current approaches in the area of phylogenetic analysis are limited to very small collections of closely related genomes using low-resolution data (typically a few hundred syntenic blocks); moreover, these approaches typically do not include duplication and loss events. We describe a maximum likelihood (ML) approach for phylogenetic analysis that takes into account genome rearrangements as well as duplications, insertions, and losses. Our approach can handle high-resolution genomes (with 40,000 or more markers) and can use in the same analysis genomes with very different numbers of markers. Because our approach uses a standard ML reconstruction program (RAxML), it scales up to large trees. We present the results of extensive testing on both simulated and real data showing that our approach returns very accurate results very quickly. In particular, we analyze a dataset of 68 high-resolution eukaryotic genomes, with from 3,000 to 42,000 genes, from the eGOB database; the analysis, including bootstrapping, takes just 3 hours on a desktop system and returns a tree in agreement with all well supported branches, while also suggesting resolutions for some disputed placements.

AN ANALYTICAL COMPARISON OF MULTILOCUS METHODS UNDER THE MULTISPECIES COALESCENT: THE THREE-TAXON CASE

Sebastien Roch
UW-Madison

Incomplete lineage sorting (ILS) is a common source of gene tree incongruence in multilocus analyses. Numerous approaches have been developed to infer species trees in the presence of ILS. Here we provide a mathematical analysis of several coalescent-based methods. The analysis is performed on a three-taxon species tree and assumes that the gene trees are correctly reconstructed along with their branch lengths. It suggests that maximum likelihood (and some equivalents) can be significantly more accurate in this setting than other methods, especially as ILS gets more pronounced.

POST-NGS: ANALYSIS OF -OMES GENERATED BY NGS
ACCEPTED PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

LSHPLACE: FAST PHYLOGENETIC PLACEMENT USING LOCALITY-SENSITIVE HASHING

Daniel G. Brown and Jakub Truszkowski

David R. Cheriton School of Computer Science University of Waterloo
Waterloo ON N2L 3G1 Canada

We consider the problem of phylogenetic placement, in which large numbers of sequences (often next-generation sequencing reads) are placed onto an existing phylogenetic tree. We adapt our recent work on phylogenetic tree inference, which uses ancestral sequence reconstruction and locality-sensitive hashing, to this domain. With these ideas, new sequences can be placed onto trees with high fidelity in strikingly fast runtimes. Our results are two orders of magnitude faster than existing programs for this domain, and show a modest accuracy tradeoff. Our results offer the possibility of analyzing many more reads in a next-generation sequencing project than is currently possible.

ChIPMODULE: SYSTEMATIC DISCOVERY OF TRANSCRIPTION FACTORS AND THEIR COFACTORS FROM CHIP-SEQ DATA

Jun Ding
Department of EECS, University of Central Florida, 4000 central Florida Blvd Orlando, FL 32816, USA Email: jding@cs.ucf.edu

Xiaohui Cai
Shanghai Center for Bioinformation Technology, 100 Qinzhou Rd, Bldg.1, Fl.12 Shanghai, 200235, China Email: xhcai@scbit.org

Ying Wang
Department of EECS, University of Central Florida, 4000 central Florida Blvd Orlando, FL 32816, USA Email: ying2010@knights.ucf.edu

Haiyan Hu
Department of EECS, University of Central Florida, 4000 central Florida Blvd Orlando, FL 32816, USA Email: haihu@cs.ucf.edu

Xiaoman Li
Burnett School of Biomedical Science, University of Central Florida, 4000 central Florida Blvd Orlando, FL 32816, USA Email: xiaoman@mail.ucf.edu

We have developed a novel approach called ChIPModule to systematically discover transcription factors and their cofactors from ChIP-seq data. Given a ChIP-seq dataset and the binding patterns of a large number of transcription factors, ChIPModule can efficiently identify groups of transcription factors, whose binding sites significantly co-occur in the ChIP-seq peak regions. By testing ChIPModule on simulated data and experimental data, we have shown that ChIPModule identifies known cofactors of transcription factors, and predicts new cofactors that are supported by literature. ChIPModule provides a useful tool for studying gene transcriptional regulation.

DETECTING HIGHLY DIFFERENTIATED COPY-NUMBER VARIANTS FROM POOLED POPULATION SEQUENCING

Daniel R. Schrider
Department of Biology and School of Informatics and Computing, Indiana University

David J. Begun
Department of Evolution and Ecology, University of California, Davis

Matthew W. Hahn
Department of Biology and School of Informatics and Computing, Indiana University

Copy-number variants (CNVs) represent a functionally and evolutionarily important class of variation. Here we take advantage of the use of pooled sequencing to detect CNVs with large differences in allele frequency between population samples. We present a method for detecting CNVs in pooled population samples using a combination of paired-end sequences and read-depth. Highly differentiated CNVs show large differences in the number of paired-end reads supporting individual alleles and large differences in read-depth between population samples. We complement this approach with one that uses a hidden Markov model to find larger regions differing in read-depth between samples. Using novel pooled sequence data from two populations of Drosophila melanogaster along a latitudinal cline, we demonstrate the utility of our method for identifying CNVs involved in local adaptation.

TEXT AND DATA MINING FOR BIOMEDICAL DISCOVERY
ACCEPTED PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

ATHENA: A TOOL FOR META-DIMENSIONAL ANALYSIS APPLIED TO GENOTYPES AND GENE EXPRESSION DATA TO PREDICT HDL CHOLESTEROL LEVELS

Emily R. Holzinger †
Center for Human Genetics Research, Vanderbilt University Nashville, TN 37232, USA
Email: emily.r.holzinger@vanderbilt.edu

Scott M. Dudek
Center for Systems Genomics, Pennsylvania State University University Park, PA 16803, USA Email: sud23@psu.edu

Alex T. Frase
Center for Systems Genomics, Pennsylvania State University University Park, PA 16803, USA Email: alex.frase@psu.edu

Ronald M. Krauss
Children's Hospital Oakland Research Institute Oakland, CA 94609, USA
Email: rkrauss@chori.org

Marisa W. Medina
Children's Hospital Oakland Research Institute Oakland, CA 94609, USA
Email: mwmedina@chori.org

Marylyn D. Ritchie
Center for Systems Genomics, Pennsylvania State University University Park, PA 16803, USA Email: marylyn.ritchie@psu.edu

Technology is driving the field of human genetics research with advances in techniques to generate high-throughput data that interrogate various levels of biological regulation. With this massive amount of data comes the important task of using powerful bioinformatics techniques to sift through the noise to find true signals that predict various human traits. A popular analytical method thus far has been the genome-wide association study (GWAS), which assesses the association of single nucleotide polymorphisms (SNPs) with the trait of interest. Unfortunately, GWAS has not been able to explain a substantial proportion of the estimated heritability for most complex traits. Due to the inherently complex nature of biology, this phenomenon could be a factor of the simplistic study design. A more powerful analysis may be a systems biology approach that integrates different types of data, or a meta-dimensional analysis. For this study we used the Analysis Tool for Heritable and Environmental Network Associations (ATHENA) to integrate high-throughput SNPs and gene expression variables (EVs) to predict high-density lipoprotein cholesterol (HDL-C) levels. We generated multivariable models that consisted of SNPs only, EVs only, and SNPs + EVs with testing r-squared values of 0.16, 0.11, and 0.18, respectively. Additionally, using just the SNPs and EVs from the best models, we generated a model with a testing r-squared of 0.32. A linear regression model with the same variables resulted in an adjusted r-squared of 0.23. With this systems biology approach, we were able to integrate different types of high-throughput data to generate meta-dimensional models that are predictive for the HDL-C in our data set. Additionally, our modeling method was able to capture more of the HDL-C variation than a linear regression model that included the same variables.

STATISTICAL EPISTASIS NETWORKS REDUCE THE COMPUTATIONAL COMPLEXITY OF SEARCHING THREE-LOCUS GENETIC MODELS

Ting Hu, Angeline S. Andrews, Margaret R. Karagas, and Jason H. Moore
Dartmouth College

The rapid development of sequencing technologies makes thousands to millions of genetic attributes available for testing associations with various biological traits. Searching this enormous high-dimensional data space imposes a great computational challenge in genome-wide association studies. We introduce a network-based approach to supervise the search for three-locus models of disease susceptibility. Such statistical epistasis networks (SEN) are built using strong pairwise epistatic interactions and provide a global interaction map to search for higher-order interactions by prioritizing genetic attributes clustered together in the networks. Applying this approach to a population-based bladder cancer dataset, we found a high susceptibility three-way model of genetic variations in DNA repair and immune regulation pathways, which holds great potential for studying the etiology of bladder cancer with further biological validations. We demonstrate that our SEN-supervised search is able to find a small subset of three-locus models with significantly high associations at a substantially reduced computational cost.

EVALUATION OF LINEAR CLASSIFIERS ON ARTICLES CONTAINING PHARAMACOKINETIC EVIDENCE OF DRUG-DRUG INTERACTIONS

A. Kolchinsky
School of Informatics and Computing, Indiana University

A. Lourenço
Institute for Biotechnology & Bioengineering, Centre of Biological Engineering, University of Minho Braga, Portugal

L. Li
Department of Medical and Molecular Genetics, Indiana Univeristy School of Medicine Indianapolis, IN, USA

L. M. Rocha
School of Informatics and Computing, Indiana University Bloomington, IN, USA

Background. Drug-drug interaction (DDI) is a major cause of morbidity and mortality. DDI research includes the study of different aspects of drug interactions, from in vitro pharmacology, which deals with drug interaction mechanisms, to pharmaco-epidemiology, which investigates the effects of DDI on drug efficacy and adverse drug reactions. Biomedical literature mining can aid both kinds of approaches by extracting relevant DDI signals from either the published literature or large clinical databases. However, though drug interaction is an ideal area for translational research, the inclusion of literature mining methodologies in DDI workflows is still very preliminary. One area that can benefit from literature mining is the automatic identification of a large number of potential DDIs, whose pharmacological mechanisms and clinical significance can then be studied via in vitro pharmacology and in populo pharmaco-epidemiology.

Experiments. We implemented a set of classifiers for identifying published articles relevant to experimental pharmacokinetic DDI evidence. These documents are important for identifying causal mechanisms behind putative drug-drug interactions, an important step in the extraction of large numbers of potential DDIs. We evaluate performance of several linear classifiers on PubMed abstractsand, separately, on sentences taken from PubMed abstracts, under different feature transformation and dimensionality reduction methods. In addition, we investigate the performance benefits of including various publicly-available named entity recognition features, as well as a set of internally-developed pharmacokinetic dictionaries.

Results. We found that several classifiers performed well in distinguishing relevant and irrelevant abstracts. We found that the combination of unigram and bigram textual features gave better performance than unigram features alone, and also that normalization transforms that adjusted for feature frequency and document length improved classification. .For some classifiers, such as linear discriminant analysis (LDA), proper dimensionality reduction had a large impact on performance. Finally, the inclusion of NER features and dictionaries was found not to help classification.

DETECTION OF PROTEIN CATALYTIC SITES IN THE BIOMEDICAL LITERATURE

Karin Verspoor, Andrew MacKinlay, Judith D. Cohn, Michael E. Wall

This paper explores the application of text mining to the problem of detecting protein functional sites in the biomedical literature, and specifically considers the task of identifying catalytic sites in that literature. We provide strong evidence for the need for text mining techniques that address residue-level protein function annotation through an analysis of two corpora in terms of their coverage of curated data sources. We also explore the viability of building a text-based classifier for identifying protein functional sites, identifying the low coverage of curated data sources and the potential ambiguity of information about protein functional sites as challenges that must be addressed. Nevertheless we produce a simple classifier that achieves a reasonable ~69% F-score on our full text silver corpus on the first attempt to address this classification task. The work has application in computational prediction of the functional significance of protein sites as well as in curation workflows for databases that capture this information.

ABERRANT PATHWAYS
ACCEPTED PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

FROM UNCERTAIN PROTEIN INTERACTION NETWORKS TO SIGNALING PATHWAYS THROUGH INTENSIVE COLOR CODING

Haitham Gabr, Alin Dobra, Tamer Kahveci

Discovering signaling pathways in protein interaction networks is a key ingredient in understanding how proteins carry out cellular functions. These interactions however can be uncertain events that may or may not take place depending on many factors including the internal factors, such as the size and abundance of the proteins, or the external factors, such as mutations, disorders and drug intake. In this paper, we consider the problem of finding causal orderings of nodes in such protein interaction networks to discover signaling pathways. We adopt color coding technique to address this problem. Color coding method may fail with some probability. By allowing it to run for sufficient time, however, its confidence in the optimality of the result can converge close to 100%. Our key contribution in this paper is elimination of the key conservative assumptions made by the traditional color coding methods while computing its success probability. We do this by carefully establishing the relationship between node colors, network topology and success probability. As a result our method converges to any confidence value much faster than the traditional methods. Thus, it is scalable to larger protein interaction networks and longer signaling pathways than existing methods. We demonstrate, both theoretically and experimentally that our method outperforms existing methods.

NEXT-GENERATION ANALYSIS OF CATARACTS: DETERMINING KNOWLEDGE DRIVEN GENE-GENE INTERACTIONS USING BIOFILTER, AND GENE-ENVIRONMENT INTERACTIONS USING THE PHENX TOOLKIT

Sarah A. Pendergrass
Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 503 Wartik Lab University Park, PA 16802, USA Email: sap29@psu.edu

Shefali S. Verma
Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab University Park, PA 16802, USA Email: szs14@psu.edu

Emily R. Holzinger
Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab University Park, PA 16802, USA Email: Emily.R.Holzinger@vanderbilt.edu

Carrie B. Moore
Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab University Park, PA 16802, USA Email: ccb12@psu.edu

John Wallace
Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab University Park, PA 16802, USA Email: sud23@psu.edu

Scott M. Dudek
Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab University Park, PA 16802, USA Email: sud23@psu.edu

Wayne Huggins
RTI International Research Triangle Park, NC,USA Email: whuggins@rti.org TERRIE KITCHNER Marshfield Clinic Marshfield, WI, USA Email: Kitchner.Terrie@mcrf.mfldclin.edu

Carol Waudby
Marshfield Clinic Marshfield, WI, USA Email: WAUDBY.CAROL@mcrf.mfldclin.edu

Richard Berg
Marshfield Clinic Marshfield, WI, USA Email: Berg.Richard@mcrf.mfldclin.edu

Catherine A. McCarty
Essential Rural Health Duluth, MN, USA Email: CMcCarty@eirh.org

Marylyn D. Ritchie
Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab University Park, PA 16802, USA Email: Marylyn.ritchie@psu.edu

Investigating the association between biobank derived genomic data and the information of linked electronic health records (EHRs) is an emerging area of research for dissecting the architecture of complex human traits, where cases and controls for study are defined through the use of electronic phenotyping algorithms deployed in large EHR systems. For our study, 2580 cataract cases and 1367 controls were identified within the Marshfield Personalized Medicine Research Project (PMRP) Biobank and linked EHR, which is a member of the NHGRI-funded electronic Medical Records and Genomics (eMERGE) Network. Our goal was to explore potential gene-gene and gene-environment interactions within these data for 529,431 single nucleotide polymorphisms (SNPs) with minor allele frequency > 1%, in order to explore higher level associations with cataract risk beyond investigations of single SNP-phenotype associations. To build our SNP-SNP interaction models we utilized a prior-knowledge driven filtering method called Biofilter to minimize the multiple testing burden of exploring the vast array of interaction models possible from our extensive number of SNPs. Using the Biofilter, we developed 57,376 prior-knowledge directed SNP-SNP models to test for association with cataract status. We selected models that required 6 sources of external domain knowledge. We identified 5 statistically significant models with an interaction term with p-value < 0.05, as well as an overall model with p-value < 0.05 associated with cataract status. We also conducted gene-environment interaction analyses for all GWAS SNPs and a set of environmental factors from the PhenX Toolkit: smoking, UV exposure, and alcohol use; these environmental factors have been previously associated with the formation of cataracts. We found a total of 288 models that exhibit an interaction term with a p-value ≤ 1x10-4 associated with cataract status. Our results show these approaches enable advanced searches for epistasis and gene-environment interactions beyond GWAS, and that the EHR based approach provides an additional source of data for seeking these advanced explanatory models of the etiology of complex disease/outcome such as cataracts.

