PACIFIC SYMPOSIUM ON BIOCOMPUTING 2012

ABSTRACT BOOK

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

ACCEPTED PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

IDENTIFICATION OF ABERRANT PATHWAY AND NETWORK ACTIVITY FROM HIGH-THROUGHPUT DATA

SSLPRED: PREDICTING SYNTHETIC SICKNESS LETHALITY

Nirmalya Bandyopadhyay, Sanjay Ranka and Tamer Kahveci

CISE Department, University of Florida
Gainesville, FL 32611, USA

Two genes in an organism have a Synthetic Sickness Lethality (SSL) interaction, if their joint deletion leads to a lower than expected fitness. Synthetic Gene Array (SGA) is a technique that helps in identifying SSL values for pairs of genes in a given set of genes. SSL interactions are useful to discover the co-expressed gene groups in the regulatory and signaling networks. Also, they are used to unravel the pair of pathways (subset of physically interacting genes) that substitute the functions of each other. Generating an SGA entry is costly as it requires producing and monitoring a double mutant (a progeny with two mutated genes). Generating a comprehensive SGA can be very expensive as the number of gene pairs is quadratic in the number of genes of the corresponding organism. In this paper, we develop a new method SSLPred to predict the SSL interactions in an organism. Our method is built on the concept of Between Pathway Models (BPM), where majority of the SSL pairs span across the two functionally complementing pathways. We develop a regression based approach that learns the mapping between the gene expressions of single deletion mutant to the corresponding SGA entries. We compare our method to the one by Hescott et al. for predicting the GI (Genetic Interaction) score of Saccharomyces cerevisiae (S. cerevisiae) on four benchmark datasets. On different experimental setups, on average SSLPred performs significantly better compared to the other method.

Predicting the Effects of Copy-Number Variation in Double and Triple Mutant Combinations

Gregory W. Carter†
The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA

Michelle Hays
Institute for Systems Biology, 401 Terry Avenue North, Seattle, WA 98109, USA

Song Li
Institute for Systems Biology, 401 Terry Avenue North, Seattle, WA 98109, USA

Timothy Galitski
Millipore Corporation, 290 Concord Road, Billerica, MA 01821, USA, and
Institute for Systems Biology, 401 Terry Avenue North, Seattle, WA 98109, USA

The study of genetic interactions is a powerful tool in inferring structure and function of biological networks. To date, genetic interaction studies have been dominated by pair-wise gene deletion screens. However, classical genetic analysis and natural genetic variation involve diverse gene forms ranging from null alleles to copy number variants. Moreover, genetic variation is typically multifactorial. Addressing multiple combinatorial genetic variations ranging in gene activity is therefore of critical value. We approach this problem using genetic network modeling that quantitatively encodes how genes influence the activity of one another and phenotype outcomes. A network model was initially inferred from linear decomposition of gene expression data. We used this network to predict the effects of combining multi-copy and deletion mutations of specific gene pairs and a gene triplet. Predicted expression patterns across hundreds of genes were experimentally validated. Prediction success was critically dependent on how a multi-copy gene interacted with other genes in the network model. This strategy provides a template for the inference, prediction, and testing of genetically complex hypotheses involving diverse genetic variation.

INTEGRATIVE NETWORK ANALYSIS TO IDENTIFY ABERRANT PATHWAY NETWORKS IN OVARIAN CANCER

Li Chen 1,2, Jianhua Xuan 1, Jinghua Gu 1, Yue Wang 1, Zhen Zhang 2, Tian-Li Wang 2, Ie-Ming Shih 2

1 The Bradley Department of Electrical and Computer Engineering, Virginia Tech 900 N. Glebe Road, Arlington, VA 22203, USA
2 Department of Pathology, Johns Hopkins Medical Institutions 1550 Orleans Street, Baltimore, MD 21231, USA

Ovarian cancer is often called the 'silent killer' since it is difficult to have early detection and prognosis. Understanding the biological mechanism related to ovarian cancer becomes extremely important for the purpose of treatment. We propose an integrative framework to identify pathway related networks based on large-scale TCGA copy number data and gene expression profiles. The integrative approach first detects highly conserved copy number altered genes and regards them as seed genes, and then applies a network-based method to identify subnetworks that can differentiate gene expression patterns between different phenotypes of ovarian cancer patients. The identified subnetworks are further validated on an independent gene expression data set using a network-based classification method. The experimental results show that our approach can not only achieve good prediction performance across different data sets, but also identify biological meaningful subnetworks involved in many signaling pathways related to ovarian cancer.

Role of Synthetic Genetic Interactions in Understanding Functional Interactions Among Pathways

Shahin Mohammadi, Giorgos Kollias, and Ananth Grama

Department of Computer Science, Purdue University, West Lafayette

Synthetic genetic interactions reveal buffering mechanisms in the cell against genetic perturbations. These interactions have been widely used by researchers to predict functional similarity of gene pairs. In this paper, we perform a comprehensive evaluation of various methods for predicting co-pathway membership of genes based on their neighborhood similarity in the genetic network. We clearly delineate the scope of these methods and use it to motivate a rigorous statistical framework for quantifying the contribution of each pathway to the functional similarity of gene pairs. We then use our model to infer inter-dependencies among KEGG pathways. The resulting KEGG crosstalk map yields significant insights into the high-level organization of the genetic network and is used to explain the effective scope of genetic interactions for predicting co-pathway membership of gene pairs. A direct byproduct of this effort is that we are able to identify subsets of genes in each pathway that act as `ports' for interaction across pathways.

Discovery of Mutated Subnetworks Associated with Clinical Data in Cancer

Fabio Vandin, Patrick Clay, Eli Upfal, Benjamin J. Raphael

Department of Computer Science, and Center for Computational Molecular Biology
Brown University

A major goal of cancer sequencing projects is to identify genetic alterations that determine clinical phenotypes, such as survival time or drug response. Somatic mutations in cancer are typically very diverse, and are found in different sets of genes in different patients. This mutational heterogeneity complicates the discovery of associations between individual mutations and a clinical phenotype. This mutational heterogeneity is explained in part by the fact that driver mutations, the somatic mutations that drive cancer development, target genes in cellular pathways, and only a subset of pathway genes is mutated in a given patient. Thus, pathway-based analysis of associations between mutations and phenotype are warranted. Here, we introduce an algorithm to find groups of genes, or pathways, whose mutational status is associated to a clinical phenotype without prior definition of the pathways. Rather, we find subnetworks of genes in an gene interaction network with the property that the mutational status of the genes in the subnetwork are significantly associated with a clinical phenotype. This new algorithm is built upon HotNet, an algorithm that finds groups of mutated genes using a heat diffusion model and a two-stage statistical test. We focus here on discovery of statistically significant correlations between mutated subnetworks and patient survival data. A similar approach can be used for correlations with other types of clinical data, through use of an appropriate statistical test. We apply our method to simulated data as well as to mutation and survival data from ovarian cancer samples from The Cancer Genome Atlas. In the TCGA data, we discover nine subnetworks containing genes whose mutational status is correlated with survival. Genes in four of these subnetworks overlap known pathways, including the focal adhesion and cell adhesion pathways, while other subnetworks are novel.

INTRINSICALLY DISORDERED PROTEINS: ANALYSIS, PREDICTION AND SIMULATION

Quasi-anharmonic analysis reveals intermediate states in the nuclear co-activator receptor binding domain ensemble

Virginia M. Burger[1;4], Arvind Ramanathan[2]; Andrej J. Savol[1;4], Christopher B. Stanley[3], Pratul K. Agarwal[2] and Chakra S. Chennubhotla[4]

1 Joint Carnegie Mellon University-University of Pittsburgh Ph.D. Program in Computational Biology,
2 Computational Biology Institute and Computer Science and Mathematics Division,
3 Neutron Scattering Science Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA
4 Department of Computational and Systems Biology, University of Pittsburgh, Pennsylvania 15260, USA

The molten globule nuclear receptor co-activator binding domain (NCBD) of CREB binding protein (CBP) selectively recruits transcription co-activators (TCAs) during the formation of the transcription preinitiation complex. NCBD:TCA interactions have been implicated in several cancers, however, the mechanisms of NCBD:TCA recognition remain uncharacterized. NCBD:TCA intermolecular recognition has challenged traditional investigation as both NCBD and several of its corresponding TCAs are intrinsically disordered. We use explicit solvent molecular dynamics simulations of up to 40 s to examine the conformational diversity of ligand-free NCBD. We introduce two novel techniques (a) dihedral quasi-anharmonic analysis (dQAA) and (b) hierarchical graph-theoretic clustering to quantify the conformational heterogeneity of ligand-free NCBD. With this integrated approach we find that three of four ligand-bound states are natively accessible to the ligand-free NCBD simulations with a root-mean squared deviation (RMSD) less than 2 A° . These conformations are accessible via diverse pathways while a rate-limiting barrier must be crossed in order to access the fourth bound state.

EFFICIENT CONSTRUCTION OF DISORDERED PROTEIN ENSEMBLES IN A BAYESIAN FRAMEWORK WITH OPTIMAL SELECTION OF CONFORMATIONS

Charles K. Fisher
Committee on Higher Degrees in Biophysics, Harvard University
Cambridge, Massachusetts 02139-4307, United States Email: ckfisher@fas.harvard.edu

Orly Ullman
Department of Chemistry, Massachusetts Institute of Technology
Cambridge, Massachusetts 02139-4307, United States Email: orly@mit.edu

Collin M. Stultz*
Harvard-MIT Division of Health Sciences and Technology, Department of Electrical Engineering and Computer Science, and the Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139-4307, United States

Constructing an accurate model for the thermally accessible states of an Intrinsically Disordered Protein (IDP) is a fundamental problem in structural biology. This problem requires one to consider a large number of conformations in order to ensure that the model adequately represents the range of structures that the protein can adopt. Typically, one samples a wide range of structures in an attempt to obtain an ensemble that agrees with some pre-specified set of experimental data. However, models that contain more structures than the available experimental restraints are problematic as the large number of degrees of freedom in the ensemble leads to considerable uncertainty in the final model. We introduce a computationally efficient algorithm called Variational Bayesian Weighting with Structure Selection (VBWSS) for constructing a model for the ensemble of an IDP that contains a minimal number of conformations and, simultaneously, provides estimates for the uncertainty in properties calculated from the model. The algorithm is validated using reference ensembles and applied to construct an ensemble for the 140-residue IDP, monomeric α- synuclein.

INTRINSIC PROTEIN DISORDER AND PROTEIN-PROTEIN INTERACTIONS

Wei-Lun Hsu, Christopher Oldfield, Jingwei Meng, Fei Huang, Bin Xue#, Vladimir N. Uversky#, Pedro Romero, and A. Keith Dunker

Department of Biochemistry and Molecular Biology,
Indiana University School of Medicine
#Department of Molecular Medicine, University of South Florida

Intrinsically disordered proteins often bind to more than one partner. In this study, we focused on 11 sets of complexes in which the same disordered segment becomes bound to two or more distinct partners. For this collection of protein complexes, two or more partners of each disordered segment were selected to have less than 25% amino acid identity at structurally aligned positions. As it turned out that most of the examples so selected had similar 3D structure, the studied set was reduced to just these similar-fold cases. Based on the analyses of the interacting partners, the average sequence identity of the partners' binding regions showed substantially higher conservation as compared to the nonbinding regions: The residue identities, averaged over the 11 sets of partner proteins, were as follows: binding residues, 42 ± 6%; nonbinding residues 20 ± 3%; nonbinding buried residues 26 ± 5%; and nonbinding surface residues 16 ± 3%. The higher sequence identity of the binding residues compared to the other sets of residues provides evidence that these observed interactions are likely to be meaningful biological interactions, not artifacts. Since many of the features of the various interactions indicate that the disordered binding segments were likely to have been disordered before binding, these results also add further weight to the existence and function of intrinsically disordered regions inside cells.

SUBCLASSIFYING DISORDERED PROTEINS BY THE CH-CDF PLOT METHOD

Fei Huang, Christopher Oldfield, Jingwei Meng, Wei-Lun Hsu, Bin Xue, Vladimir N. Uversky, Pedro Romero and A. Keith Dunker

Intrinsically disordered proteins (IDPs) are associated with a wide range of functions. We suggest that sequence-based subtypes, which we call flavors, may provide the basis for different biological functions. The problem is to find a method that separates IDPs into different flavor / function groups. Here we discuss one approach, the (Charge-Hydropathy) versus (Cumulative Distribution Function) plot or CH-CDF plot, which is based the combined use of the CH and CDF disorder predictors. These two predictors are based on significantly different inputs and methods. This CH-CDF plot partitions all proteins into 4 groups: structured, mixed, disordered, and rare. Studies of the Protein Data Bank (PDB) entries and homologous show different structural biases for each group classified by the CH-CDF plot. The mixed class has more order-promoting residues and more ordered regions than the disordered class. To test whether this partition accomplishes any functional separation, we performed gene ontology (GO) term analysis on each class. Some functions are indeed found to be related to subtypes of disorder: the disordered class is highly active in mitosis-related processes among others. Meanwhile, the mixed class is highly associated with signaling pathways, where having both ordered and disordered regions could possibly be important.

FUNCTIONAL ANNOTATION OF INTRINSICALLY DISORDERED DOMAINS BY THEIR AMINO ACID CONTENT USING IDD NAVIGATOR

Ashwini Patil
Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai
Minato-ku, Tokyo 108-8639, Japan Email: ashwini@hgc.jp

Shunsuke Teraguchi
Host Defense Lab, WPI Immunology Frontier Research Center (IFReC), Osaka University,3-1
Yamadaoka, Suita, Osaka 565-0871, Japan   Email: teraguch@ifrec.osaka-u.ac.jp

Huy Dinh
Systems Immunology Lab, WPI Immunology Frontier Research Center (IFReC), Osaka
University, 3-1 Yamadaoka, Suita, Osaka 565-0871, Japan   Email: dinh@ifrec.osaka-u.ac.jp

Kenta Nakai
Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai
Minato-ku, Tokyo 108-8639, Japan Email: knakai@ims.u-tokyo.ac.jp

Daron M Standley
Systems Immunology Lab, WPI Immunology Frontier Research Center (IFReC), Osaka
University, 3-1 Yamadaoka, Suita, Osaka 565-0871, Japan   Email: standley@ifrec.osaka-u.ac.jp

Function prediction of intrinsically disordered domains (IDDs) using sequence similarity methods is limited by their high mutability and prevalence of low complexity regions. We describe a novel method for identifying similar IDDs by a similarity metric based on amino acid composition and identify significantly overrepresented Gene Ontology (GO) and Pfam domain annotations within highly similar IDDs. Applications and extensions of the proposed method are discussed, in particular with respect to protein functional annotation. We test the predicted annotations in a large-scale survey of IDDs in mouse and find that the proposed method provides significantly greater protein coverage in terms of function prediction than traditional sequence alignment methods like BLAST. As a proof of concept we examined several disorder-containing proteins: GRA15 and ROP16, both encoded in the parasitic protozoa T. gondii; Cyclon, a mostly uncharacterized protein involved in the regulation of immune cell death; STIM1, a protein essential for regulating calcium levels in the endoplasmic reticulum. We show that the overrepresented GO terms are consistent with recently-reported biological functions. We implemented the method in the web server IDD Navigator. IDD Navigator is available at http://sysimm.ifrec.osaka-u.ac.jp/disorder/beta.php.

MODULATING PROTEIN–DNA INTERACTIONS BY POST-TRANSLATIONAL MODIFICATIONS AT DISORDERED REGIONS

Dana Vuzman
Department of Structural Biology, Weizmann Institute of Science, Rehovot, 76100, Israel
Email: dana.golbin@weizmann.ac.il

Yonit Hoffman
Department of Structural Biology, Weizmann Institute of Science, Rehovot, 76100, Israel
Email: yonit.hoffman@weizmann.ac.il

Yaakov Levy
Department of Structural Biology, Weizmann Institute of Science, Rehovot, 76100, Israel
Email: Koby.Levy@weizmann.ac.il

Intrinsically disordered regions, particularly disordered tails, are very common in DNA-binding proteins (DBPs). The ability of disordered tails to modulate specific and nonspecific interactions with DNA is tightly linked to their being rich in positively charged residues that are often non-randomly distributed along the tail. Perturbing the composition and distribution of charged residues in the disordered regions by post-translational modifications, such as phosphorylation and acetylation, may impair the ability of the tail to interact nonspecifically with DNA by reducing its DNA affinity. In this study, we analyzed datasets of 3398 and 8943 human proteins that undergo acetylation or phosphorylation, respectively. Both modifications are common on the disordered tails of DBPs (3.1 ± 0.2 (0.07 ± 0.007) and 2.0 ± 0.2 (0.02 ± 0.003) acetylation and phosphorylation sites per tail (per tail residue), respectively). Phosphorylation sites are abundant in disordered regions and particularly in flexible tails for both DBPs and non-DBPs. While acetylation sites are also frequently occurred in the disordered tails of DBPs, in non-DBPs they are often found in ordered regions. This difference may indicate that acetylation has different function in DBPs and non-DBPs. Post-translational modifications, which often take place at disordered sites of DBPs, can modulate the interactions of proteins with DNA by changing the local and global properties of the tails. The effect of the modulation can be tuned by adjusting the number of modifications and the cross-talks between them.

MICROBIOME STUDIES: ANALYTICAL TOOLS AND TECHNIQUES

Estimating population diversity with unreliable low frequency counts

John Bunge
Department of Statistical Science, Cornell University, Ithaca, NY 14853, USA E-mail: jab18@cornell.edu www.northeastern.edu/catchall

Dankmar Bohning
Department of Mathematics and Statistics, University of Reading, Reading RG6 6BX, UK E-mail: d.a.w.bohning@reading.ac.uk

Heather K. Allen
Food Safety and Enteric Pathogens Research Unit, National Animal Disease Center, Agricultural Research Service, Ames, Iowa, 50010, USA E-mail: heather.allen@ars.usda.gov

James A. Foster
Department of Biological Sciences, University of Idaho, Moscow, ID 83844, USA E-mail: foster@uidaho.edu

We consider the classical population diversity estimation scenario based on frequency count data (the number of classes or taxa represented once, twice, etc. in the sample), but with the proviso that the lowest frequency counts, especially the singletons, may not be reliably observed. This arises especially in data derived from modern high-throughput DNA sequencing, where errors may cause sequences to be incorrectly assigned to new taxa instead of being matched to existing, observed taxa. We look at a spectrum of methods for addressing this issue, focusing in particular on fitting a parametric mixture model and deleting the highest-diversity component; we also consider regarding the data as left-censored and effectively pooling two or more low frequency counts. We find that these purely statistical "downstream" corrections will depend strongly on their underlying assumptions, but that such methods can be useful nonetheless.