COMPUTATIONAL DRUG REPOSITIONING
ACCEPTED PROCEEDINGS PAPER WITH POSTER PRESENTATION

PREDICTIVE SYSTEMS BIOLOGY APPROACH TO BROAD-SPECTRUM, HOST-DIRECTED DRUG TARGET DISCOVERY IN INFECTIOUS DISEASES

Ramon M. Felciano, Sina Bavari, Daniel R. Richards, Jean-Noel Billaud, Travis Warren, Rekha Panchal, Andreas Krämer

Knowledge of immune system and host-pathogen pathways can inform development of targeted therapies and molecular diagnostics based on a mechanistic understanding of disease pathogenesis and the host response. We investigated the feasibility of rapid target discovery for novel broad-spectrum molecular therapeutics through comprehensive systems biology modeling and analysis of pathogen and host-response pathways and mechanisms. We developed a system to identify and prioritize candidate host targets based on strength of mechanistic evidence characterizing the role of the target in pathogenesis and tractability desiderata that include optimal delivery of new indications through potential repurposing of existing compounds or therapeutics. Empirical validation of predicted targets in cellular and mouse model systems documented an effective target prediction rate of 34%, suggesting that such computational discovery approaches should be part of target discovery efforts in operational clinical or biodefense research initiatives. We describe our target discovery methodology, technical implementation, and experimental results. Our work demonstrates the potential for in silico pathway models to enable rapid, systematic identification and prioritization of novel targets against existing or emerging biological threats, thus accelerating drug discovery and medical countermeasures research.

EPIGENOMICS
ACCEPTED PROCEEDINGS PAPER WITH POSTER PRESENTATION

A POWERFUL STATISTICAL METHOD FOR IDENTIFYING DIFFERENTIALLY METHYLATED MARKERS IN COMPLEX DISEASES

Surin Ahn
Email: surin.ahn@gmail.com

Tao Wang †
Department of Epidemiology and Population Health, Albert Einstein College of Medicine of Yeshiva, 1300 Morris Park Ave, Bronx, NY 10461

DNA methylation is an important epigenetic modification that regulates transcriptional expression and plays an important role in complex diseases, such as cancer. Genome-wide methylation patterns have unique features and hence require the development of new analytic approaches. One important feature is that methylation levels in disease tissues often differ from those in normal tissues with respect to both average and variability. In this paper, we propose a new score test to identify methylation markers of disease. This approach simultaneously utilizes information from the first and second moments of methylation distribution to improve statistical efficiency. Because the proposed score test is derived from a generalized regression model, it can be used for analyzing both categorical and continuous disease phenotypes, and for adjusting for covariates. We evaluate the performance of the proposed method and compare it to other tests including the most commonlyused t-test through simulations. The simulation results show that the validity of the proposed method is robust to departures from the normal assumption of methylation levels and can be substantially more powerful than the t-test in the presence of heterogeneity of methylation variability between disease and normal tissues. We demonstrate our approach by analyzing the methylation dataset of an ovarian cancer study and identify novel methylation loci not identified by the t-test.

PERSONALIZED MEDICINE: FROM GENOTYPES AND MOLECULAR PHENOTYPES TOWARDS THERAPY

ACCEPTED PROCEEDINGS PAPER WITH POSTER PRESENTATION

A CORRELATED META-ANALYSIS STRATEGY FOR DATA MINING "OMIC" SCANS

Michael A Province
Division of Statistical Genomics, Washington University School of Medicine

Ingrid B Borecki
Division of Statistical Genomics, Washington University School of Medicine

Meta-analysis is becoming an increasingly popular and powerful tool to integrate findings across studies and OMIC dimensions. But there is the danger that hidden dependencies between putatively "independent" studies can cause inflation of type I error, due to reinforcement of the evidence from false-positive findings. We present here a simple method for conducting meta-analyses that automatically estimates the degree of any such non-independence between OMIC scans and corrects the inference for it, retaining the proper type I error structure. The method does not require the original data from the source studies, but operates only on summary analysis results from these in OMIC scans. The method is applicable in a wide variety of situations including combining GWAS and or sequencing scan results across studies with dependencies due to overlapping subjects, as well as to scans of correlated traits, in a meta-analysis scan for pleiotropic genetic effects. The method correctly detects which scans are actually independent in which case it yields the traditional meta-analysis, so it may safely be used in all cases, when there is even a suspicion of correlation amongst scans.

PHYLOGENOMICS AND POPULATION GENOMICS: MODELS, ALGORITHMS, AND ANALYTICAL TOOLS
ACCEPTED PROCEEDINGS PAPER WITH POSTER PRESENTATION

INFERRING OPTIMAL SPECIES TREES UNDER GENE DUPLICATION AND LOSS

M. S. Bayzid
Department of Computer Science, The University of Texas at Austin

S. Mirarab
Department of Computer Science, The University of Texas at Austin

T. Warnow
Department of Computer Science, The University of Texas at Austin

Species tree estimation from multiple markers is complicated by the fact that gene trees can differ from each other (and from the true species tree) due to several biological processes, one of which is gene duplication and loss. Local search heuristics for two NP-hard optimization problems - minimize gene duplications (MGD) and minimize gene duplications and losses (MGDL) - are popular techniques for estimating species trees in the presence of gene duplication and loss. In this paper, we present an alternative approach to solving MGD and MGDL from rooted gene trees. First, we characterize each tree in terms of its ``subtree-bipartitions" (a concept we introduce). Then we show that the MGD species tree is defined by a maximum weight clique in a vertex-weighted graph that can be computed from the subtree-bipartitions of the input gene trees, and the MGDL species tree is defined by a minimum weight clique in a similarly constructed graph. We also show that these optimal cliques can be found in polynomial time in the number of vertices of the graph using a dynamic programming algorithm (similar to that of Hallett and Lagergren), because of the special structure of the graphs. Finally, we show that a constrained version of these problems, where the subtree-bipartitions of the species tree are drawn from the subtree-bipartitions of the input gene trees, can be solved in time that is polynomial in the number of gene trees and taxa. We have implemented our dynamic programming algorithm in a publicly available software tool, available at http://www.cs.utexas.edu/users/phylo/software/dynadup/.

POST-NGS: ANALYSIS OF -OMES GENERATED BY NGS
ACCEPTED PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

USING BIOBIN TO EXPLORE RARE VARIANT POPULATION STRATIFICATION

Carrie Moore
Vanderbilt University

John Wallace
Pennsylvania State University

Alex T. Frase
Pennsylvania State University

Sarah A. Pendergrass
Pennsylvania State University

Marylyn D. Ritchie
Pennsylvania State University

Rare variants (RVs) will likely explain additional heritability of many common complex diseases; however, the natural frequencies of rare variation across and between human populations are largely unknown. We have developed a powerful, flexible collapsing method called BioBin that utilizes prior biological knowledge using multiple publicly available database sources to direct analyses. Variants can be collapsed according to functional regions, evolutionary conserved regions, regulatory regions, genes, and/or pathways without the need for external files. We conducted an extensive comparison of rare variant burden differences (MAF < 0.03) between two ancestry groups from 1000 Genomes Project data, Yoruba (YRI) and European descent (CEU) individuals. We found that 56.86% of gene bins, 72.73% of intergenic bins, 69.45% of pathway bins, 32.36% of ORegAnno annotated bins, and 9.10% of evolutionary conserved regions (shared with primates) have statistically significant differences in RV burden. Ongoing efforts include examining additional regional characteristics using regulatory regions and protein binding domains. Our results show interesting variant differences between two ancestral populations and demonstrate that population stratification is a pervasive concern for sequence analyses.

METASEQ: PRIVACY PRESERVING META-ANALYSIS OF SEQUENCING-BASED ASSOCIATION STUDIES

Angad Pal Singh, Samreen Zafer, Itsik Pe'er

Department of Computer Science, Columbia University, New York, NY - 10027.

Human genetics recently transitioned from GWAS to studies based on NGS data. For GWAS, small effects dictated large sample sizes, typically made possible through meta-analysis by exchanging summary statistics across consortia. NGS studies groupwise-test for association of multiple potentially-causal alleles along each gene. They are subject to similar power constraints and therefore likely to resort to meta-analysis as well. The problem arises when considering privacy of the genetic information during the data-exchange process. Many scoring schemes for NGS association rely on the frequency of each variant thus requiring the exchange of identity of the sequenced variant. As such variants are often rare, potentially revealing the identity of their carriers and jeopardizing privacy. We have thus developed MetaSeq, a protocol for meta-analysis of genomewide sequencing data by multiple collaborating parties, scoring association for rare variants pooled per gene across all parties. We tackle the challenge of tallying frequency counts of rare, sequenced alleles, for metaanalysis of sequencing data without disclosing the allele identity and counts, thereby protecting sample identity. This apparent paradoxical exchange of information is achieved through cryptographic means. The key idea is that parties encrypt identity of genes and variants. When they transfer information about frequency counts in cases and controls, the exchanged data does not convey the identity of a mutation and therefore does not expose carrier identity. The exchange relies on a 3rd party, trusted to follow the protocol although not trusted to learn about the raw data. We show applicability of this method to publicly available exome-sequencing data from multiple studies, simulating phenotypic information for powerful metaanalysis. The MetaSeq software is publicly available as open source.

TEXT AND DATA MINING FOR BIOMEDICAL DISCOVERY
ACCEPTED PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

ENABLING HIGH-THROUGHPUT GENOTYPE-PHENOTYPE ASSOCIATIONS IN THE EPIDEMIOLOGIC ARCHITECTURE FOR GENES LINKED TO ENVIRONMENT (EAGLE) PROJECT AS PART OF THE POPULATION ARCHITECTURE USING GENOMICS AND EPIDEMIOLOGY (PAGE) STUDY

William S. Bush*, Jonathan Boston*, Sarah A. Pendergrass, Logan Dumitrescu, Robert Goodloe, Kristin Brown-Gentry, Sarah Wilson, Bob McClellan Jr, Eric Torstenson, Melissa A. Basford, Kylee L. Spencer, Marylyn D. Ritchie, Dana C. Crawford

Genetic association studies have rapidly become a major tool for identifying the genetic basis of common human diseases. The advent of cost-effective genotyping coupled with large collections of samples linked to clinical outcomes and quantitative traits now make it possible to systematically characterize genotype-phenotype relationships in diverse populations and extensive datasets. To capitalize on these advancements, the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) project, as part of the collaborative Population Architecture using Genomics and Epidemiology (PAGE) study, accesses two collections: the National Health and Nutrition Examination Surveys (NHANES) and BioVU, Vanderbilt University's biorepository linked to de-identified electronic medical records. We describe herein the workflows for accessing and using the epidemiologic (NHANES) and clinical (BioVU) collections, where each workflow has been customized to reflect the content and data access limitations of each respective source. We also describe the process by which these data are generated, standardized, and shared for meta-analysis among the PAGE study sites. As a specific example of the use of BioVU, we describe the data mining efforts to define cases and controls for genetic association studies of common cancers in PAGE. Collectively, the efforts described here are a generalized outline for many of the successful approaches that can be used in the era of high-throughput genotype-phenotype associations for moving biomedical discovery forward to new frontiers of data generation and analysis.

INCORPORATING EXPERT TERMINOLOGY AND DISEASE RISK FACTORS INTO CONSUMER HEALTH VOCABULARIES

Michael Seedorff
University of Iowa 240 Schaeffer Hall, Iowa City, IA, USA Email: michael-seedorff@uiowa.edu

Kevin J. Peterson, BS
Department of Information Technology, Mayo Clinic 200 1st Street SW, Rochester, MN, USA Email: peterson.kevin@mayo.edu

Laurie A. Nelsen, MS
Mayo Clinic Global Business Solutions, Mayo Clinic 200 1st Street SW, Rochester, MN, USA Email: nelsen.laurie@mayo.edu

Cristian Cocos, PhD
Mayo Clinic Global Business Solutions, Mayo Clinic 200 1st Street SW, Rochester, MN, USA Email: cocos.cristian@mayo.edu

Jennifer B. McCormick, PhD, MPP
Department of General Internal Medicine, Mayo Clinic 200 1st Street SW, Rochester, MN, USA Email: mccormick.jb@mayo.edu

Christopher G. Chute, MD, DrPH
Department of Health Sciences Research, Mayo Clinic 200 1st Street SW, Rochester, MN, USA Email: chute@mayo.edu

Jyotishman Pathak, PhD
Department of Health Sciences Research, Mayo Clinic 200 1st Street SW, Rochester, MN, USA Email: pathak.jyotishman@mayo.edu

It is well-known that the general health information seeking lay-person, regardless of his/her education, cultural background, and economic status, is not as familiar with—or comfortable using—the technical terms commonly used by healthcare professionals. One of the primary reasons for this is due to the differences in perspectives and understanding of the vocabulary used by patients and providers even when referring to the same health concept. To bridge this "knowledge gap," consumer health vocabularies are presented as a solution. In this study, we introduce the Mayo Consumer Health Vocabulary (MCV)—a taxonomy of approximately 5,000 consumer health terms and concepts—and develop text-mining techniques to expand its coverage by integrating disease concepts (from UMLS) as well as non-genetic (from deCODEme) and genetic (from GeneWiki+ and PharmGKB) risk factors to diseases. These steps led to adding at least one synonym for 97% of MCV concepts with an average of 43 consumer friendly terms per concept. We were also able to associate risk factors to 38 common diseases, as well as establish 5,361 Disease:Gene pairings. The expanded MCV provides a robust resource for facilitating online health information searching and retrieval as well as building consumer-oriented healthcare applications.

WORKSHOP: COMPUTATIONAL BIOLOGY IN THE CLOUD: METHODS AND NEW INSIGHTS FROM COMPUTING AT SCALE

ACCEPTED WORKSHOP ABSTRACT

STORMSEQ: AN OPEN-SOURCE, USER-FRIENDLY PIPELINE FOR PROCESSING PERSONAL GENOMICS DATA IN THE CLOUD

Konrad J. Karczewski 1, 2, Guy Haskin Fernald 1,2, Alicia R. Martin 1,2, Michael Snyder 2, Nicholas P. Tatonetti 3, and Joel T. Dudley 4

1 Biomedical Informatics Training Program, Stanford University School of Medicine, Stanford, CA 94305
2 Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305
3 Department of Biomed-ical Informatics, Columbia University, New York, NY 10032
4 Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY 10029

The increasing public availability of personal complete genome sequencing data has ushered in an era of citizen genomics, where individuals are becoming interested in and active participants in exploring their own personal genetics. Direct-to-consumer and direct-to-physician genetic testing companies return results in the form of short reads and preliminary variant calls. However, software for performing read mapping and variant calling is constantly improving and individuals may prefer to customize and update variant calls. Such an analysis typically requires parallel computing resources, such as a large computing cluster as well as proficiency on the command-line. We present STORMSeq (Scalable Tools for Open-Source Read Mapping), a customizable and modular cloud computing solution that is usable by individuals without significant computing resources or extensive technical experience. We provide this open-access and open-source resource as a user-friendly interface in the Amazon Elastic Compute Cloud.

COMPUTATIONAL DRUG REPOSITIONING POSTERS

IDENTIFYING DRUGGABLE TARGETS BY PROTEIN MICROENVIRONMENTS MATCHING

Tianyun Liu and Russ B. Altman
Stanford University

The druggability of a target protein is its potential to be modulated by small, drug-like molecules. Druggability is an important criterion in the target selection phase of drug discovery. However, an effective standard for evaluating target druggability has not been established. We hypothesize that: (1) Known drug binding sites contain advantageous physicochemical properties for drug binding, or "druggable microenvironments", and (2) Given a new target, the presence of multiple druggable microenvironments is associated with a high likelihood of druggability. Accordingly, we developed a computational method, DrugFEATURE that evaluate target druggability by assessing the protein microenvironments in potential small molecule binding sites. We benchmarked DrugFEATURE by using two datasets. One dataset measures a target's druggability using NMR-based screening. DrugFEATURE correlates well with this metric. The second dataset is based on historical drug discovery outcomes. Using the DrugFEATURE score cutoff derived from the first, we are able to accurately discriminate druggable and difficult targets in the second. We inspected the druggable microenvironments identified by DrugFEATURE and observed associations between key chemical groups in drug molecules and particular microenvironments within the binding sites. Therefore, DrugFEATURE may provide useful insight for early stage drug discovery, by suggesting not only druggability, but also chemical fragments with high likelihood of binding. We further evaluated the druggability of ten transcription factors and identified novel druggable sites with implications for cancer therapy.
Computational drug repositioning

GENOMIC PREDICTORS OF SENSITIVITY TO 17-AAG TREATMENT FOR INDIVIDUAL BREAST CANCER PATIENTS

Stephen R. Piccolo
Department of Pharmacology & Toxicology, University of Utah

Adam L. Cohen
Division of Oncology, Department of Internal Medicine, Huntsman Cancer Institute, University of Utah

W. Evan Johnson
Division of Computational Biomedicine, Boston University School of Medicine

Philip J. Moos
Department of Pharmacology & Toxicology, University of Utah

Andrea H. Bild
Department of Pharmacology & Toxicology, University of Utah

17-allylamino-17-demethoxygeldanamycin (17-AAG) is a derivative of the naturally occurring antibiotic geldanamycin. 17-AAG has been shown to inhibit cellular growth in vitro through binding to heat shock protein 90, which is believed to be essential for malignancy; thus 17-AAG is currently being tested as an antitumor agent. Optimal repositioning of 17-AAG requires a detailed assessment of the genomic profiles that underlie tumor responsiveness. In early clinical trials, clinical efficacy has been moderate, suggesting that 17-AAG treatment is suitable for only a subset of tumors. In this study, we demonstrate a systems-biology approach for targeting this drug at patients most likely to respond to treatment. We explored treatment effects on a series of breast-cancer cell lines and observed that cell lines with HER-2 amplification were significantly more sensitive to 17-AAG treatment—; this finding confirms previous studies. We also observed that tumors of the “"basal" gene-expression subtype were least sensitive to treatment. Upon extending this analysis to a series of data sets representing primary breast tumors, we observed similar associations with genomic subtypes, but the connection between basal tumors and resistance was considerably more pronounced. An evaluation of mutation status revealed a pattern that sheds light on a potential mechanism behind sensitivity/resistance: mutations in growth-factor-receptor genes were highly prevalent in 17-AAG sensitive lines, whereas mutations within the Wnt/Beta-catenin signaling pathway were more common in resistant lines. Our approach constitutes a systematic way to associate individuals with treatment response. We believe this method can be extended to various other drugs and cancer types and thus aid researchers in their efforts to reposition drugs in a way that can be tailored to individual patients.