COMPARISONS OF DISTANCE METHODS FOR COMBINING COVARIATES AND ABUNDANCES IN MICROBIOME STUDIES

Julia Fukuyama, Paul J. McMurdie
Statistics Department, Stanford University, Stanford, CA 94305, USA E-mail:{jfukuyama,mcmurdie}@stanford.edu

Les Dethlefsen, David A. Relman
Departments of Microbiology & Immunology, and Medicine Stanford University, VA Palo Alto Health Care System 154T 3801 Miranda Avenue, Palo Alto, CA 94304

Susan Holmes
Statistics Department, Stanford University, Stanford, CA 94305, USA

This article compares different methods for combining abundance data, phylogenetic trees and clin- ical covariates in a nonparametric setting. In particular we study the output from the principal coordinates analysis on UniFrac and weighted UniFrac distances and the output from a double principal coordinate analyses(DPCoA) using distances computed on the phylogenetic tree. We also present power comparisons for some of the standard tests of phylogenetic signal between different types of samples. These methods are compared both on simulated and real data sets. Our study shows that DPCoA is less robust to outliers, and more robust to small noisy fluctuations around zero.

SEPP: SATé -Enabled Phylogenetic Placement

S. Mirarab, N. Nguyen, T. Warnow

Department of Computer Science, University of Texas at Austin

We address the problem of Phylogenetic Placement, in which the objective is to insert short molecular sequences (called query sequences) into an existing phylogenetic tree and alignment on full-length sequences for the same gene. Phylogenetic placement has the potential to provide information beyond pure “species identification” (i.e., the association of metagenomic reads to existing species), because it can also give information about the evolutionary relationships between these query sequences and to known species. Approaches for phylogenetic placement have been developed that operate in two steps: first, an alignment is estimated for each query sequence to the alignment of the full-length sequences, and then that alignment is used to find the optimal location in the phylogenetic tree for the query sequence. Recent methods of this type include HMMALIGN+EPA, HMMALIGN+pplacer, and PaPaRa+EPA. We report on a study evaluating phylogenetic placement methods on biological and simulated data. This study shows that these methods have extremely good accuracy and computational tractability under conditions where the input contains a highly accurate alignment and tree for the full-length sequences, and the set of full-length sequences is sufficiently small and not too evolutionarily diverse; however, we also show that under other conditions accuracy declines and the computational requirements for memory and time exceed acceptable limits. We present SEPP, a general “boosting” technique to improve the accuracy and/or speed of phylogenetic placement techniques. The key algorithmic aspect of this booster is a dataset decomposition technique in SATé, a method that utilizes an iterative divide-and-conquer technique to co-estimate alignments and trees on large molecular sequence datasets. We show that SATé-boosting improves HMMALIGN+pplacer, placing short sequences more accurately when the set of input sequences has a large evolutionary diameter and produces placements of comparable accuracy in a fraction of the time for easier cases. SEPP software and the datasets used in this study are all available for free at http://www.cs.utexas.edu/users/phylo/software/sepp/submission.

ARTIFICIAL FUNCTIONAL DIFFERENCE BETWEEN MICROBIAL COMMUNITIES CAUSED BY LENGTH DIFFERENCE OF SEQUENCING READS

Quan Zhang
School of Informatics and Computing, Indiana University

Thomas G. Doak
Biology Department, Indiana University

Yuzhen Ye
School of Informatics and Computing, Indiana University

Homology-based approaches are often used for the annotation of microbial communities, providing functional profiles that are used to characterize and compare the content and the functionality of microbial communities. Metagenomic reads are the starting data for these studies, however considerable differences are observed between the functional profiles--built from sequencing reads produced by different sequencing techniques--for even the same microbial community. Using simulation experiments, we show that such functional differences are likely to be caused by the actual difference in read lengths, and are not the results of a sampling bias of the sequencing techniques. Furthermore, the functional differences derived from different sequencing techniques cannot be fully explained by the read-count bias, i.e. 1) the higher fraction of unannotated shorter reads (i.e. “read length matters”), and 2) the different lengths of proteins in different functional categories. Instead, we show here that specific functional categories are under-annotated, because similarity-search-based functional annotation tools tend to miss more reads from functional categories that contain less conserved genes/proteins. In addition, the accuracy of functional annotation of short reads for different functions varies, further skewing the functional profiles. To address these issues, we present a simple yet efficient method to improve the frequency estimates of different functional categories in the functional profiles of metagenomes, based on the functional annotation of simulated reads from complete microbial genomes.

MetaDomain: a profile HMM-based protein domain classification tool for short sequences

Yuan Zhang, Yanni Sun

Michigan State University

Protein homology search provides basis for functional profiling in metagenomic annotation. Profile HMM-based methods classify reads into annotated protein domain families and can achieve better sensitivity for remote protein homology search than pairwise sequence alignment. However, their sensitivity deteriorates with the decrease of read length. As a result, a large number of short reads cannot be classified into their native domain families. In this work, we introduce MetaDomain, a protein domain classification tool designed for short reads generated by next-generation sequencing technologies. MetaDomain uses relaxed position-specific score thresholds to align more reads to a profile HMM while using the distribution of alignment positions as an additional constraint to control false positive matches. In this work MetaDomain is applied to the transcriptomic data of a bacterial genome and a soil metagenomic data set. The experimental results show that it can achieve better sensitivity than the state-of-the-art profile HMM alignment tool in identifying encoded domains from short sequences. The source codes of MetaDomain are available at http://sourceforge.net/projects/metadomain/.

MODELING HOST-PATHOGEN INTERACTIONS

STRUCTURAL MODELS FOR HOST-PATHOGEN PROTEIN-PROTEIN INTERACTIONS: ASSESSING COVERAGE AND BIAS

Eric A. Franzosa
Bioinformatics Program, Boston University, 24 Cummington Street Boston, MA 02215, USA
Email: franzosa@bu.edu

Yu Xia
Bioinformatics Program, Department of Chemistry, Department of Biomedical Engineering
Boston University, 24 Cummington Street Boston, MA 02215, USA
Email: yuxia@bu.edu

Recently, we applied structural systems biology to host-pathogen interaction and constructed the human-virus structural interaction network (SIN) based on a combination of solved structures and homology models. Subsequent analysis of the human-virus SIN revealed significant differences between antagonistic human-virus and cooperative within-human protein-protein interactions (PPIs). Although the SIN approach is advantageous due to the complementary nature of 3D structure and network data, integration of these data sources is associated with two potential issues: reduced coverage of the full human-virus PPI network, and the introduction of specific biases from structure determination. In this work, we evaluate the impact of these issues by comparing the growth and properties of human-virus and within-human PPI networks with and without structural models. We find that although the human-virus SIN is small in size, it is largely depleted for false positives, which are common in the full network. In addition, the SIN shows potential for major growth in the near future. Furthermore, compared to the full network, the coverage of viral species in the human-virus SIN is large relative to its size, suggesting that it may be a less biased sampling of the universe of human-virus PPIs. Next, we systematically compare structural versus full networks of human-virus and within-human PPIs in terms of functional, physicochemical, and network properties. We find that although there exist biases inherent to the structural approach, such biases tend to affect both human-virus and within-human PPIs equally. As a result, the significant differences between structured human-virus and within-human PPI networks are never contradicted by the full networks. Collectively, these results suggest that a structural approach to host-pathogen systems biology is not only justified, but also highly complementary to previous approaches. In particular, conclusions drawn from direct comparisons of host-virus and host-host PPIs within the SIN are minimally confounded by the inherent biases of the structural approach.

IDENTIFICATION OF CELL CYCLE-REGULATED, PUTATIVE HYPHAL GENES IN CANDIDA ALBICANS

Raluca Gordan
Division of Genetics, Department of Medicine, Brigham & Women's Hospital and Harvard Medical School, Boston, MA 02115, USA Present address: Department of Biostatistics and Bioinformatics, Institute for Genome Sciences and Policy, Duke University Email: raluca.gordan@duke.edu

Saumyadipta Pyne
Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
Email: spyne@broad.mit.org

Martha L. Bulyk
Division of Genetics, Department of Medicine, Department of Pathology, Brigham & Women's Hospital and Harvard Medical School, Boston, MA 02115, USA Harvard-MIT Division of Health Sciences & Technology (HST) Harvard Medical School, Boston, MA 02115, USA
Email: mlbulyk@receptor.med.harvard.edu

Candida albicans, a major fungal pathogen in human, can grow in a variety of morphological forms ranging from budding yeast to pseudohyphae and hyphae, and its ability to transition to true hyphae is critical for virulence in various types of C. albicans infections. Here, we identify 17 putative hyphal genes whose expression peaks during the S/G2 transition of the cell cycle in C. albicans. These genes are Candida-specific (i.e., they do not have orthologs in S. cerevisiae, a related fungal species that does not exhibit hyphal growth and is primarily non-pathogenic), and their promoters are enriched for the DNA binding site motifs of Tec1 and Rfg1, two transcription factors (TFs) known to play important roles in hyphal growth and virulence. For 5 of the 17 genes we found strong evidence in the literature that confirms our hypothesis that these genes are involved in hyphal growth and/or virulence, for 5 additional genes we found suggestive (albeit weak) evidence, while the other genes remain to be tested. It will be interesting to determine in future studies whether these 17 putative hyphal genes, whose expression peaks during the S/G2 transition, are part of a mechanism for this pathogenic fungus to 'turn on' hyphal growth late during the cell cycle, or if these genes are used to sustain hyphal growth and ensure that the cell does not transition back to yeast growth. In either case, the involvement of these genes in hyphal growth makes them putative targets for new antifungal drugs aimed at inhibiting hyphae formation in C. albicans.

Determining confidence of predicted interactions between HIV-1 and human proteins using conformal method

Ilia Nouretdinov
Computer Learning Research Centre, Royal Holloway University of London, UK

Alex Gammerman
Computer Learning Research Centre, Royal Holloway University of London, UK

Yanjun Qi
Machine Learning Department, NEC Labs America, Princeton, NJ 08540, USA

Judith Klein-Seetharaman
Research Centre Juelich, Germany / University of Pittsburgh, Pittsburgh, PA 15260, USA

Identifying protein-protein interactions (PPI's) is critical for understanding virtually all cellular molecular mechanisms. Previously, predicting PPI's was treated as a binary classification task and has commonly been solved in a supervised setting which requires a positive labeled set of known PPI's and a negative labeled set of non-interacting protein pairs. In those methods, the learner provides the likelihood of the predicted interaction, but without a confidence level associated with each prediction. Here, we apply a conformal prediction framework to make predictions and estimate confidence of the predictions. The conformal predictor uses a function measuring relative 'strangeness' interacting pairs to check whether prediction of a new example added to the sequence of already known PPI's would conform to the 'exchangeability' assumption: distribution of interacting pairs is invariant with any permutations of the pairs. In fact, this is the only assumption we make about the data. Another advantage is that the user can control a number of errors by providing a desirable confidence level. This feature of CP is very useful for a ranking list of possible interactive pairs. In this paper, the conformal method has been developed to deal with just one class - class interactive proteins - while there is not clearly defined of 'non-interactive' pairs. The confidence level helps the biologist in the interpretation of the results, and better assists the choices of pairs for experimental validation. We apply the proposed conformal framework to improve the identification of interacting pairs between HIV-1 and human proteins.

PERSONALIZED MEDICINE

Interpretome: A Freely Available, Modular, and Secure Personal Genome Interpretation Engine

Konrad J. Karczewski, Robert P. Tirrell, Pablo Cordero, Nicholas P. Tatonetti, Joel T. Dudley, Keyan Salari, Michael Snyder, Russ B. Altman, Stuart K. Kim

Training Program in Biomedical Informatics, Department of Genetics, Department of Developmental Biology, Stanford University School of Medicine, Stanford, CA 94305, USA

The decreasing cost of genotyping and genome sequencing has ushered in an era of genomic personalized medicine. More than 100,000 individuals have been genotyped by direct-to-consumer genetic testing services, such as 23andme, which offer a glimpse into the interpretation and exploration of a personal genome. However, these interpretations, which require extensive manual curation, are subject to the preferences of the company and are not customizable by the individual. Academic institutions teaching personalized medicine, as well as genetic hobbyists, may prefer to customize their analysis and not be limited to corporate-backed interpretations. We present a system for private genome interpretation, which contains all genotype information in client-side interpretation scripts, supported by server-side databases. We provide state-of-the-art analyses for teaching clinical implications of personal genomics, including disease risk assessment and pharmacogenomics. Additionally, we have implemented client-side algorithms for ancestry inference, demonstrating the power of these methods without excessive computation. Finally, the modular nature of the system allows for plugin capabilities for custom analyses. This system will allow for personal genome exploration without compromising privacy, facilitating hands-on courses in genomics and personalized medicine.

A Kinase inhibition map approach for tumor sensitivity prediction and combination therapy design for targeted drugs

Ranadip Pal and Noah Berlow

Department of Electrical and Computer Engineering, Texas Tech University
Lubbock, TX, 79409, USA

Drugs targeting specific kinases are becoming common in cancer research and are a basis for personalized cancer therapy. Some of these drugs have the capacity to target multiple kinases. Promiscuous kinase inhibitors can be effective but the ”off-target” effects can bring in toxicity for the patient. Thus the success of targeted cancer therapies with nominal harmful side effects is dependent on administering a single or multiple combinations of kinase inhibitors that targets the minimum number of kinases required to inhibit the tumor pathways. This requires a framework to predict the tumor sensitivities of a drug or drug combination based on the knowledge of the kinase inhibitors of a drug. In this article, we present a novel approach to predict the tumor sensitivities of a drug based on the generation of deterministic and stochastic Kinase Inhibition Maps. We build sensitivity maps or truth tables for a cell line from experimentally generated tumor sensitivities to kinase inhibitor drugs and use them to predict the sensitivity of a new drug or drug combinations based on known kinase inhibitor targets. We test our algorithms on a dataset of a dog osteosarcoma cell line with 317 possible kinase inhibitor targets after application of 36 targeted drugs. Our proposed algorithms are able to predict the sensitivities with high accuracy based on the given kinase inhibitor targets.

MIXTURE MODEL FOR SUB-PHENOTYPING IN GWAS

David Warde-Farley 3, Michael Brudno 2, Quaid Morris 2;4, Anna Goldenberg 1

1Genetics and Genome Biology, SickKids Research Institute, 101 College Street, Toronto, ON M5G 1L7
2Department of Computer Science, University of Toronto, 6 King's College Rd., Toronto ON M5S 3G4
3Department d'Informatique et de Recherche Operationelle, Universite de Montreal, CP 6128 succ. Centre-ville, Montreal QC H3C 3J7
4Donnelly Centre, University of Toronto, 160 College Street, Toronto ON M5S 3E1, Canada

Genome Wide Association (GWA) studies resulted in discovery of genetic variants underlying several complex diseases including Chron's disease and age-related macular degeneration (AMD). Still geneticists find that in majority of studies the size of the effect even if it is significant tends to be very small. There are several factors contributing to this problem such as rare variants, complex relationships among SNPs (epistatic effect), and heterogeneity of the phenotype. In this work we focus on addressing phenotypic heterogeneity. We introduce the problem of identifying, from GWAS data, separate genotypic markers from overlapping mixtures of clinically indistinguishable pheno-types. We propose a generative model for this scenario and derive an expectation-maximization (EM) procedure to fit the model to data, as well as a novel screening procedure designed to identify skew specific to certain phenotypic regimes. We present results on several simulated datasets as well as preliminary findings in applying the model to type 2 diabetes dataset.

TEXT AND KNOWLEDGE MINING FOR PHARMACOGENOMICS

The Extraction of Pharmacogenetic and Pharmacogenomic Relations – A Case Study Using PharmGKB

Ekaterina Buyko, Elena Beisswanger and Udo Hahn

Jena University Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Germany

In this paper, we report on adapting the JReX relation extraction engine, originally developed for the elicitation of protein-protein interaction relations, to the domains of pharmacogenetics and pharmacogenomics. We propose an intrinsic and an extrinsic evaluation scenario which is based on knowledge contained in the PharmGKB knowledge base. Porting JReX yields favorable results in the range of 80% F-score for Gene-Disease, Gene-Drug, and Drug-Disease relations.

Key words: Information Extraction, Pharmacogenetics, Pharmacogenomics, PharmGKB

LINKING PHARMGKB TO PHENOTYPE STUDIES AND ANIMAL MODELS OF DISEASE FOR DRUG REPURPOSING

**Robert Hoehndorf 1, Anika Oellrich 2, Dietrich Rebholz-Schuhmann 2,
Paul N. Schofield 3, Georgios V. Gkoutos 1**

1 Department of Genetics, University of Cambridge
Downing Street, Cambridge, CB2 3EH, UK

2 European Bioinformatics Institute
Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

3 Department of Physiology, Development and Neuroscience
University of Cambridge Downing Street, Cambridge CB2 3EG, UK, and The Jackson
Laboratory, 600, Main Street Bar Harbor ME 04609-1500, USA

The investigation of phenotypes in model organisms has the potential to reveal the molecular mechanisms underlying disease. The large-scale comparative analysis of phenotypes across species can reveal novel associations between genotypes and diseases. We use the PhenomeNET network of phenotypic similarity to suggest genotype-disease association, combine them with drug-gene associations available from the PharmGKB database, and infer novel associations between drugs and diseases. We evaluate and quantify our results based on our method's capability to reproduce known drug-disease associations. We find and discuss evidence that levonorgestrel, tretinoin and estradiolare associated with cystic fibrosis (p < 2.65 x 10^-6, p < 0:002 and p < 0:031, Wilcoxon signed-rank test, Bonferroni correction) and that ibuprofen may be active in chronic lymphocytic leukemia (p < 2.63 x 10^-23, Wilcoxon signed-rank test, Bonferroni correction). To enable access to our results, we implement a web server and make our raw data freely available. Our results are the first steps in implementing an integrated system for the analysis and prediction of drug-disease associations for rare and orphan diseases for which the molecular basis is not known.