ANALYSIS OF ANTIBACTERIAL AND TUMOR HOMING PEPTIDES USING SUPPORT VECTOR MACHINE

THAMMAKORN SAETHANG
Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa, Japan
Email: thammakorn.kmutt@gmail.com

OSAMU HIROSE
Institute of Science and Engineering, Kanazawa University, Kanazawa, Japan
Email: hirose@se.kanazawa-u.ac.jp

INGORN KIMKONG
Department of Microbiology, Faculty of Science, Kasetsart University, Bangkok, Thailand Email: fsciiok@ku.ac.th

KENJI SATOU
Institute of Science and Engineering, Kanazawa University, Kanazawa, Japan
Email: ken@t.kanazawa-u.ac.jp

In the last few decades, bacterial resistance to antibiotics has become a global public health problem. The development for a novel class of antibiotics is required. Antimicrobial peptides (AMPs), which are small cationic peptides with a broad antimicrobial activity, are one of the most promising alternatives. There have been reports that the use of AMPs is effective and safe. AMPs also have anti-tumor activity. This marvelous property of AMPs encourages researchers to study tumor homing peptides (THPs). The THP is a crucial part of drug delivery vehicle that guides therapeutic agents to eliminate specific tumor cells without destroying normal cells. In this study, we developed a new computational method for the prediction of AMPs and THPs by using our ensemble approach. We combined a number of peptide-coding schemes and used the support vector machine (SVM) as the classifier. Our method significantly outperformed the previous AMPs predictor, AntiBP2. When 5-fold cross validation was carried out on the processed dataset of AMPs, accuracy (ACC), Matthew's correlation coefficient (MCC), and specificity of our method were 2.1%, 4.3%, and 3.2% higher than those of AntiBP2, respectively. In addition, our method achieved high performance when testing on the dataset of THPs (ACC = 0.937, MCC = 0.875, sensitivity = 0.942, and specificity = 0.932). We speculate that our method may be useful in identifying AMPs and THPs for the development of new therapeutics.
Keywords: antimicrobial peptides, tumor homing peptides, machine learning, and antibiotics resistance

EPIGENOMICS POSTERS

A POWERFUL STATISTICAL METHOD FOR IDENTIFYING DIFFERENTIALLY METHYLATED MARKERS IN COMPLEX DISEASES

SURIN AHN and TAO WANG

Department of Epidemiology and Population Health
Albert Einstein College of Medicine of Yeshiva
1300 Morris Park Ave, Bronx, NY 10461

DNA methylation is an important epigenetic modification that regulates transcriptional expression and plays an important role in complex diseases, such as cancer. Genome-wide methylation patterns have unique features and hence require the development of new analytic approaches. One important feature is that methylation levels in disease tissues often differ from those in normal tissues with respect to both average and variability. In this paper, we propose a new score test to identify methylation markers of disease. This approach simultaneously utilizes information from the first and second moments of methylation distribution to improve statistical efficiency. Because the proposed score test is derived from a generalized regression model, it can be used for analyzing both categorical and continuous disease phenotypes, and for adjusting for covariates. We evaluate the performance of the proposed method and compare it to other tests including the most commonly-used t-test through simulations. The simulation results show that the validity of the proposed method is robust to departures from the normal assumption of methylation levels and can be substantially more powerful than the t-test in the presence of heterogeneity of methylation variability between disease and normal tissues. We demonstrate our approach by analyzing the methylation dataset of an ovarian cancer study and identify novel methylation loci not identified by the t-test.

USING DNASE DIGESTION DATA TO ACCURATELY IDENTIFY TRANSCRIPTION FACTOR BINDING SITES

Kaixuan Luo and Alexander J. Hartemink

Program in Computational Biology and Bioinformatics, and Department of Computer Science, Duke University, Durham, NC 27708, USA

Identifying binding sites of transcription factors (TFs) is a key task in deciphering transcriptional regulation. ChIP-based methods are used to survey the genomic locations of a single TF in each experiment. But methods combining DNase digestion data with TF binding specificity information could potentially be used to survey the locations of many TFs in the same experiment, provided such methods permit reasonable levels of sensitivity and specificity. Here, we present a simple such method that outperforms a leading recent method, CENTIPEDE, marginally in human but dramatically in yeast (average auROC across 20 TFs increases from 74% to 94%). Our method is based on logistic regression and thus benefits from supervision, but we show that partially and completely unsupervised variants perform nearly as well. Because the number of parameters in our method is at least an order of magnitude smaller than CENTIPEDE, we dub it MILLIPEDE.

SAAP-BS: STREAMLINED ANALYSIS AND ANNOTATION PIPELINE FOR BISULFITE SEQUENCING

Zhifu Sun, Saurabh Baheti, Sumit Middha, Rahul Kanwar, Andreas S. Beutler and Jean-Pierre A. Kocher

Mayo Clinic College of Medicine, Rochester, MN 55905

Epigenetic modification of histones and DNA adds heritable information to the genome. Bisulfite Sequencing (BS) is commonly used to decipher genome-wide methylation patterns and allows researchers to study and identify regions which epigenetically regulate gene expression. Analyzing BS sequencing data is challenging and specialized alignment/mapping programs are needed. Although such programs have been developed, a comprehensive solution that provides researchers with good quality and analyzable data is still lacking. Here, we present Streamlined Analysis and Annotation Pipeline for BS data (SAAP-BS) that integrates read quality assessment/clean-up, quality metrics, alignment, methylation data extraction, variant detection, annotation, reporting, and visualization. The pipeline works with the most commonly used RRBS (Reduced Representation Bisulfite Sequencing), WGS (Whole Genome Bisulfite Sequencig), or Agilent SureSelect Methyl-Seq either by single end or pair end sequencing. It is based on a MapReduce approach and runs in a parallel computational environment, making it highly efficient and scalable. For maximum flexibility the whole pipeline is made modular to allow the users to initiate it with various universally accepted input data types (Fastq, BAM, or a tab-separated list of CpG positions). The simultaneous detection of CpG methylation and underlying single nucleotide variants maximizes the usage of the sequence data and allows immediate integration of methylation and genomic variation. This package facilitates a rapid transition from sequencing reads to a fully annotated CpG methylation report to biological interpretation.

GENERAL POSTERS

CONTACTS-ASSISTED PROTEIN STRUCTURE PREDICTION

Badri Adhikari 1, Xin Deng 1, Jilong Li 1, Debswapna Bhattacharya 1, and
Jianlin Cheng 1,2,3

1 Department of Computer Science, 2 Informatics Institute and
3 C. Bond Life Science Center,
University of Missouri, Columbia, MO 65211, USA.

One recent approach for protein tertiary structure prediction from its residue sequence is to predict which residues are close to each other first, and then build a complete structure solely from this contacts information. These methods use a threshold distance to define this closeness or residue-residue contact. Instead of building a structure purely from these contacts, we summarize a contact-assisted structure prediction approach that uses only a few known contacts to improve the quality of already predicted models. Assuming that we already have some predicted structures and some known contacts, we designed and implemented an automated pipeline that starts with a predicted structure and improves the structure using the inputted contacts as constraints. The system also handles cases when some non-contact information (i.e., knowledge that two residues are not in contact) is provided as input along with or instead of contact information. Our approach for contact assisted structure prediction is a model selection and improvement process comprising of three major steps. First, we select models from a predicted model pool using a scoring scheme. We then refine these selected models using existing protein refinement tools. Finally, we improve the structure of these refined models with given residue-residue contacts information as distance restraints. Our experiment during the 10th Critical Assessment of Techniques for Protein Structure Prediction (CASP10) in 2012 shows that in most cases the quality of predicted structures is improved. The server for contact-assisted protein structure prediction is available at:
http://protein.rnet.missouri.edu/contact_assisted/index.html

THE EVOLUTIONARY LANDSCAPE OF ALTERNATIVE SPLICING IN VERTEBRATE SPECIES

Nuno L. Barbosa-Morais 1,2 Manuel Irimia 1 Qun Pan 1 Hui Y. Xiong 3 Serge Gueroussov 1,4 Leo J. Lee 3 Valentina Slobodeniuc 1 Claudia Kutter 5,6 Stephen Watt 5 Recep Çolak 1,7 TaeHyung Kim 1,7 Christine M. Misquitta-Ali 1 Michael D. Wilson 4,5,8 Philip M. Kim 1,4,7 Duncan T. Odom 5,6,9 Brendan J. Frey 1,3 Benjamin J. Blencowe 1,4

1 Banting and Best Department of Medical Research, Donnelly Centre, University of Toronto, Ontario, Canada
2 Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Portugal
3 Department of Electrical and Computer Engineering, University of Toronto, Canada
4 Department of Molecular Genetics, University of Toronto, Ontario, M5S 1A8, Canada 5 Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, UK
6 Department of Oncology, Hutchison/MRC Research Centre, Hills Road, Cambridge, CB2 0XZ, UK
7 Department of Computer Science, University of Toronto, Ontario, M5S 2E4, Canada
8 Hospital for Sick Children, Toronto, Ontario, M5G 1X8, Canada
9 Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

How species with similar repertoires of protein coding genes differ so dramatically at the phenotypic level is poorly understood. Alternative splicing (AS) has been proposed to play an important role in phenotypic differences, because it is a widespread process by which diverse mRNA and protein isoforms can be produced from individual genes. However, to date, evolutionary differences in AS complexity, regulation and function between vertebrate species have not been systematically investigated. Accordingly, in this study, we describe the first genome-wide investigation of AS differences among multiple physiologically equivalent organs from several vertebrate species spanning the major tetrapod lineages. High-throughput RNA sequencing (RNA-Seq) data were collected from seven tissues from ten vertebrate species. For each species, we considered all annotated internal exons as potential cassette type AS events and created a non-redundant database of splice junction sequences formed by inclusion or skipping of each exon. RNA-Seq reads were mapped to the junction databases to determine inclusion levels for each exon in each tissue and species. Orthology relationships between individual exons were established to allow for direct cross-species comparisons. From comparing the transcriptomes of multiple organs from vertebrate species spanning ~350 million years of evolution, we observe significant increases in AS complexity that are associated with proximity to the primate lineage. Moreover, in species separated by at least six million years, the AS profiles of physiologically equivalent organs have diverged to the extent that they are more strongly related to the identity of a species than to organ type. These species-dependent AS patterns are controlled by a largely conserved cis-regulatory code, together with specific changes in trans-acting factors. In particular, we define species-classifying AS events that are predicted to remodel protein-protein interactions involved in gene regulation and other processes. These species-specific AS events presumably evolved to reconfigure interactions involving regulatory complexes, rather than to change their binding specificity. These changes likely evolved in conjunction with the extensive shuffling of largely common code elements to drive the remarkable diversification in AS and other transcriptomic changes that underlie fundamental differences between vertebrate species.

REFINEPRO: A NOVEL CONFORMATION ENSEMBLE APPROACH TO PROTEIN STRUCTURE REFINEMENT

Debswapna Bhattacharya 1 and Jianlin Cheng 1,2,3

1 Department of Computer Science, 2 Informatics Institute and 3 C. Bond Life Science Center, University of Missouri, Columbia, MO 65211, USA.

Despite having significant advancement of computational methods in protein structure prediction during the last decade, these techniques often fail to achieve allowable prediction accuracy to be applied in solving biological problems. Bringing these low-resolution predicted models to high-resolution structures close to their native state, called the protein structure refinement problem, however, has proven to be extremely challenging and a largely unsolved problem in the field of protein structure prediction. Here, we propose a novel conformation ensemble approach to protein structure refinement, called REFINEpro aiming to improve the overall fold of the starting model together with enhancement in the general physicality. Given a target protein structure and a model ensemble consisting of numerous structures generated for the same target, the method first identifies the less conserved local regions in the initial model by consensus approach. We call these regions problematic regions (PRs). An iterative refinement is then applied in a greedy manner for each PR via generation of hybrid models by assembling better-modeled fragments corresponding to these PRs from structures in the ensemble followed by quality assessment to select the best hybrid model. Finally, atomic level energy minimization is performed on the best hybrid model to optimize the hydrogen-bonding network and to improve the local qualities in order to produce the refined structure. By performing a large-scale benchmark study on 177 proteins comprising of targets from recent Critical Assessment of Techniques for Protein Structure Prediction (CASP) and models generated through Molecular Dynamics (MD) simulation, it has been demonstrated that the protocol is capable of consistently and simultaneously improving both global and local qualities of protein models generated by both template based and ab-initio methods. To our knowledge, a fully automated ensemble based approach has not been used before in refinement problem. Also, the ability of REFINEpro to often drastically improve the overall fold of the initial models through refinement of loop and terminal regions or rearrangements of disoriented secondary structure segments, accompanied by correction of local errors makes it distinctly different from a “conservative” local sampling strategy producing improvement only in the physicality of the models instead of improvement of the global positioning of the backbone atoms. The REFINEpro web server is freely available at http://sysbio.rnet.missouri.edu/REFINEpro/.

APERTURE: A WEB-BASED TOOL FOR VISUALIZING THE REGULATORY POTENTIAL OF SNPS

WILLIAM S. BUSH

Department of Biomedical Informatics, Center for Human Genetics Research, Vanderbilt University, 2215 Garland Avenue, 519 Light Hall Nashville, TN 37232, USA Email: william.s.bush@vanderbilt.edu

Numerous studies have evaluated the effects of genetic variation in cis on gene expression. The results of these studies have great potential for translating disease associated SNPs into functional mechanisms, but tools to access the results of these studies are limited. We developed a standalone tool and website called Aperture to provide LocusZoom-style plots for eQTL associations in the context of existing GWAS phenotype associations and tissue-specific chromatin states. Aperture provides access to 28 million eQTL results. Aperture augments GWAS results by providing simul-taneous visualization of linkage disequilibrium patterns, predicted chromatin states, eQTL associations, and known disease associations, refining GWAS hits to a collection of SNPs with known functional effects.

PROTEIN MODEL QUALITY PREDICTION BY MULTICOM SERVERS

Renzhi Cao 1, Zheng Wang 1, Jilong Li 2, Jianlin Cheng 1,2,3

1Computer Science Department; 2Informatics Institute; 3 C. Bond Life Science Center, University of Missouri, Columbia, MO 65211, USA

Ranking predicted protein structures (models) based on the quality without knowing the native protein structure is of great importance for protein structure prediction. In protein structure prediction, quality assessment (QA) is used to evaluate the accuracy of a protein model. We developed and tested four model quality assessment servers: MULTICOM-REFINE, MULTICOM-CLUSTER, MULTICOM-NOVEL, MULTICOM-CONSTRUCT. All of the four QA servers can generate both global quality scores and local quality scores. The global quality score is the quality score for each protein model and the local quality score is the quality score for each residue in protein model. For MULTICOM-REFINE, the global quality score is generated by the pair-wise method, which calculates the pair-wise GDT-TS score of a pool of models and then evaluate the quality of each model; the local quality score is generated by the refined local scores, which selected a reference model set by global quality score and compared the model with each model in the reference set. MULTICOM-CLUSTER and MULTICOM-NOVEL are single-model, support vector machine (SVM)-based method. The SVM was trained to predict the local quality score of each residue and the global quality scores were generated from the local quality scores. The input features for MULTICOM-CLUSTER includes amino acids encoded by 20-digit vector of 0 and 1, the difference between secondary structure and solvent accessibility predicted from protein sequence and parsed from the model itself, and predicted contact probabilities. The feature used in MULTICOM-NOVEL is the same as MULTICOM-CLUSTER except that amino acid sequence features were replaced with the sequence profile features. The global quality scores of MULTICOM-CONSTRUCT are generated by refined pair-wise model comparison method. The local quality scores are generated by using SVM method, which uses secondary structure difference, solvent accessibility, profiles, and SOV score as the features for SVM. Our servers are available at the following website: http://sysbio.rnet.missouri.edu/QApro.html. The software packages are available at: http://sysbio.rnet.missouri.edu/multicom_toolbox/.

UNDERSTANDING TRANSCRIPTIONAL REGULATION BY CHIP-SEQ DATA ANALYSIS

Chao Cheng

1. Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire 03755, United States of America

2. Institute for Quantitative Biomedical Sciences, Norris Cotton Cancer Center, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire 03766, United States of America

Chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq) has been widely used to determine binding sites of transcription factors (TFs) and to investigate histone modifications in a genome wide scale. By integrating ChIP-seq data from the ENCODE project with other data sources such as expression data and TF binding motif information, we have developed new methods to identify TF target genes, to predict new human enhancers, to relate TF binding and histone modifications with gene expression, and to construct integrated human regulatory network. First, given the genomic binding data for a TF, target genes are often identified as those with at least one binding peaks in their promoter regions. There are several limitations associated with this peak-based method. We thus propose a probabilistic model called target identification from profiles (TIP) that quantitatively measures the regulatory relationships between TFs and target genes. For a TF, the model builds a characteristic, averaged profile of binding around the transcription start site (TSS) and then uses this to weight the sites associated with a given gene, providing a continuous-valued score relating the TF and potential target. Second, to predict novel human enhancers we develop a model that integrates chromatin features, TF binding data with sequence analysis. The model identifies a list of highly conserved DNA regions with active enhancer histone marks and enriched TF binding motifs. Experimental validation of the predicted enhancers in mouse embryos has shown a high accuracy of our model. Third, we develop statistical models to quantify the relationship between TF binding/histone modification signals with gene expression levels. Our results indicate that more than 60% of variation of gene expression can be explained by TF binding or histone modification signals around the TSS of genes. We show that TF and HM models are cell line specific: TF binding and HM signals are more predictive of gene expression in the same cell line. Moreover, the models trained solely on protein-coding genes are also predictive of expression levels of microRNAs (miRNAs), suggesting that their regulation by TFs and HMs shares a similar mechanism to that for protein-coding genes. Fourth, we present a network framework for analyzing multi-level regulation in human and worm by integrating various high-throughput datasets. The integrated network includes TF-gene, TF-miRNA, miRNA-gene regulatory interactions and further incorporates protein-protein interactions. We investigate the hierarchical structure of the network and find that TFs at different levels display significantly different features, e.g. lower-level TFs are more uniformly expressed at various tissues and tend to have more interacting partners. Network motif analysis suggests an over-representation of some notable network motifs, e.g the miRNA mediated feed-forward loop. Our methods provide useful tools to analyze ChIP-seq data for better understanding transcriptional regulation of genes.

USING SINGLE-CELL RNA-SEQ TO UNDERSTAND EARLY EMBRYO DEVELOPMENT

H. Rosaria Chiang 1, Shawn Chavez 1, Wing Wong 2, Renee A. Reijo Pera 1

1 Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, CA 94305
2 Department of Statistics, Stanford University, Stanford, CA 94305

Due to inherent difficulties in studying early human embryo development, most data regarding mammalian embryogenesis have been derived from studies in mice. While many aspects of human embryo development are conserved in mice, some of the genetic and epigenetic programs that culminate in activation of the embryonic genome differ. In addition, while 80-90% of mouse embryos develop into the blastocyst stage in vitro, only 30-50% of human embryos reach the blastocyst stage. Thus, our lab has previously identified non-invasive means to identify embryos with the potential to develop to the blastocyst stage based on the timing of the first three mitotic divisions. We are working toward profiling and comparing gene expressions of the blastomeres that are predicted to develop to the blastocyst stage and those of the blastomeres predicted to arrest. Here we describe progress towards using single-cell RNA-seq to achieve this goal.

NOVEL MODELING OF COMBINATORIAL MIRNA TARGETING IDENTIFIES SNP WITH POTENTIAL ROLE IN BONE DENSITY

Claudia Coronnello, Ryan Hartmaier, Arshi Arora, Luai Huleihel, Kusum V. Pandit, Abha S. Bais, Michael Butterworth, Naftali Kaminski, Gary D. Stormo, Steffi Oesterreich, Panayiotis V. Benos

MicroRNA genes (miRNAs) are small non-coding RNAs that regulate the expression levels of mRNAs post-transcriptionally. miRNAs are critical in many important biological processes, like development, and are important markers for many diseases. Identifying the targets of miRNAs is not an easy task. Recent developments of high-throughput data collection methods for identification of all miRNA targets in a cell are promising, but they still depend on computational algorithms to identify the exact miRNA:mRNA interactions. We present a novel algorithm, ComiR (Combinatorial miRNA targeting), which addresses a more general question, that is, whether a given mRNA is targeted by a set of miRNAs. ComiR uses miRNA expression to improve the targeting models of four target prediction algorithms. Then it combines their predicted targets using a support vector machine. By applying ComiR to single nucleotide polymorphism (SNP) data, we identified a SNP that is likely to be causally associated to osteoporosis in women. ComiR web tool is available at http://www.benoslab.pitt.edu/comir/. It generates custom predictions for H.sapiens, M.musculus, D.melanogaster and C.elegans species miRNAs.

ESTIMATING THE MOLECULAR COMPLEXITY OF SEQUENCING LIBRARIES

Timothy Daley (1) and Andrew Smith (2)

(1) Department of Mathematics and (2) Department of Molecular and Computational Biology, University of Southern California

Modern DNA sequencing experiments often interrogate hundreds of millions, up to billions of distinct molecules, possibly to achieve deep coverage genome-wide, or to deeply interrogate a specific biological sample. Low complexity DNA sequencing libraries are problematic in such experiments: much of the sequencing examines the same original molecules, and continued sequencing either provides redundant data that is discarded, or introduces biases in downstream analyses. Often investigators are presented with the decision to sequence an existing library more deeply, or to generate another library. Investigators may be presented with a choice between multiple libraries to sequence deeply, and must decide based on a preliminary ``shallow'' survey which one to pursue. The central analysis question underlying these decisions is what proportion of the sequenced reads will correspond to yet unsequenced molecules.

We describe an empirical Bayesian method for predicting the yield of previously unsequenced molecules from deeper sequencing of the same library. Our method is based on statistics derived in the context of capture-recapture experiments, but have so far been impractical in applications involving far-ranging predictions or very large data sets. We apply rational function approximations to obtain estimates that are accurate even when extrapolating more than 100-fold beyond the initial sample. In the context of sequencing experiments, these estimates allow data from sequencing a few million reads to very accurately predict the yield for sequencing several hundred million reads from the same library.

PREDICTING PROTEIN MODEL QUALITY FROM SEQUENCE ALIGNMENT BY SUPPORT VECTOR MACHINES

Xin Deng, Jianlin Cheng

Computer Science Department, University of Missouri, Columbia, MO 65211, USA

Protein sequence alignment is essential for homology-based protein structure modeling. Here, we developed a SVM (Support Vector Machine) alignment-based model selection method to predict the quality score (GDT-TS score) of a protein structure model from the features extracted from the query-template alignment used to generate the model. The input features fed into the SVM predictor include the normalized e-value of the given query-template alignment, the percentage of identical residue pairs of alignment positions, the percentage of residues of the query aligned with one in the template, and the sum of the BLOSUM scores of all aligned residues divided by the length of the aligned positions. The four input features were extracted from 482 pairwise sequence alignments generated from the CASP9 (Critical Assessment of Techniques for Protein Structure Prediction) datasets by PSI-BLAST along with the real GDT-TS scores calculated by the TM-Score program. In the training process, a SVM regression predictor was trained based on the above data to predict the GDT-TS scores of the models from the input features. Three parameters (the epsilon width of the regression tube w; the margin option c; the gamma in the RBF kernel g) were tuned during training the SVM with Gaussian radial basis kernel (RBF) regression model. The root mean square error (RMSE) and the absolute mean error (ABS) between predicted and real GDT-TS scores were calculated to assess the performance. A five-fold cross validation was adopted to select the best parameter sets based on the average RMSE and ABS on the five folds. The RMSE and ABS of the SVM trained with the best parameter values were 0.085 and 0.06, respectively. We compared the SVM alignment-based predictor with the pure e-value based method. The better performance of SVM predictor indicates that integrating sequence alignment features with a SVM can improve model selection over the pure e-value based method.

GENOME-WIDE ANALYSIS OF CORRELATION BETWEEN ECONOMIC FLUCTUATION AND IMMUNITY LEVEL

Docyong Kim, Kyunghyun Park, Doheon Lee

KAIST

Economic fluctuation is the one of reasons of human's psychological stress in contemporary capitalism. The psychological stress affect to human health. An association between the stress and level of immunity was well studied by previous research. A change of immunity level could be influenced by external psychological stress. This phenomenon could be occurred by genetic factor of each person. Hence, many studies about a relationship between environmental factor and genetic factor, namely gene-environment interaction, have been performed recently. Our hypothesis is that economic fluctuation can influence to individual health by genetic variations. In this study, therefore, stock market index was utilized for economic fluctuation. In addition, white blood cell count used as individual health level. These data were gathered from two data resource. One of them was Korea Composite Stock Price Index (KOSPI) for economic fluctuation for 6 years (2001~2006). The other was Korea Association Research and Examination (KARE) in which Single Nucleotide Polymorphism (SNP) and White Blood cell Count (WBC) in same period. At the first step, we attempted calculating correlation score and statistical validation between KOSPI and WBC of total data. In this step, there was a comparison between WBC of each person and KOSPI daily fluctuation level. The comparison has low correlation score. The next step was to identify candidate SNPs utilizing correlation score's rank by genotype of SNP. We calculated correlation of stock market daily fluctuation in same date of WBC examination after collecting WBC of person who had a specific SNP. At the result, p-value of 1,115 SNPs was satisfied with a significant value (0.05). Gene sets which mapped from extracted SNPs were tested gene enrichment using the Database for Annotation, Visualization and Integrated Discovery (DAVID). A part of 752 DAVID IDs which were mapped from 1,115 SNPs were related with pituitary and whole blood in tissue DB, and cell adhesion molecules in pathway DB. In this study, we discovered a new gene set which was related with economic fluctuation. Meta-analysis of the result can confirm genes related with immunity and psychological stress. We will attempt to validating and investigating a mechanism of environmental factor (economic fluctuation), brain (psychological stress), and human health (immune system) in further study.

SOLVENT ACCESSIBILIY AND AMINO ACID PROPENSITIES AROUND GLYCOSYLATION SITES DEPENDING ON THE KINDS OF OLIGOSACCHARIDES

Kenji Etchuya, Hirotaka Tanaka, Takeo Terasaki, Takanori Sasaki, Yuri Mukai

Meiji University

In recent years, carbohydrate chains have been considered “the third chain” in life science. Many molecules including glycoproteins and glycolipids are glycosylated in vivo. Glycosylation can maintain important life activities such as stability, solubility, signal secretion and the regulation of interactions. Glycosyltransferases recognize specific motif sequences and modify oligosaccharides. In glycation called "N-glycosylation" which includes Asn-X-Thr/Ser motifs, Asn residues are modified. In "O-glycosylation", Thr/Ser residues are modified. "C-glycosylation" has Trp-X-X-Trp motifs, and the first Trp residues are modified. Many proteins are glycosylated in the endoplasmic reticulum and the Golgi apparatus and transported outside cells. Functions of oligosaccharides in O-glycosylation are as follows; Fucose (Fuc) and glucose (Glc) were shown to be essential in the activation of Notch protein in differentiation system by experimental studies. In glycosaminoglycan (GAG), the extension by connecting tetra-saccharides (Xylose-Galactose-Galactose-Glucuronic acid) can be observed. In addition, the first xylose (Xyl) becomes the starting point of other GAG extensions and has functions related to signal transduction systems. Proteins modified by N-acetyl galactosamine (GalNAc) exist on the surface of cells and preserve cell protective function due to high viscosity by holding water molecules. Therefore, N-acetyl glucosamine (GlcNAc) exists in the nucleus and cytoplasm unlike other oligosaccharides. In the nucleus, in which transcription and translation take place, phosphorylation controls signal transmission and the enzymatic response. GlcNAc inhibits phosphorylation by the structural change of signal transmission related enzymes and receptors. Therefore GlcNAc is disintegrated in cytoplasm quickly by an existing enzyme without being extended if it becomes unnecessary. Mannose (Man) exists mainly in the brain and accounts for one-third of the glycosylated proteins. Therefore, physicochemical properties of the amino acid sequences around the modification sites depending on kinds of oligosaccharides. In this study, to find the specificity in the sequences around glycosylation sites of each kind of O-Glycosylation, the characteristics were extracted around glycosylated positions. At first, the sequences around the glycosylated positions were obtained from Uniprot KnowledgeBase / Swiss-Prot Release 2012_06 and classified into different kinds of oligosaccharides. After this, the sequences were assigned by position-specific scores and evaluated by five-fold cross-validation with various calculation domains. The optimized calculation domains reflected oligosaccharide type-dependent characteristics and domain lengths. As a result, a correlation between the accuracies and the propensities of hydrophilic and hydrophobic residues could be found. Furthermore, the necessity of solvent accessibility and charged residues around glycosylation sites was confirmed. Solvent accessibility, surface charges, protein tertiary structure and the characteristics of glycosyltransferases are estimated based on amino acid sequences around glycosylation positions.

GENESEER: A FLEXIBLE, EASY-TO-USE TOOL TO AID DRUG DISCOVERY BY EXPLORING EVOLUTIONARY RELATIONSHIPS BETWEEN GENES ACROSS GENOMES

Douglas D. Fenger, Matthew Shaw, Philip Cheung, Tim Tully

Dart NeuroScience, San Diego, CA

Homologous relationships are useful in drug discovery because they facilitate the mapping of gene/protein function between and within species, allowing functional predictions of novel or unknown genes. Early knowledge of a gene's paralogous family is also important when designing the safety screens associated with a drug discovery project. If the target has related paralogs, it is important that the drug discovery team understand the potential effects of an off target interaction. A drug discovery program might want to include a selectivity screen as part of their assay cascade to ensure that their drug is not binding close paralogs, resulting in undesirable off target effects. Now that genomic sequences for many species are readily available, bioinformatic algorithms can perform entire genome comparisons to identify these same relationships.

GeneSeer (http://geneseer.com) is a publicly available tool that leverages public sequence data, gene metadata information, and other publicly available data to calculate and display orthologous and paralogous gene relationships for all genes from several species, including yeasts, insects, worms, vertebrates, mammals, and primates such as human. GeneSeer calculates homology relationships by performing a full proteome BLAST calculation between two species. The GeneSeer interface is designed to help scientists quickly predict important drug discovery attributes such as selectivity and safety. It is a useful tool for cross-species translational mapping and enables scientists to easily translate hypotheses about gene identity and function from one species to another.

Besides describing GeneSeer's underlying methods and user-friendly interface, we will also present a validation study of GeneSeer versus Homologene, the homolog prediction tool from NCBI. The underlying scientific data for GeneSeer has been validated to be as good as, if not better than, Homologene. Finally, a comparison of features shows GeneSeer to be the most feature rich when compared to alternative orthologing tools.

ROBUST AND FAST LINEAR MIXED MODELS FOR GENOME-WIDE ASSOCIATION STUDIES WITH CONFOUNDING

David Heckerman, Christoph Lippert, Jennifer Listgarten

Microsoft Research, Los Angeles, CA, USA

Linear Mixed Models (LMMs) tackle confounding in genome-wide association studies (GWAS) by using a matrix of pairwise genetic similarities to model relatedness among subjects. The consensus until now has been that all available SNPs should be used in the determination of these similarities. In contrast, we show that selecting only certain SNPs to estimate genetic similarity improves power and P value calibration (the avoidance of inflation or deflation of the test statistic), while also dramatically reducing runtime and memory use.

Our approach is motivated by the fact that those SNPs used to estimate similarity for the LMM can be viewed as covariates in a linear regression. Given this mathematical equivalence, it becomes clear that one should use only SNPs to estimate genetic similarity that provide information about the phenotype. Such SNPs include causal SNPs or those nearby that tag causal SNPs, and SNPs that are associated by way of confounding. By conditioning on causal or tagging SNPs, we reduce the noise in the assessment of the association. By conditioning on SNPs associated because of confounding, we control for such confounding. Moreover, if a SNP is unrelated to the phenotype, it should not be in the conditioning set, otherwise it would add variance to the correction for confounding. Following this reasoning yields our approach, called FaST-LMM Select, which can be described as a method for performing variable selection on the SNPs used to estimate genetic similarity.