Integrating VA's NDF-RT drug terminology with PharmGKB: Preliminary Results

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

Laura C. Weiss
Bethel University 3900 Bethel Drive, St. Paul, MN, USA Email: laura-weiss@bethel.edu

Matthew J. Durski
Department of Health Sciences Research, Mayo Clinic 200 1st Street SW, Rochester, MN, USA
Email: durski.matthew@mayo.edu

Qian Zhu
Department of Health Sciences Research, Mayo Clinic 200 1st Street SW, Rochester, MN, USA
Email: zhu.qian@mayo.edu

Robert R. Freimuth
Department of Health Sciences Research, Mayo Clinic 200 1st Street SW, Rochester, MN, USA
Email: freimuth.robert@mayo.edu

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

Biomedical terminology and vocabulary standards play an important role in enabling consistent, comparable, and meaningful sharing of data within and across institutional boundaries, as well as ensuring semantic interoperability. The Veterans Affairs (VA) National Drug File Reference Terminology (NDF-RT) is a federally recommended standardized terminology resource encompassing medications, ingredients, and a hierarchy for high-level drug classes. In this study, we investigate the drug-disease relationships in NDF-RT and determine how PharmGKB can be leveraged to augment NDF-RT, and vice-versa. Our preliminary results indicate that with additional curation and analyses, information contained in both knowledge resources can be mutually integrated.

Discovery and Explanation of Drug-Drug Interactions via Text Mining

Bethany Percha, Yael Garten, Russ B. Altman

Stanford University

Drug-drug interactions (DDIs) can occur when two drugs interact with the same gene product. Most available information about gene-drug relationships is contained within the scientific literature, but is dispersed over a large number of publications, with thousands of new publications added each month. In this setting, automated text mining is an attractive solution for identifying gene-drug relationships and aggregating them to predict novel DDIs. In previous work, we have shown that gene-drug interactions can be extracted from Medline abstracts with high fidelity - we extract not only the genes and drugs, but also the type of relationship expressed in individual sentences (e.g. metabolize, inhibit, activate and many others). We normalize these relationships and map them to a standardized ontology. Equivalent relationships are mapped to common standards in a context-sensitive manner. In this work, we hypothesize that we can combine these normalized gene-drug relationships, drawn from a very broad and diverse literature, to infer both known and novel DDIs. Using a training set of established DDIs, we have trained a random forest classifier to score potential DDIs based on the features of the normalized assertions extracted from the literature that relate two drugs to a gene product. The classifier recognizes the combinations of relationships, drugs and genes that are most associated with the gold standard DDIs, correctly identifying 79.8% of assertions relating interacting drug pairs and 78.9% of assertions relating noninteracting drug pairs. Most significantly, because our text processing method captures the semantics of individual gene-drug relationships, we can construct mechanistic pharmacological explanations for the newly proposed DDIs. We show how our classifier can be used to explain known DDIs and to uncover new DDIs that have not yet been reported.

RANKING GENE-DRUG RELATIONSHIPS IN BIOMEDICAL LITERATURE USING LATENT DIRICHLET ALLOCATION

Yonghui Wu
Department of Biomedical Informatics, Vanderbilt University Nashville, TN 37203, USA
E-mail: yonghui.wu@Vanderbilt.Edu

Mei Liu
Department of Biomedical Informatics, Vanderbilt University Nashville, TN 37232, USA
E-mail: mei.liu@Vanderbilt.Edu

W. Jim Zheng
Department of Biochemistry, Medical University of South Carolina Charleston, SC 29425, USA
E-mail: zhengw@musc.edu

Zhongming Zhao
Department of Biomedical Informatics, Vanderbilt University Nashville, TN 37232, USA E-mail: zhongming.zhao@Vanderbilt.Edu

Hua Xu
Department of Biomedical Informatics, Vanderbilt University Nashville, TN 37232, USA E-mail: hua.xu@Vanderbilt.Edu

Drug responses vary greatly among individuals due to human genetic variations, which is known as pharmacogenomics (PGx). Much of the PGx knowledge has been embedded in biomedical literature and there is a growing interest to develop text mining approaches to extract such knowledge. In this paper, we present a study to rank candidate gene-drug relations using Latent Dirichlet Allocation (LDA) model. Our approach consists of three steps: 1) recognize gene and drug entities in MEDLINE abstracts; 2) extract candidate gene-drug pairs based on different levels of co-occurrence, including abstract level, sentence level, and phrase level; and 3) rank candidate gene-drug pairs using multiple different methods including term frequency, Chi-square test, Mutual Information (MI), a reported Kullback-Leibler (KL) distance based on topics derived from LDA (LDA-KL), and a newly defined probabilistic KL distance based on LDA (LDA-PKL). We systematically evaluated these methods by using a gold standard data set of gene-drug relations derived from PharmGKB. Our results showed that the proposed LDA-PKL method achieved better Mean Average Precision (MAP) than any other methods, suggesting its promising uses for ranking and detecting PGx relations.

ACCEPTED PROCEEDINGS PAPERS WITH POSTER PRESENTATIONS

INTRINSICALLY DISORDERED PROTEINS: ANALYSIS, PREDICTION AND SIMULATION

CORRELATION BETWEEN POSTTRANSLATIONAL MODIFICATION AND INTRINSIC DISORDER IN PROTEIN

Jianjiong Gao and Dong Xu

Department of Computer Science, C.S. Bond Life Sciences Center
University of Missouri, Columbia, Missouri 65211, USA

Protein intrinsic disorder has been shown to play an important role in some posttranslational modifications (PTM). In this paper, we systematically investigated the correlation between protein disorder and dozens of PTMs using data from UniProt/Swiss-Prot and 3-D structures solved by NMR from Protein Data Bank. We observed that many PTMs have a preference for occurrence in disordered regions, including phospho-serine/-threonine/-tyrosine, hydroxylation, sulfotyrosine, S-geranylgeranyl cysteine, deamidated glutamine, 4-carboxyglutamate, 6'-bromotryptophan and most of methylation; while a few PTMs have a preference for occurrence in ordered regions, including 4-aspartylphosphate, S-nitrosocysteine, tele-methylhistidine, FMN conjugation, 4,5-dihydroxylysine, 3-methylthioaspartic acid, most of ADP-ribosylation, and most of FAD attachment. It is also noted that acetyllysine does not show any significant preference for occurrence in either disordered or ordered regions. Further analysis of NMR structures suggested disorder-to-order transitions might be introduced by modifications of phospho-serine/-threonine, mono-/di-/tri-methyllysine, sulfotyrosine, 4-carboxyglutamate, and potentially 4-hydroxyproline. This study sheds light on the functions and mechanisms of various PTMs.

Intrinsic disorder within and flanking the DNA-binding domains of human transcription factors

Xin Guo (1), Martha L. Bulyk (2), and Alexander J. Hartemink (1)

1 Department of Computer Science, Duke University,
Box 90129, Durham, NC 27708-0129, USA E-mail: {xinguo,amink}@cs.duke.edu

2 Division of Genetics, Department of Medicine; Department of Pathology;
Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA E-mail: mlbulyk@receptor.med.harvard.edu

While the term "protein structure" is commonplace, it is increasingly appreciated that proteins may not possess a single, well-defined structure: some regions of proteins are intrinsically disordered. The role these intrinsically disordered regions (IDRs) play in protein function is an area of significant interest. In particular, because proteins containing IDRs are largely involved in processes related to molecular recognition, the question arises whether IDRs are important in these recognition events. It has been observed that IDRs are enriched in transcription factors (TFs) in comparison with other proteins, and we sought to explore this enrichment more precisely, with an eye toward functional dissection of the prevalence and locations of IDRs in different classes of TFs. Specifically, we considered the occurrences of 76 classes of DNA-binding domains (DBDs) within a comprehensive set of 1,747 human, sequence-specific TFs. For each DBD class, we analyzed whether a significant level of disorder was present within the DBD itself, the N-terminal or C-terminal sequence flanking the DBD, or both flanking sequences. We found that although the DBDs themselves exhibit significant order, the regions flanking the DBDs exhibit significant disorder, which suggests a functional role for such IDRs in TF DNA binding. These results may have important implications for studies of TFs not just in human but across all eukaryotes, and suggest future studies focused on testing the roles of N- and C-terminal flanking regions in determining or modulating the DNA binding affinity and/or specificity of the associated TFs.

Coevolved residues and the functional association for intrinsically disordered proteins

Chan-Seok Jeong and Dongsup Kim

Department of Bio and Brain Engineering
Korea Advanced Institute of Science and Technology (KAIST)

The evolution of intrinsically disordered proteins has been studied primarily by focusing on evolutionary changes at an individual position such as substitution and conservation, but the evolutionary association between disordered residues has not been comprehensively investigated. Here, we analyze the distribution of residue-residue coevolution for disordered proteins. We reveal that the degree of coevolved residues significantly decreases in disordered regions regardless of the sequence propensity, and the degree distribution of coevolved and conserved residues exclusively differs in each functional category. Consequently, the coevolution information can be useful for predicting intrinsic disorder and understanding biological functions of a disordered region from the sequence.

CRYPTIC DISORDER: AN ORDER-DISORDER TRANSFORMATION REGULATES THE FUNCTION OF NUCLEOPHOSMIN

Diana M Mitrea and Richard W Kriwacki

St Jude Children's Research Hospital, Dept. of Structural Biology

It is now well appreciated that disordered proteins and domains are prevalent in eukaryotic proteomes and that disorder is critically linked with their regulation and functionality. However, our recent observations with the multi-domain protein, nucleophosmin (Npm), suggest that the biological palette of disorder is more diverse than currently understood. The N-terminal oligomerization domain of Npm (Npm-N) can be transformed from a folded, pentameric structure to a monomeric, disordered state through changes in solution ionic strength and, importantly, through physiologically relevant post-translational modifications. Thus, it appears that Npm has been evolutionarily tuned to exist in equilibrium between disordered and ordered states. Results from us and others suggest that the function of Npm is regulated through shifts in this equilibrium via post-translational modifications. Interestingly, this polymorphic behavior is not detected using standard secondary structure and disorder prediction algorithms, which show Npm-N to be folded into beta strands, consistent with the structure of the pentameric form. We have used a combination of computational tools, including structure-based analysis, sequence analysis algorithms (NetPhos 1.0, SCRATCH, KinasePhos, GPS2.1, PONDR) and molecular mechanics energy calculations, to test the hypothesis that the polymorphic behavior of Npm-N can be understood on structural and energetic grounds. This computational strategy has resulted in the identification of unfavorable energetic "hot-spots" within the Npm-N structure which coincide with experimentally observed sites of post-translational modification. Based on these observations, we propose that Npm-N has evolved energetic switches within its structure to enable transformation to a disordered state through phosphorylation. We further propose that the transformation process is triggered by sequential phosphorylation of solvent exposed hot-spots followed by exposure and modification of additional but initially buried sites to completely shift the equilibrium to the disordered state. This regulated, shifting equilibrium is associated with control of Npm localization within the nucleolus, nucleoplasm and cytoplasm, and with its role in regulation of centrosome duplication through interactions with Crm1-Ran. More broadly, we present a general computational strategy to identify transformational hot-spots within proteins and to test the hypothesis that other proteins currently understood to be folded participate in functionally relevant order-disorder equilibria as we have observed for Npm. The identification of such polymorphic proteins would broaden the palette of protein disorder utilized in biological systems.

ON THE COMPLEMENTARITY OF THE CONSENSUS-BASED DISORDER PREDICTION

Zhenling Peng and Lukasz Kurgan

Electrical and Computer Engineering Department, University of Alberta, Edmonton, AB, Canada

Intrinsic disorder in proteins plays important roles in transcriptional regulation, translation, and cellular signal transduction. The experimental annotation of the disorder lags behind the rapidly accumulating number of known protein chains, which motivates the development of computational predictors of disorder. Some of these methods address predictions of certain types/flavors of the disorder and recent years show that consensus-based predictors provide a viable way to improve predictive performance. However, the selection of the base predictors in a given consensus is usually performed in an ad-hock manner, based on their availability and with a premise that more is better. We perform first-of-its-kind investigation that analyzes complementarity among a dozen recent predictors to identify characteristics of (future) predictors that would lead to further consensus-based improvements in the predictive quality. The complementarity of a given set of three base predictors is expressed by the differences in their predictions when compared with each other and with their majority vote consensus. We propose a regression-based model that quantifies/predicts quality of the majority-vote consensus of a given triplet of predictors based on their individual predictive performance and their complementarity measured at the residue and the disorder segment levels. Our model shows that improved performance is associated with higher (lower) similarity between the three base predictors at the residue (segment) level and to their consensus prediction at the segment (residue) level. We also show that better consensuses utilize higher quality base methods. We use our model to predict the best-performing consensus on an independent test dataset and our empirical evaluation shows that this consensus outperforms individual methods and other consensus-based predictors based on the area under the ROC curve measure. Our study provides insights that could lead to the development of a new generation of the consensus-based disorder predictors.

MICROBIOME STUDIES: ANALYTICAL TOOLS AND TECHNIQUES

PROTEOTYPING OF MICROBIAL COMMUNITIES BY OPTIMIZATION OF TANDEM MASS SPECTROMETRY DATA INTERPRETATION

Alys Hugo 1, Douglas J. Baxter 2, William R. Cannon 1, Ananth Kalyanaraman 4, Gaurav Kulkarni 4, Stephen J. Callister 3

1Computational Biology and Bioinformatics Group,
2Molecular Sciences Computing Facility,
3Biological Separations Group, Pacific Northwest National Laboratory Richland, WA 99352
4School of Electrical Engineering and Computer Science, Washington State University,
Pullman, WA, 99164.

We report the development of a novel high performance computing method for the identification of proteins from unknown (environmental) samples. The method uses computational optimization to provide an effective way to control the false discovery rate for environmental samples and complements de novo peptide sequencing. Furthermore, the method provides information based on the expressed protein in a microbial community, and thus complements DNA-based identification methods. Testing on blind samples demonstrates that the method provides 79-95% overlap with analogous results from searches involving only the correct genomes. We provide scaling and performance evaluations for the software that demonstrate the ability to carry out large-scale optimizations on 1258 genomes containing 4.2M proteins.

phyloseq: A Bioconductor package for handling and analysis of high-throughput phylogenetic sequence data

Paul J. McMurdie and Susan Holmes

Stanford University

We present a detailed description of a new Bioconductor package, phyloseq, for integrated data and analysis of taxonomically-clustered phylogenetic sequencing data in conjunction with related data types. The phyloseq package integrates abundance data, phylogenetic information and covariates so that exploratory transformations, plots, and confirmatory testing and diagnostic plots can be carried out seamlessly. The package is built following the S4 object-oriented framework of the R language so that once the data have been input the user can easily transform, plot and analyze the data. We present some examples that highlight the methods and the ease with which we can leverage existing packages.

PERSONALIZED MEDICINE

Finding Genome-Transcriptome-Phenome Association with Structured Association Mapping and Visualization in GenAMap

Ross E Curtis 1,2, Junming Yin 2, Peter Kinnaird 3, Eric P Xing 4

1) Joint Carnegie Mellon - University of Pittsburgh PhD Program in Computational Biology
2) Lane Center for Computational Biology
3) Human Computer Interaction Institute
4) Machine Learning Department
Carnegie Mellon University

Despite the success of genome-wide association studies in detecting novel disease variants, we are still far from a complete understanding of the mechanisms through which variants cause disease. Most of previous studies have considered only genome-phenome associations. However, the integration of transcriptome data may help further elucidate the mechanisms through which genetic mutations lead to disease and uncover potential pathways to target for treatment. We present a novel structured association mapping strategy for finding genome-transcriptome-phenome associations when SNP, gene-expression, and phenotype data are available for the same cohort. We do so via a two-step procedure where genome-transcriptome associations are identified by GFlasso, a sparse regression technique presented previously. Transcriptome-phenome associations are then found by a novel proposed method called gGFlasso, which leverages structure inherent in the genes and phenotypic traits. Due to the complex nature of three-way association results, visualization tools can aid in the discovery of causal SNPs and regulatory mechanisms affecting diseases. Using well-grounded visualization techniques, we have designed new visualizations that filter through large three-way association results to detect interesting SNPs and associated genes and traits. The two-step GFlasso-gGFlasso algorithmic approach and new visualizations are integrated into GenAMap, a visual analytics system for structured association mapping. Results on simulated datasets show that our approach has the potential to increase the sensitivity and specificity of association studies, compared to existing procedures that do not exploit the full structural information of the data. We report results from an analysis on a publically available mouse dataset, showing that identified SNP-gene-trait associations are compatible with known biology.

POSTER ABSTRACTS

GENERAL POSTERS

Using Bayesian Learning to Identify Protein Interactions from Mass Spectra

John Boyle, Adam Norberg, Sarah Killcoyne

We present a technique for identifying protein-protein interactions from tandem mass spectrometry based proteomics experiments. The tool, called DirtMiner, uses Bayesian learning to calculate the probability of interactions occurring between any pair of proteins. The tool uses known protein-protein interactions to perform feature selection. Features are selected, from the spectra data, which show similarity between known interaction pairs, and dissimilarity between randomly selected pairs whose interaction is unknown. The advantage of the tool is that it can accommodate the uncertainty that arrives due to: experiment reproducibility issues, as mass spectrometry experiments have a high degree of variability due to separation chemistry unpredictability and the random sampling nature of precursor ion selection in the equipment; and lack of accurate interaction data sets to learn from, as the commonly used data sets are noisy due to both the transient nature of interactions and experimental error. The tool has been used to identify stable complexes using differentially labeled protein pull-down data sets. The study used related to RNA transport proteins, and it was hypothesized that proteins that were pulled-down and showed similar heavy isotope incorporation would be part of the same stable complex. However existing techniques were not applicable to the analysis of the data due to size and variability of the data sets, the high levels of impurity and experimental factors. The DirtMiner tool was used to infer interacting pairs from the data, and the results were then validated using both cross-validation and also expert knowledge. The features that were selected were those that related to both isotope incorporation (e.g. similar ratio of heavy/light incorporation of lysine or arginine indicating stable proteins) and features indicative of interactions (e.g. charged residues such as glutamine presence in the peptide, cysteine occurrence presuming relating to a disulfide bridge). The interacting pairs were shown to have biological relevance, and the tool did successfully identify proteins involved in RNA complex formation, processing and transport. The DirtMiner tool has been built to deal with the uncertainties of both biological facts and experiment data sets. The tool has been successfully applied to the study of protein-protein interactions, and has been used to extract biologically knowledge from mass spectrometry data. DirtMiner is made available, with full source, freely to all.