We show that our approach performs well on a wide variety of human data sets, including WTCCC, GAW14, and NFBC66.

DEVELOPMENT AND USE OF ACTIVE CLINICAL DECISION SUPPORT FOR PREEMPTIVE PHARMACOGENOMICS

Hoffman JM 1, Bell GC 1, Hicks JK 1, Baker DK 3, Crews KR 1, Kornegay NM 1, Wilkinson MR 1, Lorier R 2, Stoddard A 2, Yang W 1, Smith C 1, Fernandez CA 1, Cross SJ 1, Haidar C 1, Howard SC 3, Evans WE 1, BroeckelU 2, Relling MV 1

1. Department of Pharmaceutical Sciences, St Jude Children's Research Hospital, Memphis, Tennessee
2. Department of Pediatrics, Medical College of Wisconsin, Milwaukee, Wisconsin
3. Department of Information Sciences, St Jude Children's Research Hospital, Memphis, Tennessee

Active clinical decision support (CDS) delivered through an electronic health record (EHR) includes essential computational tools to facilitate gene-based drug prescribing and other applications of genomics to patient care. Active CDS that provides concise gene-based drug prescribing recommendations at the time an affected drug is ordered is particularly important because pharmacogenetic tests can be conducted preemptively and have lifetime implications. We describe the development and use of active CDS delivered to clinicians through a widely used commercial EHR (Cerner Corporation, Kansas City, MO) as part of preemptive array-based genotyping implemented at St. Jude Children's Research Hospital. Genotyping is performed in a CLIA-approved laboratory as part of research protocol PG4KDS (http://www.stjude.org/pg4kds) using the Affymetrix DMET Plus array with a CYP2D6 copy number assay, testing for variants in 225 genes. Once clear clinical recommendations for prescribing can be made for a gene-drug pair, such as from consensus guidelines through the Clinical Pharmacogenetics Implementation Consortium (CPIC), an automated system is used to incorporate the genetic results and clinical interpretations into the EHR Pharmacogenetics Tab. High-risk phenotypes automatically populate the EHR Problem List, and the problem list is the unambiguous triggers that drive alert firing to clinicians when high-risk drugs are prescribed. By only interrupting clinicians when a high priority phenotype and high-risk drug are both present, our CDS approach is carefully crafted to limit alert fatigue, a common shortcoming of CDS. Besides these post-test alerts, pre-test alerts fire if an order is placed for a very-high-risk drug (e.g., thiopurines), but the patient does not yet have the appropriate pharmacogenetic test result in the EHR. The information presented in each alert is written to provide actionable recommendations in a standard format that can be quickly understood by busy clinicians. Decisions to migrate gene-drug pairs and the corresponding use of CDS are approved by an oversight committee, which reports to the hospital's Pharmacy and Therapeutics Committee. Because clinicians at our hospital can order single gene tests for selected genes, the CDS is configured to apply to all clinical genetic test results and intercept possible duplicate genetic tests. Our CDS can be readily modified to incorporate new genes or high-risk drugs. Through November 2012, 35 customized rules have been built and implemented, including rules for TPMT with thiopurines (azathioprine, thioguanine, and mercaptopurine) and CYP2D6 with various drugs (codeine, tramadol, amitriptyline, fluoxetine, and paroxetine). From 5/19/2011 to 11/25/2012, the pre-test alerts have fired 938 times (64 for thiopurines and 874 for codeine and other drugs metabolized by CYP2D6), and the post-test alerts have fired 660 times (640 for TPMT and 20 for CYP2D6). Our experience illustrates the feasibility of developing computational systems that provide clinicians with actionable alerts for gene-based drug prescribing at the point of care.

PHARMGKB: GENOME ANNOTATION PROJECT

DJ Klein, M Whirl-Carrillo, RB Altman, TE Klein

Department of Genetics, Stanford University, Stanford, CA 94305

PharmGKB (www.pharmgkb.org) is the foremost pharmacogenomics online resource. Pharmacogenomic information is manually curated from the literature and deposited in the knowledgebase for annotation and aggregation. PharmGKB also disseminates genotype-based drug dosing guidelines published by the Clinical Pharmacogenetics Implementation Consortium (CPIC) and FDA drug labels containing genetic information. The information contained in PharmGKB is used to annotate eight subjects' genotypes as reported by 23andMe. A pharmacogenomic profile for each subject, including the determination of the subject's haplotypes for several cytochrome P450s (CYP2C9, CYP2C19, CYP2D6, etc.), is created. The haplotypes are predicted as accurately as possible given the un-phased alleles available from the 23andMe genotyping platform. Less frequent, but potentially important, haplotypes for some subjects that are often not well studied in the pharmacogenomics literature have been identified. Preliminary genomic analysis results are presented in this poster.

Work supported by NIH grant R24 GM61374.

INDUCED-FIT DOCKING AND STRUCTURAL ANALYSIS OF THE SELECTIVE PHOSPHATIDYLINOSITIDE 3-KINASE INHIBITORS

Yoonji Lee
National Leading Research Lab (NLRL) of Molecular Modeling & Drug Design, College of Pharmacy, Division of Life & Pharmaceutical Sciences, and National Core Research Center for Cell Signaling & Drug Discovery Research, Ewha Womans University, Seoul 120-750, Korea

Junghyun Lee
National Leading Research Lab (NLRL) of Molecular Modeling & Drug Design, College of Pharmacy, Division of Life & Pharmaceutical Sciences, and National Core Research Center for Cell Signaling & Drug Discovery Research, Ewha Womans University, Seoul 120-750, Korea

Sungwoo Hong
Department of Chemistry, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 305-701, Korea)

Sun Choi
(National Leading Research Lab (NLRL) of Molecular Modeling & Drug Design, College of Pharmacy, Division of Life & Pharmaceutical Sciences, and National Core Research Center for Cell Signaling & Drug Discovery Research, Ewha Womans University, Seoul 120-750, Korea

Phosphatidylinositide 3-kinases (PI3Ks) are a family of lipid kinases and catalyze the phosphorylation of the 3-hydroxyl position of phosphatidylinositides. Since the lipid products play a pivotal role in the cellular signaling network of several essential biological processes including oncogenesis, PI3K is considered as a promising target for anti-cancer therapy. Recently, a series of imidazo[1,2-a]pyridine analogues were synthesized, and the C8-unsubstituted imidazopyridine compounds showed inhibitory activities on PI3Ks along with CDKs with high potency. Interestingly, the substitution of the C8 hydrogen of the imidazopyridine ring with fluorine resulted in the selective inhibition of PI3Ks over CDKs. The chloro- and methyl-substituted compounds also maintained the good inhibitory activity on PI3Ks and low activity on CDKs. These significant activity differences of the tested compounds even with the slight chemical modifications might be due to their unique binding at the ATP-binding site of PI3K. To explain these results, induced-fit docking studies were carried out with the available X-ray crystal structures of the tested kinases. The C8-unsubstituted compound appears to bind very nicely at the ATP binding sites of PI3K alpha, PI3K gamma, and CDK2, consistent with its good biological activity. In contrast, the C8-substituted derivatives could not bind to CDK2 even though they could bind quite well to the PI3Ks. The PI3K alpha and PI3K gamma could well tolerate the C8-substituents, however, in case of CDK2, the binding site of the C8-hydrogen atom turned out to have spatial limitation which could only accommodate the size of hydrogen atom but not enough for fluorine or bigger substituents. Furthermore, the C8-hydrogen is directly facing the carbonyl oxygen of Glu81, whose lone pair electrons might also contribute to reduce the binding affinities of halogen-containing compounds by electrostatic repulsion. Consequently, the C8-substituted compound would bump into the binding pocket due to the net effects of the steric and electrostatic factors. Through the synthesis, biological evaluation and molecular modeling studies, structural requirements for PI3K selectivity were identified, and these results could be utilized for the further design of the selective inhibitors of PI3K.

STRUCTURALLY ALIGNED LOCAL SITES OF ACTIVITY (SALSAS) COMPUTATIONAL METHOD FOR THE PREDICTION OF FUNCTION OF STRUCTURAL GENOMICS PROTEINS

Joslyn Lee, Mary Jo Ondrechen

The Protein Structure Initiative (PSI) is aimed at determining a 3D structure of proteins to generate research in the fields of protein function, identifying better therapeutics for genetic and infectious diseases, increase new experiment designs and a better methodology for protein production and crystallography. These generated protein structures yield useful information that may help in drug target design, unique genome sequences or evolutionary relationships. In the Protein Data Bank, there are over 86,487 deposited structures, 11,999 of those are structural genomics (SG) proteins and most of these are of unknown or uncertain function. A large number of solved structures from structural genomic centers contain correctly or incorrectly annotated function due to assignments based on sequenced information. We developed a computational method to assign function to these incorrectly annotated SG proteins using superfamily of proteins as known sets and compare the unknown to the established sets. The present method to predict function, Structurally Aligned Local Sites of Activity (SALSAs), uses functional active site predictors, Theoretical Microscopic Anomalous Titration Curve Shapes (THEMATICS) and Partial Order Optimum Likelihood (POOL), to first determine the important active site residues from the 3D structure. Second, these local active site structures are aligned to compare proteins of known function with SG proteins. Third, those matches are assigned scores with a similarity scoring matrix to assign correct annotation. We apply this method to the ribulose-phosphate binding barrel (RPBB) superfamily and annotate twenty-eight SG proteins. The results of this analysis yielded correct annotations for the structural genomics proteins but also a classification between the subgroups was demonstrated. The superfamily is divided into eight subgroups and their active sites are clearly differentiated. Literature information and the Catalytic Site Atlas (CSA) are used to verify the correct location of the active site for each subgroup. The advantage of THEMATICS and POOL revealed more residues of importance that distinguish between similarly catalyzed reactions seen in the OMPDC, HPS and KGPDC subgroups. Large-genomic scale annotations have been performed on a few subclasses within the superfamily but this is the first structural analysis of the superfamily. In the future we plan to automate this process.

PUF CO-REGULATORY PROTEINS: RNA-SEQ, RNA-BINDING PROTEINS, AND CONSERVED BINDING MOTIFS

Richard McEachin, Ashwini Bhasi, Trista Schagat, Aaron Goldstrohm

University of Michigan

PUF proteins regulate the abundance of mRNAs by binding a highly conserved 3'UTR motif (UGUANAUA). Due to the diversity of mechanisms seen in PUF target proteins, we hypothesize that PUF proteins do not act alone; rather, co-regulatory RNA-binding proteins have significant impacts on PUF function. To identify these co-regulators, we first identified PUF target mRNAs by knocking down PUF, identifying differentially abundant mRNAs by RNA-Seq, and identifying the subset of PUF response genes that have PUF binding motifs. This process yielded 99 likely direct (binding) targets of PUF. We screened these mRNA sequences for the canonical binding motifs of proteins in the RNA-Binding Protein Data Base (RBPDB) then used MEME to search for novel over-represented motifs. These steps yielded candidate PUF co-regulatory proteins and candidate novel binding motifs. Two interesting results include conserved poly-A binding protein sites, and a site that most closely matches the binding site for the AIRE transcription factor. Poly-A binding proteins bind the poly-A tails of mature mRNAs and have multiple effects on those mRNAs, including influencing abundance. We previously identified a potential association between PUF and poly-A tails, so identifying poly-A binding motifs in other segments of the mRNA may expand our understanding of this mechanism. The mechanism for a transcription factor such as AIRE to act as a PUF co-regulator is much less clear, though at least four transcription factors have been shown to bind both DNA and RNA in a sequence specific manner (MDM2, WT1, SmZF1, and Spi-1/PU.1). Beyond regulation of transcription, the role(s) that transcription factors may play in regulation of mRNAs are not well understood. However, this result provides tantalizing evidence that AIRE and other known regulatory proteins could serve as co-regulators with PUF proteins.

A PROPOSAL FOR WEB-BASED REVIEWS OF SUPPLEMENTS AND OTHER NUTRACEUTICALS IN AGING AND AGING RELATED DISEASE

Jackson Miller, Greg Ceniceroz, Sean Mooney

The Buck Institute for Research on Aging

There is perhaps no greater market for nutraceuticals than in aging and in aging related diseases. Vendors market products everywhere from scientific fact to scientific quackery, none with FDA approval. We are proposing to develop a forum for scientific review of the classes of these products using web-based technologies. Our currently titled resource, MedVerified, is focused on providing reviews of products and services with clear coverage on topics related to aging and age associated conditions for consumers. The main objective is to inform the user and MedVerified will provide unbiased advice and reviews to individuals without medical expertise.

MedVerified seeks to provide product reviews from experienced and knowledgeable professionals. The core of the reviews will be science based and will be disseminated to and for the public investigating matters of clinical relevance. MedVerified will be a network of consumers, volunteers and medical advisors. We are providing individual evaluations of the safety and efficacy of products and services related to aging and, if known, the organization behind them. Users of the resource will rely upon the presented information to be impartial, scientifically accurate, and unbiased from commercial interests or competing products.

Although MedVerified is presently in the concept stage, our objective is to help and inform individuals. We will be simultaneously promoting health and health related products. MedVerified will be a resource to help people make informed decisions to maintain their health as well as help them if they do become diagnosed with an age associated condition.

ANALYSIS AND DISCRIMINATION OF SUBCELLULAR LOCALIZATION BASED ON AMINO ACID SEQUENCES OF MEMBRANE PROTEINS.

Ryohei Nambu, Takanori Sasaki, Yuri Mukai

Meiji University

In order to carrying out protein functions, bio-synthesized proteins should be transported to particular organelles. Signal-peptides, which are responsible for transporting proteins to the Endoplasmic Reticulum, and nuclear localization signals, which transport proteins to the nucleus, contain information about the subcellular localization of proteins and are located in amino acid sequences. There are many variations in signal sequences of each organelle, however, many of them have not been discovered. These signals which consist of 20 to 30 residues and usually exist near the N-terminus enable proteins to be transported to each organelle. The lipid bilayers of each organelle consist of different kinds and ratios of lipid molecules, and membrane proteins and have individual characteristics. Therefore, their transport signals are thought to recognize conformation and physicochemical properties of each membrane, and the characteristics of the signals are recognized by each organelle. The transport signals are also thought to have high hydropathy because they are inserted into the lipid bilayer and have differences based on the various thicknesses of each organellar membrane. To find the characteristics of the transport signals that are localized to organellar membranes including the endoplasmic reticulum, the Golgi apparatus, and the plasma membrane from their amino acid sequences, the hydrophobic regions of membrane proteins were analyzed, and a computational discrimination method was developed in this study. Notably, there were a significant number of membrane proteins in the Golgi apparatus and the endoplasmic reticulum with important roles in the oligosaccharide modification of proteins including oligosaccharide synthases, polysaccharide-degrading enzymes and glycosyltransferases. Therefore the discrimination of organellar membrane proteins is thought to be useful for discovery of unknown proteins related to glycosylation. Data of type II membrane proteins, single-pass type membrane localized proteins which have one hydrophobic region, were extracted from Uniprot Knowledge Base/Swiss-Prot Release 2011_11. Hydropathy profiles of each protein were estimated by average hydropathy calculation. As a result, the most hydrophobic positions in each protein within the 100 amino acid residues from the N-terminus were included in annotation regions as transmembrane helices. The distance from the N-terminus to the most hydrophobic positions was different depending on the subcellular localization of the organelles. The sequences were aligned at the most hydrophobic positions. Hydropathy profiles and position-specific amino acid propensities in these regions were also different according to the positions of each organelle. Therefore, it is thought to be possible to predict subcellular localizations of membrane proteins through the characterization and optimization of hydropathy profiles and amino acid compositions around hydrophobic regions of each organelle dataset.