Patterns of exposure-associated DNA methylation alterations in normal human tissues are dependent upon CpG island context

Brock C. Christensen, Dartmouth Medical School; E. Andres Houseman, Oregon State University; Carmen J. Marsit, Dartmouth Medical School; Shichun Zheng, University of California San Francisco; Margaret R. Wrensch, University of California San Francisco; Joseph L. Wiemels, University of California San Francisco; Heather H. Nelson, University of Minnesota; Margaret R. Karagas, Dartmouth Medical School; Raphael Bueno, Brigham and Women's Hospital; David J. Sugarbaker, Brigham and Women's Hospital; John K. Wiencke, University of California San Francisco; Karl T. Kelsey, Brown University

Altered DNA methylation has been linked to cancers and many other human diseases, though the effects of exposures on methylation in normal tissues are not well characterized. DNA methylation occurs at cytosines followed by guanine known as CpG sites. CpG sites are underrepresented in the human genome, but cluster in concentrations known as CpG islands; half of human genes have a promoter-based CpG island, and CpG island methylation is associated with gene silencing. In normal cells CpG islands are unmethylated, whereas non-island CpGs are generally methylated. We investigated patterns of exposure-related DNA methylation in three normal human tissues dependent upon CpG island context using the Illumina GoldenGate array to measure methylation at 1,413 autosomal CpG sites associated with 773 cancer-related genes. Subjects (n=101) provided normal (non-diseased) tissue from pleura, peripheral blood, or lung; exposures were asbestos, alcohol consumption, and smoking history (and packyears) respectively. To discern and describe the relationships between CpGs with coordinate methylation, a modified model-based form of unsupervised clustering known as recursively partitioned mixture modeling (RPMM) was used to cluster CpGs. This approach built classes of CpGs using a mixture of beta distributions to recursively split samples into parsimoniously differentiated classes. Separate models clustered CpGs for pleural, peripheral blood, and lung tissue samples, and each model was pruned to eight terminal RPMM classes. To test the hypothesis that methylation was associated with exposures we averaged methylation values to form a single, CpG-class-specific partial methylation statistic. For each set of RPMM classes, we fit a quasi-binomial model expressing the logit mean partial methylation for each class as a linear function of exposure; standard errors were computed using generalized estimating equations (GEE), from which confidence intervals were constructed. To account for multiple comparisons we tested the omnibus null hypothesis of no association in any of the eight partial methylation classes using a Wald test statistic constructed from the GEE estimates and robust variance-covariance matrix. We observed a significant, CpG island dependent association between asbestos exposure and methylation classes in pleural samples (P=7.8E-07), as well as between alcohol consumption and methylation classes in blood samples (P=0.002), though not between smoking status or packyears smoked and methylation classes in lung samples. However, in association with the exposure variables, the pattern of methylation alterations across RPMM CpG classes was consistent across all tissue types. More specifically, in association with exposures DNA methylation decreased at CpG island CpGs, and increased at CpGs that were not in CpG islands. Strikingly, this pattern of CpG island-dependent exposure-associated methylation alterations contrasted those patterns observed in a similar analysis (previously published) of age-associated DNA methylation. Collectively, these results suggest that differential mechanisms of epigenetic dysregulation contribute to age and exposure associated methylation alterations.

Coarse-grain modeling of siRNA-binding hydrophilic-b-cationic copolymers synthesized via RAFT polymerization

James W. Delancey, Jr. 1, Andrew C. Holley 2, Wilson Hannah 3, Charles L. McCormick 2, John J. Correia 3, and Randy M. Wadkins1*

1 Department of Chemistry & Biochemistry, University of Mississippi, University, MS 38677
2 Department of Polymer Science, University of Southern Mississippi, Hattiesburg, MS 39406
3 Department of Biochemistry, Univ. of Mississippi Med.Center, Jackson, MS 39216

Introduction

RAFT polymerization is a method of polymerization that can take place in aqueous media and allows for the quantitative control of block length, micro structure, and placement of special terminal groups. These key features make this polymerization method interesting for creating delivery devices for therapeutics in biological fluids. One specific use for devices made by RAFT polymerization is the delivery of small interfering ribonucleic acids (siRNAs), which are utilized in natural RNA interference pathways. Here, we use coarse-grained molecular modeling tools to investigate how various proportions of hydroxypropylmethacrylamide (HPMA) and dimethylaminopropylmethacrylamide (DMAPMA) RAFT block copolymers spontaneously aggregate and bind siRNA

Methods

The coarse-grained molecular modeling tools used are comprised in the GROMACS molecular dynamics package, the MARTINI coarse-grain force field, and our in-house software. Models for siRNA and RAFT polymers structures were created by us based on canonical parameters

Results

We have created both atomistic structures and coarse-grained structures for both siRNA and for RAFT polymers containing various proportions of HPMA and DMAPMA. The siRNA structure was coarse-grained using a configuration of beads and an elastic network that we modeled after coarse-graining methods for DNA. The polymer structures were built using in-house software designed to align a specific number of each monomer between both termini and offsets each monomer by a standard length. The structures were then assigned atom-types and minimized. At this stage, knowing whether these polymers form stable secondary structures is of interest. A preliminary molecular dynamics simulation was performed for in vacuo conditions with 315 HPMA and 30 DMAPMA monomers and a stable secondary structure assembled in 500 ps. Our current simulations include solvent, which will reduce electrostatic charges and may affect the stability of the secondary structure. The polymer structures are being coarse-grained similarly to the method used with MARTINI on amino acid simulations. Each coarse-grain will replace approximately 4 atoms in the polymer

Acknowledgements

This material is based on the work supported by the National Science Foundation under Grant No. NSF EPS-0903787.

Indiana University founds the National Center for Genome Analysis Support: now open for business

Thomas G. Doak (IU Bloomington)
Department of Biology

Le-Shin Wu (IU Bloomington)
University Information Technology Services (UITS)

Craig A. Stewart (IU Bloomington)
Executive Director, Pervasive Technology Institute
Associate Dean, Research Technologies, Associate Director, CREST

Robert Henschel (IU Bloomington)
Manager, High Performance Applications, Research Technologies/PTI

William K. Barnett (IU Bloomington)
Director, National Center for Genome Analysis Support
Associate Director, Center for Applied Cybersecurity Research, PTI

U.S. researchers are in the midst of dramatic developments in genome sequencing capabilities, driven by the availability of high throughput, low cost next-generation gene sequencers. To address the scientific challenges presented by this new wealth of gene sequence information, the National Science Foundation (NSF) has awarded Indiana University a $1.5-million grant (NSF Award #1062432 - ABI Development: National Center for Genome Analysis Support) to establish the National Center for Genome Analysis Support (NCGAS). NCGAS will support the use of genome analysis software, store the data sets, and curate open source genome analysis software. As an example, installed assemblers now include: SOAPdenovo, Velvet, ABySS, Celera Assembler, Allpaths, Arachne 2, and BMA+1.

A specific goal is to provide dedicated access to large memory supercomputers, such as IU's new Mason system. Each Mason compute node has 500GB of random access memory, critical for data-intensive science applications such as genome assembly. Mason's integration with the new NSF-funded Extreme Science and Engineering Discovery Environment (XSEDE) will provide campus-based integration known as "campus bridging."

IU's NCGAS partners include the Texas Advanced Computing Center (TACC) and the San Diego Supercomputer Center (SDSC), and will support software running on supercomputers at TACC and SDSC, as well as other supercomputers that are part of XSEDE.

NCGAS services will include:

- Consulting services for biologists who want to undertake genome analysis on our systems
- Assistance in running genome analysis software on our systems
- Hardened and optimized genome analysis software
- An equitable and easy to use process for NCGAS allocations

Early users include Yuzhen Ye (IU Bloomington School of Informatics) Metagenomics Sequence Analysis; Michael Lynch (IUB Department of Biology) Genome Assembly and Annotation; Genome Informatics (IU Bloomington Department of Biology) Improved Genome Annotation for Animals and Plants; Tatiana Foroud (IU School of Medicine, Medical and Molecular Genetics) Imputation of Genotypes And Sequence Alignment ; Michael Lynch (IU Bloomington Department of Biology) Daphnia Population Genomics; Jeff Palmer (IUB Department of Biology), Bob Jansen (Texas), and Jeff Mower (Nebraska) Assembly of plant genomes.

Quantitative Multifactor Dimensionality Reduction Method for Detecting Gene-Gene Interaction

Jiang Gui g, Diane Gilbert-Diamond g,h, Peter Andrews h, Folkert W. Asselbergs a, Scott M. Williams b, Patricia R. Hebert c, Christopher S. Coffey d, Hans L. Hillege a, Gerjan Navis e, Douglas E. Vaughan b, Wiek H. van Gilst a,f, Jason H. Moore g,h

a Department of Cardiology, University Medical Center Groningen, Groningen, The Netherlands
b Division of Cardiovascular Medicine, Department of Medicine, Vanderbilt University Medical School, Nashville, TN, USA
c Section of Cardiovascular Medicine, Department of Medicine, Yale University School of Medicine, New Haven, CT, USA
d Department of Biostatistics, School of Public Health, University of Alabama, Birmingham, Birmingham, AL, USA
e Department of Nephrology, University Medical Center Groningen, Groningen, The Netherlands
f Department of Clinical Pharmacology, University Medical Center Groningen, Groningen, The Netherlands
g Department of Community and Family Medicine, Dartmouth Medical School, Lebanon, NH, USA
h Department of Genetics, Dartmouth Medical School, Lebanon, NH, USA

The widespread use of high-throughput methods of SNP genotyping has created a number of computational and statistical challenges. The problem of identifying SNP-SNP interactions in case-control studies has been studied extensively and a number of new techniques have been developed. Little progress has been made, however in the analysis of SNP-SNP interactions in relation to continuous data. We present an extension of the two class multifactor dimensionality reduction (MDR) algorithm that enables detection and characterization of epistatic SNP-SNP interactions in the context of Quantitative trait. The proposed Quantitative MDR (Quant-MDR) method handles continuous data by modifying MDR's constructive induction algorithm to use T Test. Quant-MDR replaces balanced accuracy with T test statistics as the score to determine the best models

We used simulation to identify the empirical distribution of Quant-MDR's testing score. We then applied Quant-MDR to genetic data from from the ongoing prospective Prevention of Renal and Vascular End-Stage Disease (PREVEND) study. We identified several two-loci SNP combinations that have strong association with patients' Tissue plasminogen activator (t-PA) expression. Quant-MDR is capable of detecting interaction models with weak main effects. These epistatic models tend to be dropped by traditional linear regression approaches. With improved efficiency to handle genome wide datasets, Quant-MDR will play an important role in a research strategy that embraces the complexity of the genotype-phenotype mapping relationship, since epistatic interactions are an important component of the genetic basis of disease.

ATHENA: A multi-functional software package for performing powerful human genetics studies on complex phenotypes

Emily Rose Holzinger, Vanderbilt University,
Scott Dudek, Vanderbilt University
Eric Torstenson, Vanderbilt University
Stephen Turner, University of Virginia School of Medicine
Will Bush, Vanderbilt University
Carrie Buchanan, Vanderbilt University
Anurag Verma, Pennsylvania State University
Gretta Armstrong, Pennsylvania State University
Marylyn DeRiggi Ritchie, Pennsylvania State University

The Analysis Tool for Heritable and Environmental Network Associations (ATHENA) is a multi-functional software package that was developed to allow for tailored analyses with the aim of identifying multi-factor susceptibility models that associate with complex phenotypes. This tool was designed to address the issue of missing heritability in complex human traits. Missing heritability refers to the fact that most studies designed to elucidate the genetic architecture of complex traits have only been able to identify variants that explain a small proportion of the estimated trait variability due to genetic factors.

This phenomenon could be a result of overly simplistic study designs that assess the association of one single variable for one type of data at a time.

A smarter study design that takes into account the underlying complexity of genetic etiology for these traits could result in the identification of genetic models that explain a portion of this missing heritability

There are several key components to ATHENA that allow for this type study. First, the filtering and analysis techniques selected to be a part of ATHENA were chosen because they allow for both categorical (i.e. SNPs) and quantitative (i.e. gene expression levels) input data.
Second, a filtering method can be selected to reduce statistical noise, which is inherently abundant in high-throughput data.

Currently, there are two filtering options: filter based on previous biological knowledge (Biofilter) or stochastic filtering via a tree-based method that allows for interactions between variables (Random Jungle).

Third, an analytical method can be selected to identify models that predict either a quantitative (i.e. lipid levels) or a categorical (i.e. case vs. control) outcome.

Currently, ATHENA employs two analytical techniques that both use computational evolution to optimize either neural networks (grammatical evolution neural networks (GENN)) or symbolic regression formulas (grammatical evolution symbolic regression (GESR)). These methods were chosen because they are able to perform a computationally-feasible, non-exhaustive analysis to find predictive models while allowing for the detection of gene-gene and gene-environment interactions.

Importantly, ATHENA was designed with a flexible framework. Future studies will involve improving the current methods and incorporating new ones in order keep up with the rapidly evolving field of human genetics research.

Biofilter: A Software Package for the Integration of Biological Domain Knowledge for Genomic Studies

NEERJA KATIYAR
CENTER FOR SYSTEMS GENOMICS, PENN STATE UNIVERSITY
EMAIL: NVK5095@PSU.EDU

ALEX FRASE
CENTER FOR SYSTEMS GENOMICS, PENN STATE UNIVERSITY
EMAIL: ATF3@PSU.EDU

JOHN WALLACE
CENTER FOR SYSTEMS GENOMICS, PENN STATE UNIVERSITY
EMAIL: JRW32@PSU.EDU

ERIC S. TORSTENSON
CENTER FOR HUMAN GENETICS RESEARCH, VANDERBILT UNIVERSITY
EMAIL: TORSTENSON@CHGR.MC.VANDERBILT.EDU

CARRIE C. BUCHANAN
CENTER FOR HUMAN GENETICS RESEARCH, VANDERBILT UNIVERSITY
EMAIL: CARRIE.C.BUCHANAN@VANDERBILT.EDU

GRETTA J ARMSTRONG
CENTER FOR SYSTEMS GENOMICS, PENN STATE UNIVERSITY
EMAIL: GRETTAD@PSU.EDU

SARAH A PENDERGRASS
CENTER FOR SYSTEMS GENOMICS, PENN STATE UNIVERSITY
EMAIL: SARAH.PENDERGRASS@PSU.EDU

MARYLYN D. RITCHIE
CENTER FOR SYSTEMS GENOMICS, PENN STATE UNIVERSITY
EMAIL: MARYLYN.RITCHIE@PSU.EDU

The identification and characterization of susceptibility genes for common, complex human disease is a difficult challenge. The initial paradigm of focusing a study on just one or a few candidate genes or proteins has limited our ability to identify novel genomic variants. The current methodologies do not allow for the modeling of complex interactions in diverse data types. Biofilter incorporates biological knowledge from different public databases and a flexible analytical framework that provides the necessary toolkit for identifying disease susceptibility factors for complex traits. The previous version of Biofilter was limited to six sources of domain knowledge (KEGG, Reactome, Gene Ontology, Database of Interacting Proteins, Protein Families Database and Netpath) to demonstrate the utility of filtering GWAS data using prior biological knowledge. The updated version of Biofilter will include integrated information from fourteen additional data sources including Transcription regulatory regions database, ORegAnno, ECRbase, UCSC, UniPathway, Biocarta, Genesigdb, Protein Networks, PharmGKB, MINT, Biogrid, BIOBASE, Genetic Association database and Metacyc. These data sources are integrated into the SQLite database for the Biofilter and the data source options are featured in the configuration file. The initial version of Biofilter is a command-line software package. To maximize the utility of the Biofilter for the scientific community, we will develop a graphical user interface to essentially create the configuration file for the system and facilitate intuitive, flexible design. In the initial version of Biofilter, biological relationships were treated as present or absent and there was no account for the strength or direction of relationship. Here, we have optimized the usage of the data sources to consider the nature of information contained using direct and indirect connections as well as strength of the effects. Generally, simulation studies are the status quo to determine performance. Such simulations are either seen to be quite challenging overly simplistic with biological prior knowledge based experiments. As such, we have performed proof-of-concept study specifically to address the performance of the Biofilter for detecting GXE

interactions associated with lipids in the Marshfield Personalized Medicine Research Project (PMRP) data. The Biofilter has two major functions: annotation and filtering. It is through this functionality that may lead to improved power for identifying gene-gene interactions or pathway associations with complex traits.

Mining of Mass Spectrometry Proteomics Experiments Through Integration of Spectral Libraries

Sarah Killcoyne, Richard B. Kreisberg, David Julian, John Boyle

Institute for Systems Biology

Here we introduce a suite of data mining tools developed to enable the design and analysis of targeted proteomic experiments, using selective reaction monitoring (SRM), which requires the mining of spectral libraries in order to identify suitable transitions for accurate and unique protein identification. These tools allow for the integration of multiple experiment technologies, which is becoming increasingly important in biomarker discovery. These integrative approaches to the identification of biomarkers require the ability to explore and visualize inferences drawn from spectral libraries, scientific literature and genomic information. This allows us identify the transitions that are most relevant to a particular disease or biological phenomena.

We have developed informatics approaches using proteomic spectral libraries (e.g. the PeptideAtlas repository of observed and validated spectra), scientific literature (e.g. MEDLINE) and genomic information (e.g. derived from studies such as The Cancer Genome Atlas). We use information theory and network inference approaches to integrate this data. Tools based on these approaches and data have been built to support: identification of assayable biomarkers through protein-disease associations (mspecLINE); visual analysis of proteomic information within a genomic context (CircAtlas). These tools use common open standards including Google Data Source, BioMart and caGrid, to support extensibility and interoperability

CircAtlas supports the mining of PeptideAtlas through the overlay of genomic information and concordance between difference feature types. The tool is being used provide an integrated view of data from TCGA which includes gene expression, copy number variance, methylation, full genome sequencing, and structural variant analyses. Concordance between features are derived through a variety of analytical techniques including random forest, bayesian network, and networks inferred using mutual information. These networks can also be visualized in Cytoscape and enriched using information drawn from PeptideAtlas. This allows for the association of protein detectability with clinically relevant genomic information.

mspecLINE combines knowledge about human disease from MEDLINE with empirical data about the detectable human proteome from spectral libraries. The mspecLINE tool allows researchers to explore relationships between human diseases and parts of the proteome that are detectable using current instrumentation. Given a disease, the tool will find proteins and peptides from PeptideAtlas that may be associated, and display relevant information from MEDLINE. These associations can be visually explored, and the results exported to the experiment design pipeline ATAQS, allowing for the development of disease-specific experiments. mspecLINE associates diseases with proteins by calculating the semantic distance between annotated terms from a controlled biomedical vocabulary. We use an established semantic distance measure that is based on the co-occurrence of disease and protein terms in the MEDLINE bibliographic database. Both mspecLINE and CircAtlas are available as web applications through the PeptideAtlas website.

Methods for understanding the temporal dynamics of the microbiome

Devin C. Koestler, Juliette C. Madan, Margaret R. Karagas, and Jason H. Moore

Pyrosequencing of 16S ribosomal DNA (rDNA) has provided a glimpse into the complexity of the human microbiome. This work has established the microbiome as dynamic and has suggested a critical role of the microbiome in modulating states of human health and disease. Most investigations of the human microbiome have been cross-sectional in nature and are typically focused on the study of single biological system (i.e., intestines, lung, skin, ect.). While there exist several well developed methodologies for analyzing data arising from such designs, statistical methods for investigating more complex study designs, such as those involving repeated-measures longitudinal collection of samples or studies involving the collection of samples from more than one biological system, are critical for obtaining a more comprehensive understanding of the dynamics of the microbiome and synergy of the microbiome across biological systems

To explore the temporal dynamics of microbial diversity, we propose using linear mixed effects models that model a transformed version of the Simpson's diversity index (SDI) as a function of time. Additionally, we utilize a generalized linear mixed effects model (GLMM) framework for examining temporal patterns in microbial richness and the extent to which these patterns are modified based on subject-specific clinical information. Lastly, we propose using a recursively partitioned mixture model (RPMM), a model-based hierarchical clustering methodology, for identifying microbiome profiles that associate with subject-specific clinical information

These methods were applied to a prospective longitudinal study of 7 infants with cystic fibrosis (CF). Respiratory (oropharyngeal sampling) and gastrointestinal (stool) samples were collected at <1 month post-birth and at 3 month intervals serially through 21 months and subsequently subjected to high-throughput pyrosequencing of the v4-v6 hypervariable region of 16S rDNA. The results from linear mixed-effects modeling demonstrated that while the diversity of both the respiratory and gastrointestinal microbiomes were increasing significantly over time, the diversity of the respiratory microbiome was increasing at a faster rate. In addition, there were 42 and 23 genera out of the 93 considered whose abundance changed significantly over time in the respiratory and gastrointestinal tracts, respectively. RPMM clustering of respiratory samples only identified microbiome profiles that were significantly associated with breastfeeding and H2-blocker therapy. A similar analysis applied to the intestinal microbiome samples only revealed microbiome profiles that were significantly associated with whether or not the subject was consuming solid-food

The methods utilized here represent a promising first step toward understanding the temporal dynamics of the microbiome.

MetRxn: A standardized knowledgebase of metabolite and reaction information spanning metabolic models and databases

Akhil Kumar & Costas D. Maranas

Penn State

MetRxn is a web based resource that integrates and standardizes metabolite and reaction descriptions by integrating information from BRENDA, KEGG, MetaCyc, HMDB and 44 organism-specific metabolic models into a single unified data set. All metabolite entries have matched synonyms, resolved protonation states and are linked to unique structures. All reaction entries are elementally and charge balanced. This is accomplished through the use of a workflow of lexicographic, phonetic, and structural comparison algorithms. MetRxn allows for the download of standardized versions of existing genome-scale metabolic models and the use of metabolic information for the rapid reconstruction of new ones. The standardization in description allows for the direct comparison of the metabolite and reaction content between metabolic models and databases and the exhaustive prospecting of pathways for bio-production. This data-set currently consists of 42540 distinct metabolites participating in 35473 reactions from all domains of life

This data-set was constructed by incorporating reaction and metabolite information from well curated databases and metabolic models representing organisms from the various kingdoms. We update the number of organisms-specific models periodically and the total now stands at 44. The curation and standardization process starts with using atomistic details for matching metabolite entries across all the databases. Atomistic details and protonation states for each metabolite are established at a constant pH of 7.2 . We subsequently use metabolites information to construct standardized atomistic descriptions for reactions. An id is provided for each uniquely identifiable entry based on atomistic details. This information is then used to identify, match and standardize information from metabolic models using various lexicographic, phonetic, and structural comparison algorithms. This standardized data-set is hosted on a MySQL database and can be accessed through a web interface at metrxn.che.psu.edu. New features currently been incorporated in the web resource include tools for simultaneous comparisons across multiple models, finding prospective pathways between the substrates and products, data export formats in SBML and BioPax. Web-services and API's allow advanced users to access and integrate MetRxn's data-set into their software and existing workflow. These Web-services are based on the same technology as provided by KEGG and BRENDA, thus allowing for a language independent (e.g., Java, Python etc.) access and seamless integration into existing software. A web-service based access allows users with minimal programming experience to access the database directly using tools such as Taverna, Galaxy etc. MetRxn also allows upload of metabolite and reaction information, while providing real-time curation and standardization capabilities. MetRxn is complementary to resource such as KEGG, BRENDA and METACYC by providing new tools for organism comparisons, pathway elucidation, reaction balancing and integration with strain design tools.

Classification of adenosine receptor antagonists using Laplacian-modified naïve Bayesian, support vector machine, and recursive partitioning

Jin Hee Lee, Sunkyung Lee, 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

Adenosine receptors (ARs) belong to the superfamily of G-protein-coupled receptors (GPCRs), and consist of four subtypes: A1, A2A, A2B, and A3. Having over 60% of sequence similarity among subtypes, ARs mediate many biological effects. These interactions between each AR and its ligands offer very broad therapeutic potential, thus the development of potent and selective synthetic modulators of ARs is important for individual therapy. AR antagonists have greater structure diversity with a variety of scaffolds such as xanthine, adenosine, or polyheterocycle. In addition, the search for selective antagonists is appealing not only for their potential therapeutic applications but also being considered as preferred molecular probes for pharmacological characterization of receptors.

In order to build reliable in silico classification models of AR antagonists, we adopted six descriptors and performed three machine learning methods: Laplacian-modified naïve Bayesian, Recursive Partitioning, and Support Vector Machine. The resulting classification models of the antagonists for each AR had excellent results that yielded high accuracy, sensitivity, specificity, area under the receiver operating characteristic (ROC) curve and Matthews correlation coefficient (MCC) values. The application of representative antagonists to these classification models for each AR demonstrated the power and utility of these models. These models could be further utilized in the prediction of potential AR antagonists in drug discovery.

Modeling Spatial Variation of Bladder Cancer Prognosis and Travel Time to Treatment Facility using Geographically Weighted Regression (GWR)

Kevin M Mwenda, Xun Shi

Dartmouth College

Tracy L Onega, Margaret R Karagas, Jason H Moore, Angeline S Andrew

Dartmouth Hitchcock Medical Center

Northern New England is known to have one of the highest bladder cancer incidence and mortality rates in the U.S. A high number of patients diagnosed with non-invasive bladder cancer experience recurrences just a few years after diagnosis. We are interested in understanding whether there is a spatial variation in the correlation between travel time to treatment facility and recurrence time within non-invasive bladder cancer cases in New Hampshire. We hypothesize that traveling farther to the treatment facility will be significantly associated with a shorter time to the first recurrence within non-invasive bladder cancer cases in rural areas of New Hampshire. Given the exact location of non-invasive bladder cancer cases in New Hampshire and their treatment facilities, among other dependent variables, we would create an origin-destination matrix to determine the total time taken for each of the cases to drive to their treatment facilities. We would then model the spatial variation of the relationship between the time to first recurrence of bladder cancer as the outcome and travel time to treatment facility using geographically weighted regression analysis – prior to that, we would determine the best set of predictors using cox regression. In addition to travel time to treatment facility, we would incorporate explanatory variables such as age, gender, smoking status, rural-urban classification codes and treatment combinations to account for the locality of the spatial relationships. Finally, we would perform error testing to validate our model. The outcome of our analysis would be a regression map and associative statistics of the spatial variation of the relationship, if any, between bladder cancer recurrence time and travel time to treatment facilities in New Hampshire. We hope that our study highlights the usefulness of GWR in providing a Geo-spatial and visualization component when modeling complex interactions in public health research, especially within the realm of epidemiology.

The Use of Sequence Data to Predict the Risk of Immunological Adverse Events in Cellular and Solid-Organ Transplantation

Octavio E. Pajaro, Francisco A. Arabia

Mayo Clinic Arizona

This study presents preliminary results of an approach to use amino acid sequence data of the human leukocyte antigens (HLA) to predict clinical outcomes measures in cellular and solid-organ transplant. Outcomes in cellular and solid-organ transplantation depend highly on the ability to minimize the immunological risk to the recipient--a risk determined primarily by differences in donor and recipient HLA genetic mismatches. In bone marrow transplantation, 75% of patients do not have a suitable HLA-matched sibling and depend on alternative donors – umbilical cord blood or haploidentical hematopoietic cell transplantation (Ballen and Spitzer 2011). Unfortunately, both methods carry a risk of graft versus host disease (GVHD) (Petersdorf et al., 2001). Furthermore, haploidentical cell transplants, although attractive because of the ease of finding donors and its cost effectiveness, carry a significantly greater risk of severe potentially fatal GVHD. In solid-organ transplantation, the risk of immune-mediated complications is present in the acute and chronic phases and ultimately determines short and long term survival.

The immunological trigger for the allograft cell-mediated immune response is initiated by the binding of a donor HLA molecule with a recipient T cell receptor (TCR). The highly polymorphic HLA region that binds to the TCR is modeled as a vector with the following properties: 1) The vector is oriented at the center of the HLA molecule's TCR binding region. 2) The magnitude and orientation of the vector is determined by the biophysical properties of the amino acids in the HLA-TCR binding region. Thus, a vector representing the biophysical properties of the HLA binding region represents each HLA allele. The underlying assumptions are that 1) the recipient HLA vectors reflect the orientation of the recipient TCR as it binds to the HLA molecule, 2) the donor HLA vector will reflect how well the recipient TCR can bind to the HLA molecule in its native orientation and 3) changing the orientation of the TCR binding to the HLA from its native orientation determines immunogenicity of the donor HLA allele. Molecular coordinates for the class I and class II alleles were obtained from the protein databank available from NCBI. The amino acid sequences for each HLA allele were obtained from the immunogenetics-HLA database available from the European Bioinformatics Institute (IMGT/HLA EBI). Data will be presented showing the ability of this approach to predict clinical outcomes in both cellular and solid-organ transplantation. This model represents the first approach to extrapolate potential biophysical properties from sequence data and translate the results into clinical outcomes.

Ballen, K. K. and T. R. Spitzer (2011). "The great debate: haploidentical or cord blood transplant." Bone marrow transplantation 46(3): 323-329

Petersdorf, E. W., J. A. Hansen, et al. (2001). "Major-histocompatibility-omplex class I alleles and antigens in hematopoietic-cell transplantation." N Engl J Med 345(25: 1794-1800.

ODIN: Advanced Text Mining in Support of the Curation Process

Fabio Rinaldi, Simon Clematide, Gerold Schneider

Institute of Computational Linguistics
University of Zurich, Switzerland

In this poster we describe ODIN (OntoGene Document Inspector), an interactive tool which supports the process of curation of the biomedical literature through integration of advanced text mining techniques, coupled with an intuitive user interface

ODIN has been developed within the scope of the SASEBio project (Semi-Automated Semantic Enrichment of the Biomedical Literature, funded by the Swiss National Science Foundation), as a collaboration between the OntoGene group at the University of Zurich and the NITAS/TMS group of Novartis Pharma AG

ODIN allows a human annotator/curator to make effective use of the results of an advanced text mining system in order to enhance the speed and effectiveness of the annotation process

The input documents (e.g. PubMed abstracts or PubMed Central full papers) are converted into an internal XML format and processed by OntoGene's NLP pipeline. Predefined entity types can be recognized by the system on the basis of terminology derived from reference databases (e.g. proteins from UniProt, genes from EntrezGene, species from the NCBI taxonomy). Additional entity sets can be easily added as needed, as we have done in recent experiments using entities derived from PharmGKB, thus including drugs and diseases

The system will disambiguate entity mentions as far as possible by providing references to database identifiers. Remaining ambiguities can be easily solved manually by the curator. All entity mentions are entirely editable: the curator can easily add or delete any of them, and also change their extent (i.e. add/remove words to its right or left) with a simple click of the mouse. Different entity views are supported, with sorting capabilities according to different criteria (entity type, entity mention, confidence score, etc.). Selective highlighting of text units (e.g. sentences containing desired entities) is supported. Additionally, extensive logging functionalities are provided. All documents and entities are fully interlinked to reference databases, for the purpose of simplified inspection. Entities can be grouped in classes (e.g. by species) and actions can be applied to whole classes, for selective editing or removal

All possible entity combinations are considered as candidate interactions, which are then scored taking into account properties of the individual entities (e.g. frequency of mention) as well as several features derived from the automated syntactic analysis of the documents

The annotated documents are handed back to the ODIN interface, which allows multiple display modalities. The curator/annotator can view the whole document with in-line annotations highlighted, or can browse candidate interactions in a separate panel, where they are presented to the user sorted according to an optimized ranking. Selection of each candidate interaction automatically highlights the original mentions in the document, thus allowing a quick decision concerning its validity

Additionally, we present results from a recent experiments in collaboration with the PharmGKB group at Stanford University, which prove the effectiveness of the ODIN system in a real curation setting. We show that the text mining component can deliver an effective ranking of candidate relationships, and that the system as a whole can enhance the productivity of the curators.

A gene-set based approach to link genomic variants to human disease using KEGG database

Heewon Seo, Jihun Kim, Yonglae Cho, Youngjo Yoon, and Ju Han Kim

Seoul National University Biomedical Informatics (SNUBI), Div. of Biomedical Informatics,
Seoul National University College of Medicine, Seoul 110-799, Korea

The availability of large genomic variants data provides a new opportunity to study the association between personal variant and susceptible disease. In order to provide a list of high risky diseases to individual, we found associating diseases to biological pathways where disease genes are enriched from KEGG pathway and KEGG disease. Assuming that disease-associated genes are affected by SNVs detected from individual and we annotated those polymorphisms using knowledgebase. First, we found 9,761 nsSNPs(non-sysnonymous SNP) from the sample and applied SIFT and PolyPhen2 prediction algorithms. Second, we listed 611 nsSNPs that the amino-acid is predicted damaging in both algorithms and then matched 477 rsIDs. Third, a given set of rsIDs was mapped to official gene symbols based on their landmark. Forth, we found a set of 405 genes and found enriched pathways. Finally, we found disease list from the KEGG Disease which is highly associated to KEGG pathway. As a result, we came across with 18 highly susceptible diseases to the sample provider. Our results could provide the preliminary way to link a set of variants to human disease with pathway.

Cross-talks in the recovery and failure of distal organs in severely injured patients

Junhee Seok, Lu Tian, Ronald Maier, Ronal Davis, Ronald Tompkins, Wing Wong and Wenzhong Xiao

Cross-talks of distal organs underlie multiple organ failures (MOF), which is a major cause of deaths of severely injured patients in intensive care units. Although previous studies suggested interactions among organs in MOF, system-level cross-talks remain poorly understood. Here, we studied the cross-talks at body system level by investigating the times of all major organs after severe traumatic injury. The Inflammation and host response to injury glue grant consortium has collected time-course information of organ failures of 1,875 severe trauma patients from 2001 to 2009. Using this large data set, we analyzed the recovery and failure relapse times of six major organs by applying multivariate survival analysis techniques. We successfully estimated the joint probability distribution of recovery and failure relapse between all possible pairs of two organs, and determined the precedency and dependency among organs. In more than 55% of patients, hematologic and hepatic system were recovered first within a week after injury, then respiratory system followed by neurological and renal system, and finally cardiovascular system. Similarly strong precedencies in organ failure relapse are observed, where hematologic and cardiovascular system often fail first. The analysis for organ recovery and failure relapse times suggested potential cross-talks among distal organs after injury. For example, the recovery and failure of respiratory system precede those of renal system, which might indicate the dependency of the renal system on the respiratory system. The overall analysis results promote a system-level view of organ cross-talks after injury.

PLATO: PLatform for the Analysis, Translation, and Organization of large-scale data

Shefali Setia, Benjamin J. Grady, John R. Wallace, Scott M. Dudek. Gretta J. Armstrong, Marylyn D. Ritchie

Genome-wide association studies (GWAS) are being conducted at an unprecedented rate in population-based cohorts and have increased our understanding of complex disease. PLATO, the PLatform for the Analysis, Translation, and Organization of large-scale data, is a filter-based method designed to bring together many statistical and data mining methods simultaneously in order to analyze gene-gene interactions between variants in GWAS datasets. In addition to analysis methods, PLATO includes filtering methods that provide a mechanism to reduce the number of SNPs to a smaller subset for analysis. Finally, PLATO includes features addressing issues associated with quality control (QC) of GWAS data including data file formats, software packages for data manipulation and analysis, sex chromosome anomalies, sample identity, sample relatedness, population substructure, batch effects, and marker quality. PLATO allows for significant flexibility when applying filters in series, parallel, or individually; it allows the specification of filters for different disease models and can be used for tests for association. Furthermore, PLATO is extensible, allowing users to easily implement their own analytical methods using a modular C++ library. Additionally, PLATO can identify multiple disease susceptibility models that may predict disease risk based on different underlying genetic models. This approach will lead to an increasing ability to identify genetic associations, thereby enhancing our understanding of complex traits.