DESENSITIZATION OF PKC TRANSLOCATION IN CO-CULTURED APLSYIA SENSORY-MOTOR NEURON PAIRS

Faisal Naqib
Department of Physiology, Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada

Carole Abi Farah
Department of Neurology and Neurosurgery, Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada

Daniel Weatherill
Department of Neurology and Neurosurgery, Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada

Christopher C. Pack
Department of Neurology and Neurosurgery, Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada

Wayne S. Sossin
Department of Neurology and Neurosurgery, Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada

The output of intracellular signaling networks can be dependent on the timing of stimuli presentation, however, the molecular mechanisms by which a cell measures elapsed time between stimulations is not well known. In the following, we develop a mathematical model to elucidate these mechanisms and validate the model experimentally. The sensory-motor neuron synapse of Aplysia is an excellent model system to investigate the intracellular signaling networks involved in memory formation. In this system, stimuli that is separated by rest periods (spaced training) leads to persistent changes in synaptic strength that depend on biochemical pathways that are different from those that occur when the stimuli lacks rest periods (massed training). In isolated sensory neurons, applications of serotonin, the neurotransmitter implicated in inducing these synaptic changes, lead to desensitization of the PKC Apl II response. Spaced applications of 5HT lead to increased desensitization compared to massed applications of 5HT. This response is reversed in sensory neurons co-cultured with motor neurons and allowed to form a synapse. Desensitization of PKC activity was also found to be dependent on PKA activity and protein translation. We developed mathematical models of the desensitization of PKC Apl II activity. We learned from our models that the sensory neuron was capable of distinguishing spaced from massed applications of 5HT by the competition of two hypothetical proteins; one with a fast rate of synthesis and degradation and the other with a delayed synthesis rate. The reversal of PKC desensitization in sensory neurons co-cultured with motor neurons is explained by a change in the competitive interaction between the two proteins. Our models suggest a mechanism by which stimuli timing information is decoded by neurons.

ESTIMATING TUMOR PURITY AND CANCER SUBPOPULATIONS FROM HIGH-THROUGHPUT DNA SEQUENCING DATA

Layla Oesper
Department of Computer Science, Brown University

Ahmad Mahmoody
Department of Computer Science, Brown University

Benjamin J. Raphael
Department of Computer Science, Brown University & Center for Computational Molecular Biology, Brown University

Background: Tumors are highly heterogeneous with individual cells in a tumor typically having different complements of somatic mutations (Gerlinger et al., 2012). Most cancer sequencing studies sequence a tumor sample containing a mixture of normal (non-cancerous) cells and various subpopulations of tumor cells, each subpopulation with a different complement of somatic mutations. Methods to infer tumor purity -- the fraction of cancerous cells in a sample -- using SNP array data have recently been introduced (Carter et al., 2012, Van Loo et al., 2012). The recent study by Nik-Zainal et al. (2012) used one such method followed by manual analysis of somatic mutations to identify a clonal (majority) population and a number of sub-clonal populations in each of several breast cancer genomes.

Methods: We describe an algorithm to infer tumor purity and clonal/sub-clonal tumor subpopulations directly from high-throughput DNA sequencing data. Copy number data derived from high coverage DNA sequencing provides information about tumor purity, as well as clonal and subclonal populations. Suppose that a mixture of cells is sequenced, with each cell differing from the normal (reference) genome by some number of copy number aberrations. We assume that the cells are partitioned into a small number of subpopulations. The genome of each subpopulation is represented by the number of copies it contains of each genomic interval, or its copy number profile, and each subpopulation has a corresponding mixture fraction. We formulate the Most Likely Mixture Decomposition Problem of finding a collection of genomes whose mixture best explains the observed sequencing data. We solve an instance of the problem using techniques from convex optimization, where the likelihood of our observed data is generated from a multinomial distribution.

Results: We applied our algorithm to 19 of the breast cancer samples analyzed by Nik-Zainal et al. Our algorithm infers the purity as well as the copy number profile and mixture fraction for each tumor subpopulation in these samples. We find that sample PD4120a contains 28% normal cells (compared to 30% estimated in Nik-Zainal et al.) and two tumor populations comprising 62% and 10% of cells in the sample. We identify clonal deletions including 1p, 4q and 16q. We also infer sub-clonal deletions of chromosomes 13 and part of 22q in 62% of cells, similar to the results reported by Nik-Zainal et al. (47% of cells), but we infer an alternate sequence of events resulting in these aberrations.

PREDICTING DRUG SIDE EFFECTS FROM AN INTEGRATIVE ANALYSIS OF GENOME-WIDE TRANSCRIPTOME AND PROTEIN INTERACTOME

Kyunghyun Park, Docyong Kim, Doheon Lee

KAIST

The side effect of drugs is one of reasons for failure in drug development. By improving bioinformatics techniques and increasing biomedical data, many researchers have tried to predict the side effects from a data-driven approach. Previous studies suggested that drug target is identified by side effect similarity and neighbors of drug target in protein-protein interaction network are used to predict cardiotoxicity as a case study. Some studies predicted effect of a drug using genome-wide transcriptome analysis. These studies used protein interaction and transcription profile respectively for similar purpose. Therefore, we suggested that an integration model of genome-wide transcriptome and protein interactome is accurate for predicting side effects. In this study, we made a rank based greedy algorithm inferring a protein sub-network from a drug target. The greedy algorithm was based on gene ranking by significant orders from gene expression data in response to drug treatment. We calculated the average AUC (area under the ROC curve) value of 10-fold cross validation and confirmed inferred side effect-associated genes in literature for an algorithm evaluation. The best average AUC of the greedy algorithm was 0.795 and 6 in top 10 side effect-associated genes were confirmed in literature. Our contribution is an integrative analysis of genome-wide transcriptome and protein interactome. Therefore, a proposed rank based greedy algorithm enables drug developers to find side effects in drug development.

INITIAL STUDY OF DARTMOUTH'S QUANTITATIVE BIOMEDICAL SCIENCES GRADUATE PROGRAM

Kristine A. Pattin
Institute for Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth

Anna C. Greene
Institute for Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth

Tor D. Tosteson
Department of Community and Family Medicine, Geisel School of Medicine at Dartmouth

Margaret R. Karagas
Department of Community and Family Medicine, Geisel School of Medicine at Dartmouth

Jason H. Moore
Institute for Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth

Graduate programs in computational biology and bioinformatics are now being created as formal training programs, though many are not multidisciplinary in nature. We discuss one such novel program, the Quantitative Biomedical Sciences (QBS) program in the Geisel School of Medicine and School of Graduate Studies at Dartmouth College established in 2010. The QBS program is an integrative program focused on the education of students (Ph.D. and eventually Masters level) in the areas of bioinformatics, biostatistics, and epidemiology and trains highly qualified quantitative students for productive careers in biomedical research and teaching through the completion of an interdisciplinary Ph.D. degree. Our philosophy is that the modern biomedical researcher must be able to speak more than one language to successfully collaborate in a highly multidisciplinary environment.

The QBS program of study begins with research rotations, a set of required courses, advanced electives, and necessary prerequisites. Required courses, known as the core competency courses, consist of two terms in each discipline: Bioinformatics, Biostatistics, and Epidemiology. First year students also enroll in a two-term core course in Integrative Biomedical Sciences that will provide a broad overview of the biomedical sciences with a focus on interdisciplinary thinking. Every term all students are required to participate in a QBS journal club that will feature student presentations and invited speakers. Training culminates in the production of a publishable dissertation based on original research in the student's chosen field of investigation.

The QBS program is currently in its second year, with 6 enrolled graduate students. Our students come into the program with varied scientific training, and the development of the curriculum is still underway to ensure the success of students with disparate educational backgrounds. We will discuss the implementation of the curriculum and proposed changes going forward. We will also review the research schedule and additional degree requirements. Student and faculty input is at the core of our curriculum design, and thus far overall feedback about the program has been extremely positive.

DE NOVO PREDICTION OF DNA-BINDING SPECIFICITIES FOR CYS2HIS2 ZINC FINGER PROTEINS

Anton Persikov and Mona Singh

Princeton University

Proteins with sequence-specific DNA binding function are important for a wide range of biological activities. De novo prediction of the DNA binding specificity of a protein from its sequence alone would be a great aid in understanding cellular networks. We introduce a method for predicting the DNA-binding specificities for Cys2His2 zinc fingers (C2H2-ZF), the largest family of DNA-binding proteins in eukaryotes. We develop a general approach, based on empirical calculations of pairwise amino acid–nucleotide interaction energies, for predicting position weight matrices (PWMs) representing the DNA-binding specificities for C2H2-ZF proteins. We show the effectiveness of the approach on a large data set of C2H2-ZF proteins with known DNA specificities. This method can be utilized for protein engineering applications and in genome-wide searches for transcription factor targets.

QUANTITATIVE ANALYSIS OF CHIP-SEQ DATA FOR CTCF PROTEIN

Joanna Raczynska, Keith Henderson, Dominika Borek, Zbyszek Otwinowski

University of Texas Southwestern Medical Center at Dallas

Chip-Seq is a Next Generation Sequencing technique that allows for genome-wide identification of protein binding sites. The procedure involves Immunoprecipitation of cross-linked chromatin fragments, followed by sequencing, read alignment and detection of read-enriched regions (peaks) that correspond to the protein binding sites. Current methods for Chip-Seq data analysis focus on peak detection without giving more attention to the peak heights. Also, the analysis is severely hindered by sequencing biases. It is especially problematic to perform an informative comparison of multiple experiments, e.g. from different cell lines, or with and without a drug treatment. The biological differences that one is looking for cannot be easily distinguished from experimental variation. Another problem in Chip-Seq is the lack of informative criteria for optimizing the experiment, like choosing the fragment length or the right amount of cross-linking. At present, one is left trying different strategies and hoping for the best, without means to judge the final outcome. Here we show how the biases can be corrected and how that leads to a meaningful interpretation of the peak heights. We also demonstrate how the histogram of peak height values is a good statistic to optimize in a Chip-Seq experiment. Our method for the bias correction relies on a Poisson regression procedure to find a set of coefficients that are then used for the calculation of weights for the reads, according to their sequence context. To make the results from different experiments directly comparable, we scaled the data using a maximum likelihood approach, as implemented in the program Scalepack (Otwinowski & Minor, Methods Enzymol. 1997). We analyzed 61 Chip-Seq data sets for CTCF protein from the Encode Project (The Encode Project Consortium, Science 2004). We focused on the comparison of peak heights between different cell lines using the CTCF motif as a reference point. We were able to identify sites that bind multiple CTCF molecules as well as sites that fall in the regions of high copy number in some cell lines.

COMPARISON OF RNA-SEQ NORMALIZATION AND DIFFERENTIAL EXPRESSION ANALYSIS METHODS USING SEQC DATA

Franck Rapaport, Raya Khanin, Yupu Liang, Azra Krek
Bioinformatics Core, Memorial Sloan-Kettering Cancer Center, New York

Paul Zumbo
Department of Physiology and Biophysics, Weill Cornell Medical College, New York; Institute for Computational Biomedicine, Weill Cornell Medical College, New York

Christopher E. Mason
Department of Physiology and Biophysics, Weill Cornell Medical College, New York; Institute for Computational Biomedicine, Weill Cornell Medical College, New York

Nicholas Socci
Bioinformatics Core, Memorial Sloan-Kettering Cancer Center, New York

Doron Betel
Division of Hematology/Oncology, Weill Cornell Medical College, New York; Institute for Computational Biomedicine, Weill Cornell Medical College, New York

RNA Sequencing (RNA-Seq) has emerged as powerful technology for study of transcriptome, motivating the development of many algorithms that aim at accurate detecting genes that are differentially expressed between conditions/phenotypes. In this study we performed a comprehensive comparison between several popular differential expression packages (DESeq, edgeR, Cuffdiff, PoissonSeq, and limma adapted to RNA-seq) using the extensively characterized SEQC samples (as part of the MACQ-III consortium) which include synthetic reference sequences as well as expression levels validated by TaqMan qPCR assays. Our comparison focused on a number of important aspects of RNA-seq differential expression analysis: Normalization: How it affects sample clustering as well as correlation with qPCR measured log-fold changes. Sensitivity and specificity: How does the list of differentially expressed compare to the genes characterized by qPCR as well as to the synthetic reference sequences. False positive rate: How many genes are falsely declared differentially expressed when comparing samples that should not exhibit any differential expression pattern (Type I error)? Sensitivity to zero read counts: Whether differential analysis of genes with zero read counts in one of the conditions has reduced sensitivity (Type II error)? Sensitivity to coverage and replication: What is the impact of the sequence coverage and the number of replicates on the detection of differential expression. We evaluated the software packages on SEQC dataset and found significant differences. While no single method outperforms all other approaches, methods based on Negative Binomial distribution of count data (DESeq, edgeR) exhibit the most consistent and robust detection of differential expression. This comparative study provides important guidelines for the use of the RNA-seq differential expression packages and points the direction for future improvements of RNA-seq analysis.

THE ROLE OF POSITIVE AND NEGATIVE FEEDBACK LOOPS OF P53 PATHWAY

Kyoungmin Roh and Stephen Proulx

University of California Santa Barbara, Ecology, Evolution and Marine Biology

The p53 gene is the most frequently mutated gene in human cancer and the p53 gene is known to control several cellular responses to DNA damage including cell cycle arrest, DNA repair, cellar senescence, and apoptosis. How does one protein coordinate these multiple responses? Is it more efficient or more robust for one protein to do all of these? Why has no backup protein evolved to reduce p53 protein's vulnerability to DNA damage? The most common answer to why evolution has resulted in a single master p53 gene is that one protein can act as the most efficient integrator of information about the different types of stress that act upon the cell. If multiple proteins receive multiple input signals, it would have to have a much more complicated communication system to integrate information about the environmental stress. However, p63 and p73 are in the same gene family as p53. They share amino acid sequence identity reaching 63% in the DNA-biding domain and have redundant functions in the regulation of gene expression. The p63 is essential for ectoderm development and the p73 regulates stress response. The p53, p63 and p73 might have overlapping functions but they all have distinct functions and mutations in p63 and p73 are rarely implicated in human cancer. However, we suspect that there is a backup protein or feedback loop that acts to reduce p53's vulnerability to DNA damage. We built deterministic p53 feedback loops and explored two additional scenarios: 1 negative p53 feedback loop, 1 negative and 1 positive p53 feedback loops, and 1 negative 2 positive p53 feedback loops. We assumed that the optimum response of p53 to DNA damage is that observed in non-cancerous cells and tried to find optimal parameter combinations using a simulated annealing algorithm. We found that both negative and positive feedback loops affect the p53 dynamics. When a positive feedback is added it results in apoptosis and cell death while addition of a negative feedback loop attenuates the p53 pathway. Also, we found that more loops make p53 is less sensitive to DNA damage. In other words, p53 feedback loops reduce the probability of apoptosis from the same level of DNA damage. We believe that our results provide new insight on the p53 feedback loops and increase understanding how to evolve p53 pathway.

SOFTWARE SUIT FOR PROCESSING AND ANALYSIS TRANSCRIPTOMICS AND GENOMICS DATA

Victor Solovyev, Igor Seledtsov, Denis Vorobyev, Vladimir Molodsov, Nicolay Okhalin

Softberry Inc.

The massive data volumes being generated by next-generation sequencing that provides fast sequencing of new genomes, genome-wide association studies, sequencing personal genomes, annotating the transcriptomes of cells, tissues and organisms and gene discovery by metagenomics studies. To process and analyze these data we have developed a software suit that consist of the following programs, pipelines and specific viewers of initial data and results of analysis.

1) ReadsMap is a fast reads aligner that quickly maps/aligns large sets of short DNA sequences. Multiple processors can be used optionally to achieve greater alignment speed. On initial stage we map "exonic" reads that demonstrate high-quality, non-interrupted alignment to a genomic sequence. At the second step, we use a modified variant of our EST_MAP program to align these "non-mapped" reads using splice site matrices and producing very accurate alignment with gaps. These reads will indentify potential exon-intron boundaries. The accuracy of RedsMap is significantly superior the known Bowtie program (especially on the spliced reads: Sn=0.985, Sp=0.973 vs Sn=0.375 and Sp=0.99 (Bowtie)).
2) TransSeq is RNASeq reads clustering program that assembles the RNA-Seq data into unique sequences of transcripts, often generating full-length transcripts for a set of alternatively spliced isoforms. The program demonstrates Sensitivity 0.97 and Specificity 0.92 in identifying known RNA transcripts on test data.
3) OligiZip assembler provides de novo reconstruction of genomic sequence or reconstruction of sequence using a reference genome.
4) GenomeMatch is fast alignment pipeline to align assembled contigs to the genomic sequence. This program can be used to align two genomic sequences finding their synteny regions.
5) Transomics pipeline initially maps reads to the genomic sequence and identifies spliced and non-spliced reads coordinates. This information is used by our FGENESH gene prediction program that includes an iterative procedure for predicting alternative splicing gene variants. It includes a module to compute a relative abundance of predicted alternative transcripts solving a system of linear equations.
6) SNP-effect reads a set of human SNPs and produces their genome/gene annotations using our database of human known and predicted genes and associated diseases. It also computes possible damaging effect of SNPs that located in protein sequences and provides available disease information.

We also developed a powerful Sequence assembling Viewer to work with the reads data and assembling results interactively and Genomic Viewer to investigate gene structures and alternatively spliced transcripts supported by RNASeq reads.

OligoZip, GenomeMatch and Transomics pipeline components and other software programs are available to run as pipelines for Unox platforms and independently at www.softwberry.com or as a part of integrated environment of the Molquest software package that can be downloaded at www.molquest.com for Windows, MAC and Linux OS.

PREDICTION OF DRUG-TARGET INTERACTION NETWORK USING FDA ADVERSE EVENT REPORT SYSTEM

Masataka Takarabe, Masaaki Kotera, Yosuke Nishimura, Susumu Goto
Institute for Chemical Research, Kyoto University

Yoshihiro Yamanishi
Medical Institute of Bioregulation, Kyushu University

Unexpected drug activities derived from off-targets are usually undesired and harmful, however they can occasionally be beneficial for different therapeutic indications. There are many uncharacterized drugs whose target proteins (including the primary target and off-targets) have been unknown. The identification of all potential targets for a given drug has become an important issue in genomic drug discovery, e.g., drug repositioning to reuse known drugs for new therapeutic indications. In this work, we defined pharmacological similarity for all possible drugs using the US Food and Drug Administration's (FDA's) adverse event reporting system (AERS). We then developed a new method to predict unknown drug-target interactions on a large scale from the integration of pharmacological similarity of drugs and genomic sequence similarity of target proteins in the framework of pharmacogenomic approach. In the results, we showed that the proposed method was applicable to a large number of drugs and it was useful especially for predicting unknown drug-target interactions which could not be expected from drug chemical structures. We made a comprehensive prediction for potential off-targets of 1,874 drugs with known targets and potential target profiles of 2,519 drugs without known targets. Our comprehensively predicted drug-target interaction networks supported us to suggest many potential drug-target interactions which were not predictable by the previous chemogenomic or pharmacogenomic approach and to increase research productivity toward genomic drug discovery.

THE PROPERTIES OF HUMAN GENOME CONFORMATION AND SPATIAL GENE INTERACTION AND REGULATION NETWORKS

Zheng Wang 1, Renzhi Cao 1, Kristen Taylor 2, Aaron Briley 2, Charles Caldwell 2, 5 and Jianlin Cheng 1, 3, 4

1 Computer Science Department, University of Missouri, Columbia, Missouri 65211, USA
2 Department of Pathology and Anatomical Sciences, School of Medicine, University of Missouri, Columbia, Missouri 65211, USA
3 Informatics Institute, University of Missouri, Columbia, Missouri 65211, USA
4 Christopher S. Bond Life Science Center, University of Missouri, Columbia, Missouri 65211, USA
5 Present Address, Translational Medicine, Eli Lilly and Company, Indianapolis, IN

The spatial conformation of a genome plays an important role in the long-range regulation of genome-wide gene expression and methylation, but has not been extensively studied due to lack of genome conformation data. The recently developed chromosome conformation capturing techniques such as the Hi-C method empowered by next generation sequencing can generate unbiased, large-scale, high-resolution chromosomal interaction (contact) data, providing an unprecedented opportunity to investigate the spatial structure of a genome and its applications in gene regulation, genomics, epigenetics, and cell biology. In this work, we conducted a comprehensive, large-scale computational analysis of this new stream of genome conformation data generated for three different human leukemia cells or cell lines by the Hi-C technique. We developed and applied a set of bioinformatics methods to reliably generate spatial chromosomal contacts from high-throughput sequencing data and to effectively use them to study the properties of the genome structures in one-dimension (1D) and two-dimension (2D). Our analysis demonstrates that Hi-C data can be effectively applied to study tissue-specific genome conformation, chromosome-chromosome interaction, chromosomal translocations, and spatial gene-gene interaction and regulation in a three-dimensional genome of primary tumor cells. Particularly, for the first time, we constructed genome-scale spatial gene-gene interaction network, transcription factor binding site (TFBS) – TFBS interaction network, and TFBS-gene interaction network from chromosomal contact information. Remarkably, all these networks possess the properties of scale-free modular networks.

COMPARISON OF KIDNEY AND LIVER TRANSCRIPTOMES USING RNA-SEQ

Xiang Qin, Michael Metzker, Hsu Chao, Harsha Doddapaneni, Donna Muzny, Richard Gibbs and Steve Scherer

Genomic variants play an important role in drug response and treatment outcomes. The RNA-seq component of the Pharmacogenomics Research Network (PGRN) is seeking to use deep RNA sequencing to investigate the association of gene expression profiles with genomic variants and drug response. The first phase of the project is to generate expression data from samples of major human organs of pharmacologic interest using next-generation sequencing technologies, and to comprehensively characterize the expression profiles including differentially expressed genes and splice variants from those normal tissues prior to case-control studies. As part of PGRN Deep Sequencing Resource, the Human Genome Sequencing Center at Baylor College of Medicine (BCM-HGSC) has completed the sequencing of 25 samples from each of four human tissues (tissues (heart, liver, kidney and adipose). Transcriptome data from kidney and liver was used to pilot RNA-seq data analyses to identify baseline gene and isoform expression levels in these two tissues. Alternative transcription start sites and splicing events were also analyzed to identify tissue-specific variants.

PERSONALIZED MEDICINE: FROM GENOTYPES AND MOLECULAR PHENOTYPES TOWARDS THERAPY POSTERS

THE PHARMACOGENOMICS BEHIND POPULATION DIFFERENCES OF ENALAPRIL RESPONSE IN SINGAPORE

Maulana Bachtiar [1], Jingbo Wang [1] and Caroline GL Lee [1,2,3]

[1] Department of Biochemistry, National University of Singapore, Singapore, SINGAPORE
[2] Division of Medical Sciences, National Cancer Centre Singapore, Singapore, SINGAPORE
[3] Duke-NUS Graduate Medical School, Singapore, SINGAPORE

Population difference in drug response is a common phenomenon. In Singapore, a multiracial South East Asian city-state, the occurrences of adverse drug reaction (ADR) often vary between patients from different ethnicity. In this study, we investigated the ethnic distribution of ADRs that was reported between 2007 and 2009 in Singapore. Among the most frequently reported ADR cases, the observed frequency of ADRs that are associated with the anti-hypertensive drug, Enalapril, is significantly different between the three major ethnic groups in Singapore: Chinese, Malays and Indians (corrected Chi Square P-Value = 2.34 x 10-5). Pharmacogenomics – the study of genetic factors behind drug response – continue to be a promising and attractive approach for addressing this challenge of population variation in drug response. Many have reported that racial differences in drug response are associated with Single Nucleotide Polymorphisms (SNPs) in the drug response genes. Here, utilizing data from the Singapore Genome Variation Project (SGVP) and HapMap (Phase 3), we analyzed the population differentiation of SNPs in Enalapril drug pathway. This includes SNPs in twenty genes belonging to both the renin-angiotensin-aldosterone system (RAAS) and angiotensin converting enzyme (ACE) inhibitor pathways. Our finding suggests that the Enalapril pathway, which has six genes carrying population-differentiated SNPs (FST scores in top five percentile), is relatively more population-differentiated in Singapore populations (Z-score SG = 0.46) compared to the general HapMap populations (Z-score HapMap = -0.98). Furthermore, there is a stronger enrichment of Enalapril genes carrying potentially functional SNPs that are also highly differentiated in the Singapore population (Z-score SG = 1.05 vs Z-score HapMap = -0.76). SNPs that can potentially affect splicing of BDKRB2 and CTSG genes in the Enalapril pathways are among those that are extremely population-differentiated in Singapore. In BDKRB2, several population-differentiated SNPs are also found to have been positively selected, which suggests for their possible functional significance in the drug pathway. These results demonstrate the potential usefulness of pharmacogenomics approach in elucidating the genetics factors behind population differences of ADRs that are associated with Enalapril.

BIOMARKER ROBUSTNESS REVEALS THE PDGF NETWORK AS DRIVING DISEASE OUTCOME IN OVARIAN CANCER PATIENTS IN MULTIPLE STUDIES

Rotem Ben-Hamo and Sol Efroni

The Mina and Everard Goodman Faculty of Life Science
Bar Ilan University, Keren Hayesod St., Ramat-Gan, 52900, Israel

Background: Ovarian cancer causes more deaths than any other gynecological cancer. Identifying the molecular mechanisms that drive disease progress in ovarian cancer is a critical step in providing therapeutics, improving diagnostics, and affiliating clinical behavior with disease etiology. Identification of molecular interactions that stratify prognosis is key in facilitating a clinical-molecular perspective.

Results: The Cancer Genome Atlas has recently made available the molecular characteristics of more than 500 patients. We used the TCGA multi-analysis study, and two additional datasets and a set of computational algorithms that we developed. The computational algorithms are based on methods that identify network alterations and quantify network behavior through gene expression. We identify a network biomarker that significantly stratifies survival rates in ovarian cancer patients. Interestingly, expression levels of single or sets of genes do not explain the prognostic stratification. The discovered biomarker is composed of the network around the PDGF pathway. The biomarker enables prognosis stratification.

Conclusion: The work presented here demonstrates, through the power of gene-expression networks, the criticality of the PDGF network in driving disease course. In uncovering the specific interactions within the network, that drive the phenotype, we catalyze targeted treatment, facilitate prognosis and offer a novel perspective into hidden disease heterogeneity.

CAGI: THE CRITICAL ASSESSMENT OF GENOME INTERPRETATION, A COMMUNITY EXPERIMENT TO EVALUATE PHENOTYPE PREDICTION

Steven E. Brenner 1, Susanna Repo 1,†, John Moult 2, CAGI Participants

1 Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720. brenner@compbio.berkeley.edu
2 IBBR, University of Maryland, Rockville, MD 20850. jmoult@umd.edu
†Currently at: EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK

he Critical Assessment of Genome Interpretation (CAGI, \'kā-jē\) is a community experiment to objectively assess computational methods for predicting the phenotypic impacts of genomic variation. In this assessment, participants are provided genetic variants and make predictions of resulting phenotype. These predictions are evaluated against experimental characterizations by independent assessors. The CAGI experiment culminates with a community workshop and publications to disseminate results, assess our collective ability to make accurate and meaningful phenotypic predictions, and better understand progress in the field. A long-term goal for CAGI is to improve the accuracy of phenotype and disease predictions in clinical settings.

The CAGI 2011 experiment consisted of 11 diverse challenges exploring the phenotypic consequences of genomic variation. In two challenges, CAGI predictors applied the state-of-the-art methods to identify the effects of variants in a metabolic enzyme and oncogenes. This revealed the relative strengths of each prediction approach and the necessity of customizing such methods to the individual genes in question; these challenges also offered insight into the appropriate use of such methods in basic and clinical research. CAGI also explored genome-scale data, showing unexpected successes in predicting Crohn's disease from exomes, as well as disappointing failures in using genome and transcriptome data to distinguish discordant monozygotic twins with asthma. Complementary approaches from two groups showed promising results in predicting distinct response of breast cancer cell lines to a panel of drugs. Predictors also made measurable progress in predicting a diversity of phenotypes present in the Personal Genome Project participants, as compared to the CAGI predictions from 2010.

CAGI 2012 challenges are presently open and we welcome participation from the community. Current information is available at the CAGI website at http://genomeinterpretation.org.

HARNESSING GENE-ENVIRONMENT INTERACTIONS TO IDENTIFY FUNCTIONAL TARGETS FOR MOLECULAR INTERVENTION IN PHENOTYPE

Julien Gagneur [1,2,*], Oliver Stegle [3,*], Chenchen Zhu [1], Petra Jakob [1], Dana Pe'er [4], Lars Steinmetz [1]

1 European Molecular Biology Laboratory, Heidelberg, Germany
2 Gene Center, Ludwig-Maximilians-Universität München,Germany
3 Max Planck Institute for developmental biology, Tübingen, Germany
4 Columbia University, New York, USA * equal contribution

Universal therapies for disease are increasingly difficult to find, motivating the search for personal treatments that are tailored to the genetic constitution and environmental exposures of the patient. To achieve this goal, it is important to identify and intervene in the precise molecular pathways that are affected in a given patient and cause disease.

Given the wealth of genetic and molecular profiling data being generated, the key challenge resides in developing suitable analytical strategies to predict causal molecular regulators that relay genetic and environmental signals to disease phenotypes.

Here, we propose new molecular signatures that are predictive of genes with causal roles in phenotype, and demonstrate their effectiveness for identifying functional molecular targets for intervention. We assessed these signatures using yeast as a model system, predicting functional molecular targets for specific growth QTLs in five environmental conditions. We comprehensively validated the predictions using high-throughput approaches, building on genome-wide deletion collections. Altogether, our results show that exploiting condition-specific genetic effects substantially increases the predictive accuracy over methods based on genetic or environmental variations alone.

Beyond proposing a new route towards identifying personal molecular targets from high-throughput omics data, our results contribute to the wider understanding of molecular systems and their role in mediating condition-specific genetic effects to cause complex traits and disease.

PREDICTING THE EFFECTS OF 3N INDELS

Jing Hu
Department of Mathematics and Computer Science, Franklin & Marshall College
Lancaster, PA, USA

Pauling C. Ng
Computational & Mathematical Biology, Genome Institute of Singapore, Singapore

Small insertions/deletions (indels 20 bp or less) account for nearly 23~24% of known Mendelian disease mutations. It is the second largest class of mutation type that leads to disease, following SNPs which account for over half of known Mendelian disease mutations. There are two types of indels in coding regions of the genome: frameshifting indels (i.e., indels that have a length that are not divisible by 3 which may cause frameshifts) and 3n indels (i.e., indels that have a length that are divisible by 3 which may cause amino acid insertions/deletions or block substitutions).Therefore, two different prediction methods are needed to predict the effects of different types of indels. In 2012, we have published SIFT Indel, a decision tree based method for predicting frameshifting indels. In our current project, we are extending SIFT indel to include predictions for 3n indels. The prediction method is still based on decision tree learning algorithm, but uses different feature set. We initially extracted 30 features and after a heuristic feature selection process, 4 features were chosen. The final decision tree algorithm achieved 85% sensitivity, 80% specificity, 83% accuracy, 0.657 MCC and 0.858 AUC (area under the curve) using 10-fold cross-validation. We applied the method to 3n indels identified from the human genomes sequenced by 1000 Genomes Project and Exome Project, and found that a higher percentage of rare indels is predicted to be gene-damaging compared to common indels. We are now updating the web server and the tool for 3n indels will be available at http://sift-dna.org/www/SIFT_indels2.html.

PHYLOGENOMICS AND POPULATION GENOMICS: MODELS, ALGORITHMS, AND ANALYTICAL TOOLS POSTERS

WHOLE-PROTEOME PHYLOGENY OF PROKARYOTES BY VARIABLE LENGTH EXACT SEQUENCE MATCH DECAY

Raquel Bromberg, Zbyszek Otwinowski

Advances in sequencing have allowed a large number of genomes to be sequenced in their entirety, with many more genomes to come in the near future. As more genomes are sequenced, many current tools for phylogeny will break down. Most alignment-based methods will not scale with the growth of information. Additionally, phylogenies constructed from whole genomes or proteomes are more robust than phylogenies done from single genes or sets of genes; for instance, a single case of horizontal gene transfer can completely misplace an organism in a phylogenetic tree made from an alignment of orthologs, whereas the same horizontal transfer, in a whole- genome or proteome phylogeny, will have almost no effect. As whole genome sequences continue to accumulate in the databanks and the 'data deluge' comes ever closer, it is not one-step methods that produce "perfect" phylogenies that should be sought, but rather methods that are scalable and whose results can be trusted to be correct within some reasonable measure. We present a whole-proteome method for constructing prokaryotic phylogenies. Our method relies upon the fact that organisms that are closely related will have more exact sequence matches, and over longer sequence lengths, in their proteomes or genomes than organisms that are more distantly related. Our algorithm counts the total number of exact matches for a range of sequence lengths between any pair of proteomes in an input set. Unlike with many other methods including those that depend upon sequence alignments, the input set for our method can be arbitrarily large. Our phylogenetic distances are derived from the rate of decay of the total number of matching sequences between pairs of organisms for increasing sequence lengths. The matrix of these distances is then converted into a tree via neighbor-joining. We have run our method on 137 archaea and 2003 bacteria separately and compared our classification with other methods, including other whole-genome or proteome methods, as well as the NCBI taxonomy, and have found that the resulting trees are in agreement.

SHORT BRANCHES CAUSE ARTIFACTS IN BLAST SEARCHES

Amanda A. Dick, Tim J. Harlow, and J. Peter Gogarten

Department of Molecular and Cell Biology, University of Connecticut
91 North Eagleville Road Unit 3125, Storrs, Ct 06236-3125

Long Branch Attraction (LBA) is a well-known artifact in phylogenetic reconstruction when dealing with branch length heterogeneity. Here we show another artifact, Short Branch Attraction (SBA), in blast searches that also results from branch length heterogeneity, but this time it is the short branches that are attracting. The SBA artifact is reciprocal and can be returned 100% of the time when multiple branches differ in length by a factor of more than two. SBA is an issue when reciprocal top scoring blast hit analyses are used to detect Horizontal Gene Transfers (HGT)s and can lead researchers to believe that there has been a HGT event when only vertical descent has occurred. SBA is also responsible for the changing results of top scoring blast hit analyses as the database grows, because more slowly evolving taxa, or short branches, are added over time, introducing more potential for SBA artifacts. SBA can be detected by examining reciprocal best blast hits among a larger group of taxa, including the known closest phylogenetic neighbors; therefore one should look for this artifact when conducting best blast hit analyses.