The motivation for PLATO is twofold:

First, any single underlying analytical scheme reveals only some important results and multiple filters reveal different subsets of important results. Once results are obtained they can be viewed in light of the results from other filters in order to best understand the implications of the patterns in the genetic data. The potential to use multiple filters does not force a priori assumptions about the mode of action of the genetic components of a phenotype allowing the most general possible analysis strategy and interpretation. This is critical as it is rare that we know what type of effect we are attempting to detect in disease gene association studies. PLATO gives the ability to evaluate the association in the context of many different models and select the optimum solution for the dataset at hand, while controlling Type I error rate through permutation testing.

Second, it is hypothesized that the genetic architecture of complex disease includes interactions between many genes as well as the environment. In GWAS datasets, searching for interactions is a computational challenge and thus filtering the full set of GWAS SNPs to a smaller subset is critical in the when searching for the presence of interactions. PLATO accomplishes both of these goals

Flexible, modular analysis methods like PLATO provide a framework for thorough genetic association analysis. The underlying goal is to improve our understanding of architecture of complex traits.

A probabilistic fragment-based protein structure prediction algorithm

David Simoncini, Francois Berenger, Rojan Shrestha, Taeho Jo and Kam Y. J. Zhang

Advanced Science Institute, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

Fragment-based approach has been shown to be an efficient method to solve the protein structure prediction problem. In this two-phase strategy, the coarse-grained modeling, which predicts the backbone of the decoys, is critical. However, The size of the conformational search space makes the sampling algorithms unlikely to focus on the region where the native protein structure belongs. We propose a new algorithm, EdaFold, which can enhance the generation of near-native protein models during the coarse-grained prediction phase of fragment-based approaches. Our method uses the principle from the Estimation of Distributions Algorithms to define probability mass functions over the databases used by fragment-based approaches. For fair comparison, we use the same energy function as Rosetta and show that EdaFold can generate models with lower energy and on average closer to the native protein structure on a benchmark of 20 proteins.

Profiling of histone modifications in normal development and MLL-rearranged leukemia

Amit U Sinha 1, Scott A Armstrong 1,2

1Division of Hematology/Oncology, Children's Hospital Boston and Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA
2Harvard Stem Cell Institute, Boston, MA

Epigenetic changes not only have a critical role in normal development but also have been implicated in cancer. Innovation in sequencing technologies has allowed us to study epigenetics on a genome wide level. We performed a comprehensive analysis of H3K4me3, H3K27me3, H3K36me3 and H3K79me2 histone modifications in hematopoietic stem cells, granulocyte macrophage progenitors, and leukemia stem cells arising from leukemias generated by an MLL-AF9 fusion oncogene in a mouse model. We observed that each modification is uniquely associated with changes in gene expression and has a unique morphology near the gene. The histone modifications share a unique relationship with each other and these relationships are maintained on a global level in all cell types. Consistent with previous reports, H3K79me2 alone showed aberrant high levels at a small group of genes that were the targets of the MLL-AF9 fusion oncogene. Knock-out of Dot1l, the only known methyltransferase of H3K79me2, in mouse models leads to a decrease in expression of MLL fusion target genes and increased survival.

Mycobacterial whole genome resequencing, network analysis, and structural analysis

Michael Strong 1, Eve Farias-Hesson 1, Paul Reynolds 1, Daniel LaFlamme 1, Rebecca Davidson 1, Preveen Ramamoorthy 1, Sarah Totten 1, Vitus Dzekedze 2, Kimberly Harbach 1, Nick Walter 2, and Charles Daley 1

1. National Jewish Health, Denver, CO
2. University of Colorado, Denver, CO

Mycobacterial infections are responsible for a variety of respiratory infections, ranging from tuberculosis (TB), which kills approximately 1.6 million people each year, to other non-tuberculous mycobacterial (NTM) infections. In order to garner a better understanding of mycobacterial pathogens at the genomic level, we have undertaken a multifaceted approach to investigate clinically relevant M. tuberculosis and non-tuberculous mycobacterial (NTM) strains, using a combination of whole genome sequencing, network inference and analysis, gene expression analysis, and structural bioinformatics. The goal of this integrated approach is to garner a better understanding of the genetic and molecular mechanisms contributing to important clinical phenotypic characteristics ranging from pathogenesis to drug resistance. Among our experiments, we have performed whole genome resequencing of approximately twenty NTM and TB isolates. We have also performed gene targeted resequencing of M. tuberculosis strains to determine the molecular determinants of fluoroquinolone and pyrazinamide resistance in M. tuberculosis strains, and have identified some mutations that appear to confer resistance to certain fluoroquinolones but that remain susceptible to related fluoroquinolones. Combined with network analysis and structural modeling, we hope to not only garner a better understanding of the molecular mechanisms of drug resistance and virulence, but also to identify novel putative targets for new intervention strategies.

Random Forest Multi-Task Learning For Predicting HIV-1, Human Interactions

Oznur Tastan
Microsoft Research New England Cambridge, MA 01239, USA
Email: oznur@cs.cmu.edu

Jaime Carbonell
Language Technologies Institute, Carnegie Mellon University Pittsburgh, PA 15213, USA
Email: jgc@cs.cmu.edu

Judith Klein-Seetharaman
Research Institute Juelich, 52425 Juelich, Germany / Department of Structural Biology, University of Pittsburgh, Pittsburgh, PA 15260, USA
Email: judithks@cs.cmu.edu

We have presented at PSB a supervised model for the task of predicting HIV-1,human protein-protein interactions using the Random Forest classification algorithm [1]. In this formulation, all viral proteins' interactions were pooled together and a single model was trained to predict the host-virus interactome. However, as different viral proteins undertake different functions in the viral replication cycle, the properties of their interactions may vary with each specific viral protein, and are thus drawn from different distributions. The negative effects of neglecting this issue on the prediction can empirically be seen in the predictions, where the highest scoring predictions are those relating to Tat due to its overrepresentation in the training data. Pooling the training data disregards the differences between viral proteins and biases the predictions. To overcome this shortcoming, we here present an alternative approach where we learn different models for each viral protein. The small sample size for the interactions of each of the viral proteins, however, impedes the construction of separate models. We overcome this data scarcity issue through a multi-tasking learning strategy. In this model, a separate learning task was defined for each of the viral proteins, but all tasks shared their training data proportional to their task relatedness. We defined relatedness according to the overall biological functions of the viral proteins. To implement the multi-task approach, we modified the bootstrapping step of the Random Forest classifier. This modification leads to more accurate predictions for 8 of the 10 HIV-1 viral proteins as compared to those derived from the single model learned for the HIV-1, host protein interactions. Thus, considering the differences in data distribution across different viral proteins improves the accuracy of the resulting interactome

1. O. Tastan, Y. Qi, J. G. Carbonell, J. Klein-Seetharaman, Pacific Symposium on Biocomputing 2009, 516 (2009).

Using Genetic Algorithms for the Inference of Motifs That Are Represented In Only a Subset of Sequences of Interest

Jeffrey A. Thompson, Clare Bates Congdon

Department of Computer Science, University of Southern Maine

OVERVIEW: In this work, we present GAMID, an extension of GAMI. GAMID is designed to be used for motif inference in noncoding DNA for co-expressed genes or for divergent species. In these cases, we would like to allow the inferred motif to be present in only a subset of the input data. This paper describes the approach and presents preliminary results

BACKGROUND: GAMI (Genetic Algorithms for Motif Inference) uses a Genetic Algorithms search to identify putative functional elements in noncoding DNA. The system was designed to identify putative functional elements following the notion that elements that have been conserved across evolution are more likely to be functional, and therefore, seeks to find highly conserved motifs in the data. In previous work, GAMI has been shown to be adept at finding highly conserved elements in long sequence lengths (e.g., 100kb) and across several dozen sequences

A limitation of GAMI is that it seeks to find evidence of the motif in all the sequences of the data, and this is not appropriate when investigating co-regulated genes. In this case, a functional element might appear in only a subset of the sequences. Similary, in a dataset of highly divergent species, a functional element might appear in closely related species, but not appear in all sequences in the dataset

The aim of this work is to extend GAMI with the capability of selectively ignoring sequences when evaluating certain motifs, if it seems likely that doing so will be beneficial in identifying a putative functional element. We refer to this process as “dropout”. It is our hypothesis that this will allow GAMI to determine motifs that exist in only a subset of the input sequences and that we will therefore be able to study a more complete picture of the functional areas in regulatory sequences. This information should be useful to researchers performing analyses of regulatory regions, since our approach can show the affinity of motifs for subsets of genes, which may help in understanding differential gene regulation in both orthologous and co-expressed genes

PRELIMINARY RESULTS: We refer to the extended GAMI as GAMI with Dropout (GAMID). We ran both GAMI and GAMID with DNA sequences from promoters of genes co-expressed in pregnancy, and promoters of genes regulated by the NF-KappaB transcription factor. For both data sets we knew of some putative and experimentally validated transcription factor binding sites (TFBSs) that were present in subsets of the sequences that we used to evaluate the performance of GAMID. These initial results support our hypothesis that dropout is an effective technique. GAMID appears to be better able to identify sparsely distributed motifs and, in the case of known TFBSs, the gene subsets regulated by them.

HLA class I Predictions by Targeted Assembly of NGS Shotgun Reads

Rene L Warren, Gina Choe, Mauro Castellarin, Sarah Munro, Richard Moore, Robert A Holt

The Human Leukocyte Antigen (HLA) genomic region comprises genes that encode cell surface proteins that present peptide antigens to T cells. Due to its importance in discriminating, at the cellular level, self from non-self, HLA is key to many aspects of human physiology and medicine, including organ transplantation, autoimmunity, vaccination, infectious disease and cancer

We are developing a computational method for HLA class I allele prediction directly from next-generation sequence (NGS) data sets. Unlike existing approaches, all of which restrict information by amplifying specific HLA sub-regions, our proposed approach of mining HLA information retroactively from NGS shotgun data is value-added, that is, it does not incur any additional time or cost for generating HLA-specific data when a shotgun data set is already available. The approach, named HPTASR, relies on targeted de novo assembly of NGS sequence reads as its cornerstone algorithm for assembly of HLA allele sequences from relatively low coverage NGS data. Resulting sequence contigs and known, public HLA-I reference allele sequences are aligned against each other and candidate HLA alleles are identified, scored, and their probability evaluated. Using this method, we predicted and validated HLA alleles for 16 colorectal cancer RNA-seq and 3 ovarian cancer exon capture patient samples. Traditional PCR-based HLA typing using corresponding gDNA as template was used to corroborate HPTASR predictions. From low-coverage RNA-seq data sets (~9.8M reads per subjects), we obtained both a detection rate and accuracy of 100% at the two-digit allele group resolution (e.g. A*01) and 97% at the four-digit allele resolution (e.g. A*01:01), respectively. Exon capture data gave a lower performance, offering a detection rate and accuracy of 89% and 94% for allele groups and 83% and 89% for HLA alleles, respectively. We discuss how different NGS data types affect HPTASR predictions

Computational HLA predictions by localized assembly of NGS shotgun read data is expected to have broad applications in research and clinical settings as the already widespread sequencing of whole genomes, exomes and transcriptomes continues to increase.

Determination and Inference of Transcription Factor DNA Sequence Specificity Across Eukarya

Matthew T. Weirauch 1, Ally Yang 1, Hong Zheng 1, Harm van Bakel 1, Atina Cote 1, Ishminder K. Mann 1, and Timothy R. Hughes 1,2

1Banting and Best Department of Medical Research
2Department of Molecular Genetics, University of Toronto

Knowledge of the sequence specificities of transcription factors (TFs) is central to the understanding of gene regulation and genome function. Yet the sequence specificities of most transcription factors (TFs) in most eukaryotes remain unknown. It is commonly assumed that proteins with similar DNA-binding domains (DBDs) will possess similar specificity, but to our knowledge the accuracy and limitations of this approach have not been rigorously examined, hindering its broad application. Moreover, current TF sequence specificity data is heterogeneous

Protein Binding Microarrays provide a rapid and uniform method for determining TF sequence specificity. Using an aggregate of several hundred published and unpublished PBM experiments, we find that similarity in the DBD typically correlates well with similarity in DNA sequence preference. Often there appears to be a distinct amino acid identity threshold for individual DBD types, above which two proteins will have virtually identical DNA-binding sequence specificity. Below this threshold, it is difficult to predict systematically what sequences new proteins will bind

This simple observation suggests a scheme for broadly characterizing DBD sequence specificities across the eukaryotic kingdom: we should prioritize “hubs” in the “network” of DBD identity, with the goal of ensuring that all DBDs are represented by an example within the amino acid similarity threshold for each class, and with a sufficiently dense distribution to allow for cross-validation of inferred sequence preferences. We have recently begun to execute such a scheme, prioritizing DBDs using a strategy that (a) favors DBDs with the largest number of sequence “neighbors”, (b) aims to broadly survey all major branches of eukarya, (c) is designed to examine a spectrum of amino-acid sequence identities among DBD amino acid sequences. After analysis of only a few hundred carefully-chosen DBDs, we now present evidence that the sequence specificity of nearly half of all eukaryotic DBDs are known or can be inferred.

Resolution-by-Proxy: A Simple Measure for Assessing and Comparing the Overall Quality of NMR Protein Structures

David Wishart, Mark Berjanskii, Jianjun Zhou, Yongjie Liang, Guohui Lin

Dept. of Computing Science, University of Alberta, Edmonton, AB, Canada

In X-ray crystallography, resolution is often used as a good indicator of structural quality. Crystallographic resolution is related directly to the number of X-ray experimental observables and correlates well with coordinate errors of protein structures. In protein NMR, there is no equivalent of X-ray resolution. Instead, the precision of a NMR structural ensemble is commonly used as a "resolution surrogate" to decide if the ensemble should be considered a high-, medium- or low-resolution structure. However, ensemble precision has different properties and meaning than X-ray resolution. The lack of common measurement techniques to assess the resolution or coordinate uncertainty of X-ray and NMR structures has made it difficult to compare structures solved by these two different methods. This problem is sometimes approached by calculating "equivalent resolution" from various protein structure quality metrics. However, existing protocols are not very robust as they typically calculate equivalent resolution from a small number (<5) of protein parameters. In an effort to improve the situation, we have developed a protocol to calculate equivalent resolution from 25 protein features. The new method demonstrates significantly better performance (correlation coefficient of 0.92, mean absolute error of 0.28 Å) than any existing predictors of equivalent resolution Because the method uses coordinate data as a proxy for X-ray diffraction data, we call this measure "Resolution-by-proxy" or ResProx. Using an extensive array of tests, we also demonstrate that ResProx exhibits all of the characteristic features expected of a robust resolution function. Furthermore, we are able to demonstrate that ResProx can be used to identify under-restrained, poorly refined or incorrectly modeled NMR structures and can discover structural defects that the other equivalent resolution methods can't detect. We have developed a web server, called "Resprox", to calculate equivalent resolution values for NMR structures using only PDB coordinate data as input. It is available at http://www.resprox.ca.

MetaDomain: a profile HMM-based protein domain classification tool for short sequences

Yuan Zhang and Yanni Sun

Michigan State University

Protein homology search provides basis for functional profiling in metagenomic annotation. Profile HMM-based methods classify reads into annotated protein domain families and can achieve better sensitivity for remote protein homology search than pairwise sequence alignment. However, their sensitivity deteriorates with the decrease of read length. As a result, a large number of short reads cannot be classified into their native domain families. In this work, we introduce MetaDomain, a protein domain classification tool designed for short reads generated by next-generation sequencing technologies. MetaDomain uses relaxed position-specific score thresholds to align more reads to a profile HMM while using the distribution of alignment positions as an additional constraint to control false positive matches. In this work MetaDomain is applied to the transcriptomic data of a bacterial genome and a soil metagenomic data set. The experimental results show that it can achieve better sensitivity than the state-of-the-art profile HMM alignment tool in identifying encoded domains from short sequences.

IDENTIFICATION OF ABERRANT PATHWAY AND NETWORK ACTIVITY FROM HIGH-THROUGHPUT DATA

A BIOLOGICALLY INFORMED METHOD FOR DETECTING ASSOCIATIONS WITH RARE VARIANTS

Carrie C. Buchanan
Center for Human Genetics Research, Vanderbilt University, 519 Light Hall Nashville, TN 37232, USA Email: carrie.c.buchanan@vanderbilt.edu

John Wallace
Center for Systems Genomics, Penn State University, 512 Wartik University Park, PA 16802, USA Email: jrw32@psu.edu

Alex Frase
Center for Systems Genomics, Penn State University, 512 Wartik University Park, PA 16802, USA Email: atf3@psu.edu

Eric S. Torstenson
Center for Human Genetics Research, Vanderbilt University, 519 Light Hall Nashville, TN 37232, USA Email: torstenson@chgr.mc.vanderbilt.edu

Marylyn D. Ritchie
Center for Systems Genomics, Penn State University, 512 Wartik University Park, PA 16802, USA Email: marylyn.ritchie@psu.edu

With the recent flood of genome sequence data, there has been increasing interest in rare variants and methods to detect their association to disease. Many of these methods are based on collapsing strategies, which bin variants based on allele frequency and functional association. To date, most previously published methods have been limited to candidate gene studies. We propose a novel method to collapse rare variants based on incorporating biological information. This is a useful collapsing strategy because it can be expanded to whole-genome data and can apply multiple levels of burden testing, including: functional regions, genes, and/or pathways. It can be used as a framework to identify gene-gene (GxG) or gene-environment (GxE) interactions. Also there is the potential to integrate large complex data sets (i.e. expression profiles, metabolomics, etc) with sequence data into future analyses. The focus of this presentation is to introduce the functionality of BioBin, a biologically informed method to collapse rare variants and detect associations with a particular phenotype. BioBin has been tested using low coverage data from the 1000 Genomes Project, targeted exome data from the 1000 Genomes Project, and in simulated data. It has demonstrated promising utility in a head-to-head comparison with other popular collapsing methods. Although BioBin is still in developmental stages, it will be a useful and very flexible tool in analyzing sequence data and uncovering novel associations with complex disease.