POST-NGS: ANALYSIS OF -OMES GENERATED BY NGS POSTERS

miRGATOR V3.0: A MICRORNA PORTAL FOR DEEP SEQUENCING, EXPRESSION PROFILING, AND MRNA TARGETING

Sooyoung Cho 1*, Insu Jang 2*, Yukyung Jun 1*, Suhyeon Yoon 1, Minjeong Ko 1, Yeajee Kwon 1, Ikjung Choi 1, Hyeshik Jang 3, Daeun Ryu 1, Byungwook Lee 2, V. Narry Kim 3, Wan Kyu Kim 1§, Sanghyuk Lee 1,2§

1 Ewha Research Center for Systems Biology (ERCSB), Ewha Womans University, Seoul 120-750, Korea
2 Korean Bioinformation Center (KOBIC), KRIBB, Daejeon 305-806, Korea
3 School of Biological Sciences, Seoul National University, Seoul, 151-742, Korea

Biogenesis and molecular function are two key subjects in the field of microRNA (miRNA) research. Deep sequencing has become the principal technique in cataloging of miRNA repertoire and generating expression profiles in an unbiased manner. Here, we describe the miRGator v3.0 update that compiled the deep sequencing miRNA data available in public and implemented several novel tools to facilitate exploration of massive data. The miR-seq browser supports users to examine short read alignment with the secondary structure and read count information available in concurrent windows. Features such as sequence editing, sorting, ordering, import and export of user data, would be of great utility for studying iso-miRs, miRNA editing and modifications. miRNA-target relation is essential for understanding miRNA function. Coexpression analysis of miRNA and target mRNAs, based on miRNA-seq and RNA-seq data from the same sample, is visualized in the heat-map and network views where users can investigate the inverse correlation of gene expression and target relations, compiled from various databases of predicted and validated targets. By keeping datasets and analytic tools up-to-date, miRGator, available at http://mirgator. kobic.re.kr, should continue to serve as an integrated resource for biogenesis and functional investigation of miRNAs.

BENCHMARKING THE EFFECTIVENESS OF ALGORITHMS FOR DETECTING FULL SPLICE FORMS FROM RNA-SEQ DATA

Katharina Hayer, Angel Pizarro, John Hogenesch, Gregory Grant

Background: High throughput RNA sequencing has dramatically increased the resolution of RNA abundance measurements. However, the data comes in the form of small reads, typically on the order of 100 bases long. It is straightforward to obtain accurate information about exons and exon/exon junctions. However, a solution remains elusive to the problem of combining this local information to obtain an accurate determination of exactly which full length splice forms are expressed and at which levels. There are several published algorithms offering solutions the problem, one of which, CUFFLINKS, has been widely accepted as the current gold standard. The accuracy of these methods on real data is difficult to assess, however, carefully designed simulated data can provide effective upper bounds on the accuracy. Methods: We have developed an RNA-Seq simulator called "Benchmarker for Evaluating the Effectiveness of RNA-Seq Software" (BEERS), which was originally used for benchmarking alignment algorithms [1]. The simulator is based on sampling from a set of gene models and mimicking the discrete operations that produce a paired-end reads from a realistic distribution of fragment lengths. In the current work we have used this simulator to assess the accuracy of CUFFLINKS and other full length transcript reconstruction algorithms. The first data set tests the dependence of the accuracy on the number of alternate splice forms that are expressed, ranging from one to ten, where two alternate forms of the same gene always share terminal exons. We used a perfect 100% accurate alignment as well as alignments with TopHat, RUM and GSNAP. Results: We present our findings as graphs of the false-positive and false-negative rates for the various tests. We found that in the simplest case, when there is only one splice form and it is abundantly expressed, sequenced without error, and perfectly aligned to a non-polymorphic genome, that the algorithms are fairly capable at detecting it. However, when factors are introduced such as multiple splice forms or alignment artifacts, the accuracy drops off precipitously. These results allow us to make the surprising conclusion that on real data the current published algorithms may not adequately address current needs, underscoring the need for further algorithmic development and funding. [1] Comparative Analysis of RNA-Seq Alignment Algorithms and the RNA-Seq Unified Mapper (RUM) Gregory Grant, Michael Farkas, Angel Pizarro, Nicholas Lahens, Jonathan Schug, Brian Brunk, Christian Stoeckert Jr, John Hogenesch and Eric Pierce. Bioinformatics (2011) 2011 Sep 15;27(18):2518-28.

NEXT-GENERATION SHORT READ SEQUENCE ANALYSIS OF PRIMARY TUMOR XENOGRAFTS

Michael Jones, Joshua Korn, David Ruddy, Hui Gao, Bella Gorbatcheva, John Monahan and Michael Morrissey

Transplantable human tumors in mice are commonly used as preclinical models for efficacy testing, pharmacokinetics, and target discovery. Sequence analysis of these xenografts is often complicated by mouse stromal (and other tissue) invasion. Computational methods can be developed to remove artifacts caused by mouse DNA contamination. We present a short read filtering method that removes artifacts and is powerful enough to limit false negatives. This mouse sequence filtering algorithm will be described along with the rationale and evidence of reduction of variant artifacts. Additionally, the extent of mouse sequence infiltration as well as variant changes as a consequence of transplantation and tumor passaging was investigated through short read sequence analysis, and will be discussed.

TEXT AND DATA MINING FOR BIOMEDICAL DISCOVERY POSTERS

IMPROVING CLASSIFICATION PERFORMANCE OF REAL AND PSEUDO MIRNA PRECURSORS

Xuan Tho Dang
Graduate School of Natural Science and Technology, Kanazawa University Kanazawa, Japan
Email: thodx@hnue.edu.vn

Osamu Hirose
Institute of Science and Engineering, Kanazawa University Kanazawa, Japan
Email: hirose@se.kanazawa-u.ac.jp

Kenji Satou
Institute of Science and Engineering, Kanazawa University Kanazawa, Japan
Email: ken@t.kanazawa-u.ac.jp

MicroRNAs (miRNAs) are short (~22nt) non-coding RNAs that play an indispensable role in gene regulation of many biological processes. Most of current computational, comparative, and non-comparative methods commonly classify human precursor microRNA (pre-miRNA) hairpins from both genome pseudo hairpins and other non-coding RNAs (ncRNAs). Although there were a few approaches achieving promising results in applying class imbalance learning methods, this issue has still not solved completely and successfully yet by the existing methods because of imbalanced class distribution in the datasets. For example, SMOTE is a famous and general over-sampling method addressing this problem, however in some cases it cannot improve or sometimes reduces classification accuracy. Therefore, we developed a novel over-sampling method named incremental-SMOTE to distinguish human pre-miRNA hairpins from both genome pseudo hairpins and other ncRNAs. Experimental results on pre-miRNA datasets from Batuwita et al. showed that our method achieved better Sensitivity and G-mean than the control (no over-sampling), SMOTE, and several successors of modified SMOTE including safe-level-SMOTE and borderline-SMOTE. In addition, we also applied the novel method to five imbalanced benchmark datasets from UCI Machine Learning Repository and achieved improvements in Sensitivity and G-mean. These results suggest that our method outperforms SMOTE and several successors of it in various biomedical classification problems including miRNA classification.

Keywords: imbalanced dataset, over-sampling, SMOTE, miRNA classification

COMPUTATIONAL NANO-DISSECTION IDENTIFIES CELL-LINEAGE SPECIFIC GENES WITH KEY ROLES IN RENAL HEALTH

Casey Greene
Department of Genetics, The Geisel School of Medicine at Dartmouth

Wenjun Ju
Department of Nephrology, University of Michigan Health System

Felix Eichinger
Department of Nephrology, University of Michigan Health System

Olga Troyanskaya
The Lewis-Sigler Institute for Integrative Genomics, Princeton University

Matthias Kretzler
Department of Nephrology, University of Michigan Health System

Cell-lineage-specific transcripts are believed to be essential for differentiated tissue function, mediate acquired chronic diseases, and are implicated in hereditary organ failure. To identify genes with cell-lineage-specific expression in lineages not accessible by experimental micro-dissection, we developed a genome-scale iterative method, “in silico nano-dissection.” Our method uses machine learning to leverage high-throughput functional-genomics data from tissue homogenates. We applied nano-dissection to chronic kidney disease (CKD) and identified transcripts specific to podocytes, key cells in the glomerular filter responsible for hereditary and most acquired chronic kidney disease. We systematically evaluated genes predicted as podocyte specific by our method and found that its accuracy exceeded predictions from in vivo fluorescence-tagged murine podocytes. Nano-dissection is broadly applicable to define lineage specificity in many functional and disease contexts. Our nano-dissection webserver allows biologists to apply our approach to their cell lineages of interest.

DIGSEE: DISEASE GENE SEARCH ENGINE WITH EVIDENCE SENTENCES

Jeongkyun Kim
School of Information and Communications
Gwangju Institute of Science and Technology, Korea

Hee-Jin Lee
Department of Computer Science, KAIST, Korea

Jong C. Park
Department of Computer Science, KAIST, Korea

Jung-jae Kim
School of Computer Science, Nanyang Technological University, Singapore

Hyunju Lee
School of Information and Communications
Gwangju Institute of Science and Technology, Korea

With a rapid accumulation of biological data for cancer and the development of computational methods to predict cancer-related genes, it became possible to generate a long list of candidate genes that might be involved in cancer. To validate these candidate genes, one of routine processes is to check whether these genes are previously reported in the biomedical literature by searching PubMed. In many cases, this process requires labor intensive manual evaluation to find evidence sentences from a long list of literatures. To automate this process, we developed a search engine to find evidences whether or not input genes are previously known as cancer related genes. For this purpose, we hypothesized that if a gene symbol and biological events such as gene expression changes and regulation of genes are shown together in the same sentence from a given literature, the literature might have a high chance to support that the gene is related to the cancer. For this task, we extracted gene symbols using ABNER and biological events using a Turku event extraction system. Then, we built inverted indexes for gene symbols and biological events. When users input gene symbols to our system, articles with evidence sentences containing both gene symbols and biological events are shown. These articles are sorted based on our proposed ranking function; each evidence sentence is scored for its relevance in cancer. For this purpose, we collected positive and negative training evidence sentences from nine events. Then, using these training sentences, we built ten linguistic based rules and constructed a Bayesian learning classifier. The accuracy of the proposed rules was tested using independent test evidence sentences. In a current version, our search engine was constructed for supporting cancers.

AN APPROACH TO EXTEND DOMAIN-DOMAIN INTERACTION NETWORKS BASED ON PROTEIN DOMAIN FUNCTIONAL SIMILARITY

Tu Kien T. Le
Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa, Japan
Email: kienltt@hnue.edu.vn

Osamu Hirose
Institute of Science and Engineering, Kanazawa University, Kanazawa, Japan
Email: hirose@se.kanazawa-u.ac.jp

Kenji Satou
Institute of Science and Engineering, Kanazawa University, Kanazawa, Japan
Email: ken@t.kanazawa-u.ac.jp

Identification of interacting domains among proteins is an important step to elucidate hidden functions for protein-protein interactions. There have been developed a number of computational methods for predicting domain-domain interactions from known protein-protein interactions or three-dimensional structures of protein complexes. However, high-throughput experimental techniques for determining the existence of protein-protein interactions often contain a large number of false positives while known structures of protein complexes available is still limited. In this study, we develop a new framework in predicting domain-domain interactions based on a link-prediction approach. By using a latent learning model which incorporates functional similarity of domain pairs into learning process, the prediction of unknown domain-domain interactions is reduced to a matrix completion task. One advantage of this approach is high scalability especially when our understanding about protein domains is still incomplete. The experimental results showed that our method achieved AUC score around 90%, and in addition, domain-domain interactions predicted by our method have a high fraction sharing rate with those of other state-of-the-art methods.

APPLICATION OF AN UNDER-SAMPLING METHOD TO BETA-TURN PREDICTION

Lan Anh T. Nguyen
Graduate School of Natural Science and Technology, Kanazawa University Kanazawa, Japan
Email: lananh@stu.kanazawa-u.ac.jp

Osamu Hirose
Institute of Science and Engineering, Kanazawa University Kanazawa, Japan
Email: hirose@se.kanazawa-u.ac.jp

Kenji Satou
Institute of Science and Engineering, Kanazawa University Kanazawa, Japan
Email: ken@t.kanazawa-u.ac.jp

Beta-turn is one of the most important reverse turns because of its role in protein folding. Due to its importance, computational approaches for predicting beta-turns have been actively studied. Many of the computational methods are based on machine-learning techniques such as support vector machines. One main problem of these approaches is the class imbalance: the number of beta-turn residues is much smaller than that of non-beta-turn residues. To relax the problem of the class imbalance, we propose a new under-sampling algorithm based on hypothesis margins. In the poster presentation, we will demonstrate an effectiveness of our approach through a number of experimental results, including the comparison with the method reported by Tang et al. (2011). Keywords: beta-turns prediction, class imbalance, under-sampling, hypothesis margin.

ANALYSIS OF NOUN PHRASES EXTRACTED FROM BIOMEDICAL TEXTS FOR SEMANTIC CATEGORY PREDICTION

Kenji Satou
Institute of Science and Engineering, Kanazawa University, Kanazawa, Japan
Email: ken@t.kanazawa-u.ac.jp

Due to the vast amount of papers published in the field of life science, it is quite important today to develop a method helpful for searching and extracting scientific knowledge from them. Among various efforts for it, the task of named entity recognition is an important basis to extract higher level of information, e.g. protein-protein interaction, gene regulation, and gene-disease relationship. Therefore, a number of studies are reported for recognition and extraction of named entities from biomedical texts. However, once a set of named entities was extracted and established, it is just used as a dictionary in text mining. In contrast, we focused on the analysis of the internal structures of named entities written as noun phrases. Some kind of adjective in a noun phrase, including present and past participle of verb, can give a clue to guess the meaning of its modificand. For example, “phosphorylated” mainly modify various kinds of molecules. On the other hand, adjectives like “small” and “important” have quite general meaning and cannot limit the semantic category of a noun. In this study, we analyzed large amount of noun phrases extracted from biomedical texts, using machine learning techniques including clustering. Through the analysis, a set of adjectives providing important information about semantic category of modified noun were identified. Using the information, we tried to search for nouns semantically similar to a given noun, for evaluating the possibility of semantic category prediction for emerging nouns.

D-IMPACT: AN EFFECTIVE METHOD TO IMPROVE THE CLUSTERING PERFORMANCE ON GENE EXPRESSION DATA

Vu Anh Tran
Graduate School of Natural Science and Technology, Kanazawa University Kanazawa, Japan
Email: tvatva2002@gmail.com

Osamu Hirose
Institute of Science and Engineering, Kanazawa University Kanazawa, Japan
Email: hirose@se.kanazawa-u.ac.jp

Kenji Satou
Institute of Science and Engineering, Kanazawa University Kanazawa, Japan
Email: ken@t.kanazawa-u.ac.jp

Clustering is a classical exploratory technique to discover the gene functions and disease mechanisms from the gene expression data. Clustering algorithm usually assigns samples to clusters based on the distance matrix to identify the underlying patterns of gene expression data. However, due to the high dimensionality and the small number of samples, it is very difficult to apply the clustering algorithms to analyze the gene expression data. The classical clustering algorithms, i.e., k-means, hierarchical agglomerative clustering (HAC), failed to cluster these data. To improve the clustering performance on gene expression data, we applied D-IMPACT, a data preprocessing method inspired by the clustering algorithm IMPACT to preprocess the data. D-IMPACT detects noise/outlier and enhances the distance matrix based on the attraction and density. We then employed the clustering algorithm k-means to cluster the dataset preprocessed by D-IMPACT, and then assigned the noisy samples to clusters detected by k-means. We tested the method above on a gene expression dataset from Gene Expression Omnibus: GSE9712. The results show that our method greatly improved the clustering result (Rand Index (RI) = 0.8030, Adjusted Rand Index (ARI) = 0.4302) compared to the clustering results on original dataset of HAC (RI = 0.6515 and ARI = 0.3297), CHAMELEON (RI = 0.5909 and ARI = -0.1208), and k-means (RI = 0.6515 and ARI = 0.1391). This result shows that our method clearly outperformed HAC, CHAMELEON, and k-means and is effective for improving the performance of clustering on gene expression data.

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

V

W

X

Y

Z