An integrated approach to infer transcription factor and microRNA regulatory networks

Li-Wei Chang, Andreu Viader, Jacqueline E. Payton, Nobish Varghese, Jeffrey Milbrandt, Rakesh Nagarajan

Washington University

Regulation of cell fate and developmental timing as well as maintenance of cellular and tissue homeostasis requires precise spatiotemporal control of gene expression, which is modulated by key transcription factors (TF) and microRNAs (miRNA). Elucidation of TF and miRNA regulatory networks is crucial to understand how normal gene regulatory mechanisms are disrupted in human disease. Computational methods have been developed to infer TF or miRNA mediated gene regulation, but few methods integrate these results to construct integrated TF and miRNA gene regulatory networks. In this study we present a systematic approach, IntegraNet, which combines mRNA and miRNA expression data, chromatin immunoprecipitation with sequencing (ChIP-Seq) data and computation TF and miRNA target prediction. We applied IntegraNet to expand a previously published miRNA-mRNA regulatory network as well as to infer the TF and miRNA regulatory network involved in Schwann cell (SC) response to nerve injury. Further analysis of regulatory network motifs in this SC injury response network provided insight on cooperative regulation of this process by TFs and miRNAs. This work demonstrates a systematic method for gene regulatory network inference that may be used to gain new insights on local and global features of gene regulation by TFs and miRNAs.

Concordant Gene Set Enrichment Analysis of Two Large-Scale Expression Data Sets

Yinglei Lai

The George Washington University

The recent large-scale technologies like microarrays and RNA-seq allow us to collect genome-wide expression profiles for biomedical studies. Genes showing significant differential expression are potentially important biomarkers. A gene set enrichment analysis enables us to identify groups of genes (e.g. pathways) showing coordinate differential expression. Genes and gene sets showing consistent behavior in two related studies can be of great biological interest. However, since the sample sizes are usually small but the numbers of genes are large, it is difficult to identify truly differentially expressed genes and determine whether a gene or a gene set behaves concordantly in two related studies. We have recently shown that the mixture model based approach can be an efficient solution for the concordant analysis of differential expression in two two-sample large-scale expression data sets. The advantage of the mixture model based approach is that the probability of a particular behavior (up-regulated or down-regulated) can be estimated for a given gene. Thus, it is feasible to address how likely this gene shows a concordant behavior in both data sets. In this study, we extend this approach for the concordant gene set enrichment analysis.

Finding perturbed signaling pathways in TCGA ovarian cancers

Songjian Lu and Xinghua Lu

Dept of Biomedical Informatics, University of Pittsburgh

Cancers are caused by somatic mutations that lead to hallmark changes in cellular signaling systems. Large-scale efforts have been devoted to identify somatic and germ-line mutations from a large number of tumor samples, including the Cancer Genome Atlas (TCGA) project and the international network of cancer genomic projects. It is not uncommon that cancer cells accumulate a large number of mutations during development; some are cancer-causing (driver mutations) while others have no relation to cancers (passenger mutations). A major thrust in cancer genomic research is to identify driver mutations and reconstruct perturbed signaling pathways that underlie the hallmark behaviors of cancers. This body of information will shed light on the disease mechanisms of cancers, reveal novel drug targets, and more importantly guide patient treatment based on personal genetic information. In this study, we address the task of revealing perturbed signaling pathways by integrating genomic mutation data with functional genomic data. The main idea underlying our approach is to use differential expression gene modules as the readouts of signaling pathway perturbations, which enable us to reconstruct a signaling pathway by finding the mutations that are strongly associated with a gene expression module. We developed a framework that unifies ontology-guided knowledge mining and graph-based data mining to achieve the goal.

Our methods were able to discover perturbation of many well-known cancer signaling pathways, and we conjecture that some of our results may help to discover novel pathways in cancers.

NetGestalt: Data integration over hierarchically and modularly organized networks

Zhiao Shi 1,2, Jing Wang 3, Bing Zhang 3,4

1 Advanced Computing Center for Research & Education,
2 Department of Electrical Engineering and Computer Science,
3 Department of Biomedical Informatics,
4 Department of Cancer Biology, Vanderbilt University, Nashville, TN 37235, USA

Advancements in high-throughput omics technologies are revolutionizing the field of complex disease studies. As a prime example, The Cancer Genome Atlas (TCGA) project will generate multiple types of genomic data from 500 cases of human cancer for each of the 25 selected tumor types by 2014. Until computational tools are available for biologists and clinicians to independently interpret the vast amount of interconnected data, the potential of this type of large collaborative projects will not be fully realized. To fill the gap between data generation and investigators' ability to interpret the data, we have developed NetGestalt, a web-based application that enables the integration of multidimensional omics data within the context of a protein interaction network.

Because molecular alterations at DNA, RNA, and protein levels exert their effects primarily through changing the activity of proteins and their participating networks, protein-protein interaction network has become a powerful model for the visualization and integration of different types of molecular data. However, the standard graph-based network visualization becomes inadequate as network size and data complexity increase.

We address this challenge through exploiting the inherent hierarchical architecture of the protein interaction network. By using only the horizontal dimension of a webpage to layout genes according to the hierarchical network architecture, it allows users to simultaneously compare and correlate information from experimental data, network modules, and existing knowledge rendered as tracks along the vertical dimension of the webpage, similar to the widely used genome browsers. However, without constraining the system to genomic sequence-based coordinates, NetGestalt is able to reveal functional relationship between different genes as encoded in the network. We also employed efficient software architecture to enable fast track rendering process and smooth navigation between different resolution scales from individual genes to the whole network. The potential of NetGestalt was demonstrated using the recently published TCGA ovarian cancer data with multiple types of genomic measurements on around 500 tumor samples and corresponding normal controls.

In summary, we have developed NetGestalt, a novel data integration framework that allows simultaneous presentation of large scale experimental and annotation data from various sources in the context of a biological network to facilitate data visualization, analysis, interpretation and hypothesis generation. The NetGestalt-based data browser can facilitate biologists and clinicians to translate the vast amount of data into novel biological discoveries and better therapeutics.

Data-driven prediction of drug effects and interactions

Nicholas P. Tatonetti and Russ B. Altman

Stanford University

Adverse drug events remain a leading cause of morbidity and mortality in the United States and around the world. In addition, nearly 30% of investigated drugs fail clinical trials due to unexpected adverse events. Large collections of adverse drug event reports are maintained by the Food and Drug Administration and other organizations. These databases represent an opportunity to study the full range of drug effects. Currently, hypotheses about novel drug side effects are generated through quantitative signal detection. These methods compare the expected reporting frequencies between drugs and side effects to the actual frequencies. However, uncharacterized biases in spontaneous reporting systems, such as prescription bias, patient demographic biases, concomitant drug use, and comorbidities, significantly limit the effectiveness of these algorithms. Here we present a novel data-driven method for correcting for these factors that outperforms existing methods. We present the first database of off-label effects (OFFSIDES) and the first comprehensive database of drug-drug interaction side effects (TWOSIDES). To demonstrate the biological utility of these new databases we replicated a study that used side effect data to predicted novel targets of drugs and used these databases to discover drug class interactions and validated 47 ($p < 0.0001$) class wide interaction effects. We found that patients taking thiazides and selective serotonin reuptake inhibitors together are 1.6-1.9 ($p< 0.0001$) times more likely to develop arrhythmias than patients taking either drug class alone.

INTRINSICALLY DISORDERED PROTEINS: ANALYSIS, PREDICTION AND SIMULATION

Emergent Structure and Function of Disordered Protein Assemblies in the Nuclear Pore Complex

David Ando (UC Merced), Yong Woon Kim (KAIST), Roya Zandi (UC Riverside), Michael Colvin (UC Merced), Michael Rexach (UC Santa Cruz) and Ajay Gopinathan (UC Merced)

The nuclear pore complex (NPC) is an important macromolecular structure that gates the aqueous pores between the cytoplasm and nucleoplasm of cells and controls all nucleo-cytoplasmic transport and communication such as the import of proteins from the cytoplasm and the export of RNA from the nucleus. The NPC forms a barrier that maintains a tight seal against cytoplasmic particles larger than 4 nm while simultaneously allowing the facilitated transport of specially "tagged" particles up to 40 nm in diameter, at speeds comparable to free diffusion! The key to the selectivity is hypothesized to be due to a large number of NPC proteins that fill the pore and potentially interact with each other and the cargo. However, despite numerous studies on the structure and properties of individual NPC proteins, the actual structure of the complex within the nuclear pore and its mechanism of operation are open questions with leading models of nuclear pore transport assuming vastly different morphologies for the NPC protein complex filling the nuclear pore. Here, we use a bottom-up approach, using coarse-grained simulations and applying the physics of polymer brushes to understand the three dimensional architecture of the complex. Our results indicate that individual nucleoporins have a diblock polymer character and that this leads to unique polymer brush morphologies. We also show that there exist transitions between distinct brush morphologies (open and closed states of the gate), which can be triggered by the presence of cargo with specific surface properties. This has led to development of the Discrete Gate Model - an experimental data driven theoretical model. The resulting transport mechanism, that we propose, is fundamentally different from existing models and points to a novel form of gated transport in operation within the nuclear pore complex. Our results can also be extended to designing and optimizing novel forms of biomimetic transport based on this mechanism.

Bringing order to protein disorder through comparative genomics and genetic interactions

Jeremy Bellay, Sangjo Han, Magali Michaut, TaeHyung Kim, Michael Costanzo, Brenda J Andrews, Charles Boone, Gary D Bader, Chad L Myers and Philip M Kim

Intrinsically disordered regions are widespread, especially in proteomes of higher eukaryotes. Recently, protein disorder has been associated with a wide variety of cellular processes and has been implicated in several human diseases. Despite its apparent functional importance, the sheer range of different roles played by protein disorder often makes its exact contribution difficult to interpret. We attempt to better understand the different roles of disorder using a novel analysis that leverages both comparative genomics and genetic interactions. Strikingly, we find that disorder can be partitioned into three biologically distinct phenomena: regions where disorder is conserved but with quickly evolving amino acid sequences (flexible disorder); regions of conserved disorder with also highly conserved amino acid sequences (constrained disorder); and, lastly, non-conserved disorder. Flexible disorder bears many of the characteristics commonly attributed to disorder and is associated with signaling pathways and multi-functionality. Conversely, constrained disorder has markedly different functional attributes and is involved in RNA binding and protein chaperones. Finally, non-conserved disorder lacks clear functional hallmarks based on our analysis. Our new perspective on protein disorder clarifies a variety of previous results by putting them into a systematic framework. Moreover, the clear and distinct functional association of flexible and constrained disorder will allow for new approaches and more specific algorithms for disorder detection in a functional context. Finally, in flexible disordered regions, we demonstrate clear evolutionary selection of protein disorder with little selection on primary structure, which has important implications for sequence-based studies of protein structure and evolution.

Bioinformatics Tools for Analysis of Mammalian Biomolecular Networks and Infectious Disease Drug Targets

F.S.L. Brinkman*[1], K. Breuer[1], Luisa Chan[1], Bhav Dhillon[1], Amir Foroushani[1,2], E.E. Gill[1], S.J. Ho Sui[1], M.R. Laird[1], R. Lo[1], Mike Peabody[1], G.L. Winsor[1], M.D. Whiteside[1], N.Y. Yu[1], R.E.W. Hancock[3], and D.J. Lynn [2]

[1]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada, V5A 1S6

[2] Animal & Bioscience Research Department, AGRIC, Teagasc, Grange, Dunsany, Co. Meath, Ireland

[3]Department of Microbiology and Immunology, University of British Columbia, British Columbia, Canada, V6T 1Z3

We report recent improvements to a collection of software tools that aid the study of pathogens, their hosts, and the development of anti-infective drugs. To facilitate systems-level analysis of the host response to infection, InnateDB (www.innatedb.com) is a database and analysis platform of all available human and mouse interactome data, integrated with >17,000 additional experimentally-verified interactions (manually-curated from >4,000 papers to date), plus tools to facilitate network or pathway-based analyses. For pathogen analysis we are expanding the www.pseudomonas.com database structure to facilitate viewing of novel RNA cleavage data, RNA-seq data, SNP data, and ortholog analysis. We have also released new versions of PSORTdb (www.psort.org/psortb/) for identifying bacterial cell surface and secreted proteins of interest as drug targets and for their potential role in pathogen-host interactions. We have coupled this with an analysis of pathogen-specific genes, to identify candidate anti-infective drug targets and drugs which we are studying further in the laboratory. By better understanding the complex interplay of factors that influence both pathogen and host during the infection process, plus identifying anti-infective drugs that disarm (vs kill), a pathogen (potentially having less selection for drug resistance), or boost immunity, we hope to improve upon current approaches for infectious disease control.

Key words: Pathogen, Systems biology, Drug discovery (if more needed: Pathways, Microbial, Bacteria, Human)

Coevolved residues and the functional association for intrinsically disordered proteins

Chan-Seok Jeong and Dongsup Kim

Department of Bio and Brain Engineering
Korea Advanced Institute of Science and Technology (KAIST)

The evolution of intrinsically disordered proteins has been studied primarily by focusing on evolutionary changes at an individual position such as substitution and conservation, but the evolutionary association between disordered residues has not been comprehensively investigated. Here, we analyze the distribution of residue-residue coevolution for disordered proteins. We reveal that the degree of coevolved residues significantly decreases in disordered regions regardless of the sequence propensity, and the degree distribution of coevolved and conserved residues exclusively differs in each functional category. Consequently, the coevolution information can be useful for predicting intrinsic disorder and understanding biological functions of a disordered region from the sequence.

Correlations among Different Prediction Methods for Intrinsically Disordered Proteins

Fan Jin 1*, Zhirong Liu 1, 2

1 Center for Theoretical Biology, Peking University, Beijing 100871, China
2 College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China

Intrinsically disordered proteins (IDP) are biologically active proteins that do not have stable secondary and/or tertiary structures. The discovery of IDPs is regarded as an unexpected surprise, which may challenge the structure-function paradigm. Accurate prediction of a protein's predisposition to be intrinsically disordered is a necessary prerequisite for the further understanding of principles and mechanisms of protein folding and function.1

In the absence of experimental assignments, the empirical charge/hydropathy correlation for the prediction of IDP sequences provides perhaps the most intuitive description of gross polypeptide conformation, which can predict IDPs well.2 On the other hand,, Galzitskaya et al. illustrate that a one-dimensional parameter, mean packing density, is a useful value that can predict IDPs better than any other parameters.3 It is very strange that the one dimensional parameter of packing density that lack physical meanings can predict better than the two dimensional parameter of charge/hydropathy correlation. We use simulation methods to investigate this puzzle. Specifically, we use a continuous HPQ model4 to investigate the relationship between the two prediction methods and analyze the correlation between them as well as a new parameter called pairwise energy5.

From comparisons on the three parameters mentioned above, we reveal that the essence of IDPs predictions based on the amino acid composition is to find a best projected direction in the twenty-dimensional space. We use linear fit to determine the best direction. The correlation analysis showed that the best direction correlates better with pairwise energy and packing density than HQ Projection.

In summary, packing density is better than HQ parameter in predicting IDPs because it is more closed to the best direction.

Reference
1. H. J. Dyson and P. E. Wright, Nat, Rev, Mol, Cell Biol. 2005. 197:208.
2. V. N. Uversky, J.R. Gillespie and A. L. Fink, Proteins: Struct., Funct., Genet. 2000. 415:427.
3. O. V. Galzitskaya, S. O. Garbuzynskiy and M. Y. Lobanov, PLos Comput. Biol. 2006. 1639:1648.
4. H. S. Ashbaugh and H. W. Hatch, J. Am. Chem. Soc. 2008. 9536:9542
5. D. Zsuzsanna,et al., J. Mol. Biol. 2005, 347, 827:839

MICROBIOME STUDIES: ANALYTICAL TOOLS AND TECHNIQUES

Alterations in the colon microbiota induced by the gastrointestinal nematode Trichuris suis

Robert W. Li 1, Sitao Wu 2, Weizhong Li 2, Dolores Hill 3, Joseph F. Urban, Jr. 4

1USDA-ARS, Bovine Functional Genomics Laboratory, Beltsville, MD
2Center for Research in Biological Systems, University of California, San Diego, CA
3USDA-ARS, Animal Parasitic Diseases Laboratory,
4USDA-ARS, Diet, Genomics, and Immunology Laboratory, Beltsville, MD

Helminth parasites ensure their survival by regulating host immunity through mechanisms that dampen inflammation. These properties have been recently exploited therapeutically to treat human diseases. The bio-complexity of the intestinal lumen suggests that interactions between parasite and intestinal microbiota would also influence inflammation. In this study, we characterized the microbiota in the porcine proximal colon in response to Trichuris suis (whipworm) infection using 16S rDNA-based and whole genome shotgun (WGS) sequencing. A 21-day T. suis infection in pigs induced a profound change in the composition of the proximal colon microbiota. Among the 15 phyla identified, the abundance of Proteobacteria and Deferribacteres was changed in infected pigs. Among these, the populations of approximately 13% of genera were significantly altered by infection. Changes in relative abundance of Succinivibrio and Mucispirillum, for example, may relate to alterations in carbohydrate metabolism and niche disruptions in mucosal interfaces induced by parasitic infection, respectively. Of note, infection by T. suis led to a significant shift in the metabolic potential of the proximal colon microbiota where 26% of all metabolic pathways indentified were affected. Besides carbohydrate metabolism, lysine biosynthesis was repressed as well. Furthermore, changes in pathogen associated pathways suggest that T. suis infection could impact subsequent infection and modulate host-pathogen interactions. Our findings should facilitate development of strategies for parasitic control in pigs and humans and optimize successful helminth therapy to reduce inflammation.

SEPP: Sate Enabled Phylogenetic Placement

S. Mirarab, N. Nguyen, T. Warnow

Department of Computer Science, University of Texas at Austin

We address the problem of Phylogenetic Placement, in which the objective is to insert short molecular sequences (called query sequences) into an existing phylogenetic tree and alignment on full-length sequences for the same gene. Phylogenetic placement has the potential to provide information beyond pure “species identification” (i.e., the association of metagenomic reads to existing species), because it can also give information about the evolutionary relationships between these query sequences and to known species. Approaches for phylogenetic placement have been developed that operate in two steps: first, an alignment is estimated for each query sequence to the alignment of the full-length sequences, and then that alignment is used to find the optimal location in the phylogenetic tree for the query sequence. Recent methods of this type include HMMALIGN+EPA, HMMALIGN+pplacer, and PaPaRa+EPA. We report on a study evaluating phylogenetic pl acement methods on biological and simulated data. This study shows that these methods have extremely good accuracy and computational tractability under conditions where the input contains a highly accurate alignment and tree for the full-length sequences, and the set of full-length sequences is sufficiently small and not too evolutionarily diverse; however, we also show that under other conditions accuracy declines and the computational requirements for memory and time exceed acceptable limits. We present SEPP, a general “boosting” technique to improve the accuracy and/or speed of phylogenetic placement techniques. The key algorithmic aspect of this booster is a dataset decomposition technique in SATe´, a method that utilizes an iterative divide-and-conquer technique to co-estimate alignments and trees on large molecular sequence datasets. We show that SATe´-boosting improves HMMALIGN+pplacer, placing short sequences more accurately when the set of input sequences has a large evolutionary diameter and produces placements of comparable accuracy in a fraction of the time for easier cases. SEPP software and the datasets used in this study are all available for free at http://www.cs.utexas.edu/users/phylo/software/sepp/submission

This poster will be an extension of our SEPP publication, which has already been accepted for publication with an oral presentation. The poster will contain updated results on the application of SEPP toward the taxon identification of short metagenomic reads.

Improved Characterization of Microbiomes and Pathogens in Human Clinical Samples by Suppressing Sequence Backgrounds in Next Generation Sequence Libraries

Joseph S. Schoeniger, Steven S. Branda, Kamlesh Patel, Kelly Williams, Owen Solberg, Hanyoup Kim, Stanley A. Langevin, Zachary Bent, Anupama Sinha, Bryan Carson, J. Bryce Ricken, Elisa La Bauve, Victoria A. VanderNoot, Ronald F. Renzi, Michael Bartsch, Numrin Thaitrong, Pamela Lane, Deanna Curtis, Robert Meagher, Julia N. Kaiser, Julia A. Fruetel, Todd W. Lane

Sandia National Laboratores

Next Generation Sequencing (NGS) of human clinical samples is a promising approach for rapid identificatoin and characterization of emerging pathogens and improved understanding the role of the microbiome in infectious disease. Unfortunately, nucleic acids from the dominant clinical sample types used in public health monotoring of respiratory or febrile outbreaks (e.g., blood/plasma, nasopharyngeal swabs) are often composed of more than 95% human background, as well bacterial rRNA, sequences from fairly invariant "typical" flora, etc.. While bioinformatic identification (and removal) of ubiquitous human sequences and typical microbiome components is useful, it does nothing to address the fundamental inefficiency of the method. In order to increase the sensitivity of NGS for these applications and/or lower the cost per sample analyzed, it would be desireable to remove these "uninteresting" sequences prior to sequencing. The ability to get more "reads on target" also means that newer more rapid NGS platorms with reduced capacity may be employed to produce sequencing results in under a day. At least two approaches exist for suppression of such sequences during library prep: Cot filtration (also called normalization) or capture. Unfortunately, as implemented on the bench, these are also technically challenging and time consuming, especially for minute sample quantities (e.g., plasma RNA). We have developed automated methods for implementing suppression and library prep using microfluidic platforms, as well as bioinformatic tools for assessing the extent of suppression and its efects on sequence content. We present here bioinformatic results of different suppression approaches, applied separately and in tandem, on both control samples and clinical samples. Typical suppression of human sequence levels, and enhancement of trace pathogen sequences, of around one half to one order of magnitude can be achieved in a few hours. Both labware and software solutions to background removal and improving sensitivity and throughput can be applied and optimized in an integrated manner.

Pathogen detection in clinical samples by high-throughput sequencing

Owen Solberg, Milind Misra, Joe Schoeniger, Kelly Williams

A main task in sequence-based detection of pathogens in clinical samples is to quickly identify and remove human sequences, a problem benefitting from the large phylogenetic separation between humans and their pathogens. We have developed a computational pipeline for identifying likely pathogens in high-throughput sequence data from clinical samples. The pre-alignment phase of our pipeline removes minimum-quality tails and rejects sequences with low average quality, removes artifacts involving primer sequences used in library preparation, and rejects low-complexity sequences. The alignment phase begins with Burrows-Wheeler (BW) alignments that provide quick identifications (first for human, then for bacterial, viral and fungal sequences) for the bulk of reads. With the smaller remaining readset, more sensitive BLAST-based alignment is applied to reach more divergent sequences. The post-alignment phase assigns taxonomy to the reads hit during alignment and summarizes abundances at multiple taxonomic levels. The summary often provides strong evidence for taxonomic identification of threat organisms

We have generated sequence data from nasopharyngeal samples from patients with respiratory complaints, and applied our pipeline. Unlike more protected regions of the human body, the nasopharynx contains great diversity of organisms, and variability among patients. Some samples yield clear diagnoses (eg, large fractions of Streptococcus), whereas others yield near-equal distributions of multiple potential pathogens.

MODELING HOST-PATHOGEN INTERACTIONS

Computational Analysis of Novel Drug Opportunities (CANDO)

The CANDO team.

Invited speaker: V. Ram Samudrala

We have a developed a novel and unique computational multitarget fragment-based docking with dynamics protocol to implement a comprehensive and efficient drug discovery pipeline with higher efficiency, lowered cost, and increased success rates, compared with current approaches. We are applying this pipeline to evaluate how all approved drugs bind to all known disease target protein structures. The top ranked predictions are then immediately and directly verified in human clinical studies, sometimes with only a minimal amount of in vitro sanity checks. Our goal is to identify and repurpose the entire repertoire of known drugs as new therapies with an emphasis on underserved diseases. The project represents an integration of our group's applied research on therapeutic discovery, building upon basic protein and proteome structure, function, and interaction prediction research. In addition to repurposing drugs, our compound-structure matrix will provide an atomic level mechanistic understanding of how all known drugs interact with all their targets, antitargets, and off targets, as well as entire pathways and organisms, to cause their desired (and undesired) effects. The talk will focus on the breaking research being generated by the Computational Analysis of Novel Drug Opportunities (CANDO) drug discovery pipeline, which is funded in part by a 2010 NIH Director's Pioneer Award, and its application to targeting host pathogen interactions (as opposed to individual host or pathogen proteins). Further information and details are available from the project web site located at cando.compbio.washington.edu .

Transcriptional correlates of disease outcome and immune response of non-human primates to hemorrhagic fever viruses

Sara Garamszegi [1], Judy Y. Yen [2], Anna Honko [5], Joan B. Geisbert [6], Kathleen H. Rubins [7], Thomas W. Geisbert [6], Yu Xia [1,3,4], John H. Connor [2], Lisa E. Hensley [5]

[1] Bioinformatics Program, [2] Department of Microbiology, School of Medicine, [3] Department of Biomedical Engineering, and [4] Department of Chemistry, Boston University, Boston, MA; [5] US Army Medical Research Institute of Infectious Diseases, Fort Detrick, MD; [6] Department of Microbiology and Immunology, University of Texas, Medical Branch, Galveston, TX; [7] National Aeronautics and Space Administration, Houston, TX

Viral hemorrhagic fever (VHF) can be caused by four unrelated viral families; however, clinical symptoms such as fever, headache, vomiting, immune dysregulation and coagulopathies can occur in all variants of the disease, making the viral diseases clinically indistinguishable. Treatment varies depending on the infectious agent: e.g., Lassavirus (LASV) and other Arenaviruses will respond to the antiviral drug ribavirin if given during early stages of infection; in contrast, there are no current FDA-approved therapies or treatments for the highly infectious Ebolavirus (EBOV). It is imperative to identify unique markers of individual viral diseases. Clinical symptoms are not a good metric for identification of the infectious agent, since symptoms are similar for most VHFs. Therefore, high-throughput technology, such as microarray analysis, can provide a more accurate assessment of immune response in the infected host. We have assessed the genome-wide transcriptional response of circulating peripheral blood mononuclear cells (PBMCs) of non-human primates (NHPs) infected with EBOV. We have identified transcriptional correlates of immune response, and have found that these transcriptional responses change depending on several factors, such as: (1) treatment of NHPs with anticoagulant drugs will provide partial protection against EBOV challenge, will increase the mean survival time of infected animals, and will alter the transcriptional profile; and (2) NHPs that survive EBOV challenge due to treatment have unique gene expression profiles that are correlated with disease outcome. We have identified ~200 statistically significant genes that accurately distinguish disease outcome in EBOV-infected animals, i.e. classification and distinction between surviving and non-surviving NHPs. These genes are associated with inflammatory response, cell growth and proliferation, T cell death, and inhibition of viral replication. Within this gene set, there were two subsets of genes which are independently controlled by the transcription factors C/EBPA and the tumor suppressor and antiviral response mediator p53. The gene expression profile is consistent with decreased activity of C/EBPA and increased activity of p53 in survivors compared to non-surviving animals. Both transcription-factor controlled gene subsets were able to distinguish between surviving and non-surviving animals

We also assessed the transcriptional response of circulating immune cells of NHPs to LASV infection, using whole-genome Agilent microarrays. Sequential sampling over the entire disease course showed that there are specific genomic signatures of the immune response to LASV infection, including the up-regulation of toll-like receptor signaling pathways and innate antiviral transcription factors. We have further characterized the response to infection through the analysis of immune cell subsets to define the responses of specific cell types. Our results suggest that specific transcriptional profiles can be associated with immune response to EBOV or LASV challenge, and that it is possible to identify unique subsets of genes which provide predictive signatures or biomarkers for clinical discrimination, diagnosis or prognosis.

Using multiple microenvironments to find similar ligand-binding sites: application to kinase inhibitor binding

Tianyun Liu 1, Russ B. Altman 1, 2

Department of Genetics 1 and Department of Bioengineering 2
Stanford University

The recognition of small-molecular binding sites in protein structures is important for understanding off-target side effects and for recognizing potential new indications for existing drugs. Current methods focus on the geometry and chemical interactions within putative binding pockets, but may not recognize distant similarities where dynamics or modified interactions allow one ligand to bind apparently divergent binding pockets. We first introduce an algorithm that seeks similar microenvironments within two binding sites, and assesses overall binding site similarity by the presence of multiple shared microenvironments. The method has relatively weak geometric requirements and uses multiple biophysical and biochemical measures to characterize the microenvironments (to allow for diverse modes of ligand binding). We term the algorithm PocketFEATURE, since it focuses on pockets using the FEATURE system for characterizing microenvironments. We apply PocketFEATURE to evolutionarily distant kinases, for which the method recognizes several proven distant relationships, and predicts unexpected shared ligand binding. Using experimental data from ChEMBL and Ambit, we show that at high significance level, 40 kinase pairs are predicted to share ligands. Some of these pairs offer new opportunities for inhibiting two proteins in a single pathway. Based on PocketFEATURE, we proposed an empirical method for evaluating druggability. Our method uses the FEATURE algorithm to evaluate how many microenvironments in a binding pocket look similar to microenvironments found in drug-binding pockets. Based on the degree of overlap, we predict the likelihood that a drug could bind the pocket. Our preliminary results show that we are able to detect druggable sites reliably.

Network-based interrogation of host-response to infectious respiratory viruses

Mitchell HD 1, JE McDermott 1, A Eisfeld 2, A Sims 3, S Belisle 4, L Gralinski 3, G Neumann 2, TO Metz 1, Y Kawaoka 2 , R Baric 3, MG Katze 4, and KM Waters 1

1 Pacific Northwest National Laboratory, Richland, Washington, USA, 99352
2 Influenza Research Institute, School of Veterinary Medicine, University of Wisconsin-Madison, Madison, Wisconsin
3 University of North Carolina, Chapel Hill, NC.
4 Department of Microbiology, University of Washington, Seattle, WA.

Characterization of the host response to viral infection requires sophisticated experimental and computational approaches to distinguish high pathogenicity from low pathogenicity outcomes. Traditional high-throughput analyses of genomic data focus on the identification of differentially expressed components but rarely integrate multiple data types for systems-level anchoring of molecular response to phenotype. In this study, an integrated network analysis approach was used to identify the primary mediators of the immune response during viral infection, as well as the specific mechanisms pathogenic organisms use to target and evade this response. We have completed a systems-level characterization of human epithelial cells responding to H5N1 avian influenza with high pathogenicity and SARS coronavirus as well as viral strains with low pathogenicity over a time course. We integrated global transcriptomic and proteomic measurements using novel network integration and topological feature analysis to compare the responses between viruses with different pathogenicity and identify host targets for therapeutic intervention. We describe the computational challenges for the network analysis and present preliminary findings of experimental validation for predicted targets.

PERSONALIZED MEDICINE

Drug-drug interaction monitoring system based on Personalized Medical Record

Su Youn Baik, Suehyun Lee, Yu Rang Park and Ju Han Kim

Seoul National University Biomedical Informatics (SNUBI), Interdisciplinary Program of Medical Informatics and Systems Biomedical Informatics Research Center, Div. of Biomedical Informatics, Seoul National University College of Medicine, Seoul 110-799, Korea

Drug-drug interaction that can be caused by co-administration more than two different drugs can bring positive effect such as synergistic drug response. But sometimes it makes drug less effective, and moreover brings about serious adverse reactions. There are many studies to previously prevent the negative effects of drug-drug interaction, but they have limitations that scope of the monitoring is restricted within one medical prescription or prescriptions written from same hospital. Therefore we develop 'drug-drug interaction monitoring system(DDIMS)' based on Continuity of Care Record (CCR). When patients input their CCR, DDIMS check the patient's medication records for contraindicated drug combinations comparing with 'Drug interaction database(DID)' that we constructed and directly alert patient to possible danger. This patient-centered DDIMS can systemically monitor the patient's medication records wherever they take drugs, multiple hospitals or pharmacies. In this study, we focus on the drug-drug interaction but the monitoring scope of DDIMS can be extended to drug-disease interaction, drug-age interaction and so on.

Ancestry-deconvolution and medical annotation of personal genomes from people of mixed ancestry

Francisco M. De La Vega, Andres Moreno-Estrada, Archie Russell, Jake K. Byrnes, Jeffrey M. Kidd, Simon Gravel, Martin G. Reese and Carlos D. Bustamante

Due to colonialism, the slave trade, and more recently, affordable access to modern means of transportation, populations of recent mixed genetic ancestry are widespread throughout the world. Understanding how to analyze and interpret admixed personal genomes will be critical to interpret genetic susceptibility risk for those individuals. We extended and vetted a PCA-based admixture deconvolution approach that assigns genomic segments to a number of potential ancestral or source populations. Here, present the analyzes of genomes of admixed individuals from North America previously genotyped by the International HapMap 3 project: five of African-American descent (ASW), five of Mexican-American descent (MXL), and two of Puerto Rican decent (PUR). Two of these genomes where sequenced with the SOLiD™ System, and the rest are part of a set released to the public domain by Complete Genomics. We utilized our algorithm to assign segments of the paternal and maternal haplotypes to three potential ancestral populations: West African, European, or Native American. Intersection of the variants from each genome with the OMIM, HGMD, PharmGKB, NHGRI GWAS, and locus specific databases using the Omicia Genome Analysis System identified an average of ~70 non-synonymous coding alleles per genome which are annotated to have an effect on human phenotypes. We observe variants in admixed genomes that are otherwise rare in the major source populations being harbored in segments originating in a different ancestral population (e.g. an allele that confers susceptibility to immunoglobulin A nephropathy which is rare in Europeans and Hispanics, but common in Africa appears in an African-ancestry segment in a MXL sample). This demonstrates that, on an individual level, the average proportion of the genome derived from a given ancestry is less informative than the specific ancestry assignment at key genomic positions. Our results also underscore the need to consider local genomic ancestry in interpreting personal genomes, and makes it critical that efforts to map the genetic basis of common disease undertake variant discovery and association mapping in individuals of diverse ancestries relevant to the population of medical genetic interest.

Gene Selection using a High-Dimensional Cox Model in Microarray Data Analysis

Akihiro Hirakawa
Department of Management Science, Tokyo University of Science

Shuhei Kaneko
Department of Management Science, Tokyo University of Science

Chikuma Hamada
Department of Management Science, Tokyo University of Science

Mining of gene expression data to identify genes associated with patient survival is an ongoing problem. Such genes are used to achieve prognoses that are more accurate and to improve treatment strategies. The Cox proportional hazards model is the most popular method to evaluate the relationship between gene expression and survival. In the general setting with $n > p$, the coefficients are estimated by maximizing Cox's partial likelihood, but this does not work when $p > n$. The least absolute shrinkage and selection operator (Lasso) is often used for parameter estimation when $p > n$. The Lasso shrinks some of the coefficients to 0 and the amount of shrinkage is determined by tuning parameter, which is often determined by cross-validation. The model determined by the cross-validation contains many false positives whose coefficients are actually 0. Inclusion of such genes in the prediction model may yield an inaccurate prediction of patient survival. In this study, we propose a method for estimating the false-positive rate (FPR) for Lasso estimates in a high-dimensional Cox model. We performed a simulation study to examine the precision of the FPR estimate by the proposed method. The precision of the FPR estimate would be satisfactory in practice, although it was slightly underestimated in almost all scenarios. We applied the proposed method to real data compromising the overall survival in 240 diffuse large B-cell lymphoma patients and the expression of 7399 genes. Six genes were associated with overall survival using the Lasso. In this case, the FPR estimate was 40.2%, indicating that only 4 ($= 6 \times (1 - 0.402)$) genes may be true positives. We categorized 240 patients into 4 groups based on the prognostic index calculated using the expression levels of 6 genes. Similar categorization was performed using the expression levels of 4 genes. The Kaplan-Meier curves for the 4 groups in each categorization were similar. The 2 genes excluded in the latter categorization were not strongly correlated to overall survival. When we develop prediction models for patient prognosis using gene expression, fewer genes should be included in the model. The proposed method can identify the predictive genes.

SIFT Indel: an on-line tool for predicting the effects of frameshifting 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

There are approximately 50-280 frameshifting (FS) indels in each human genome. How can so many frameshifting indels exist in each human, and what are the implications for a human's phenotype? To answer this question, we created SIFT Indel, a prediction method for frameshifting indels. This is an extension to the SIFT algorithm which predicts the effect of amino acid substitutions and was introduced in 2001 by the senior author of this paper. We initially extracted 20 features and after a heuristic feature selection process, 4 features were chosen. The final decision tree algorithm achieved 84% accuracy, with 90% sensitivity and 81% precision using 10-fold cross-validation. We applied SIFT Indel to human FS indels identified from the human genomes sequenced by the 1000 Genomes Project and by Complete Genomics and found that 73-85% of rare indels (frequency < 0.05) are predicted to alter gene function, whereas common FS indels are less likely to have an effect (33%-39% predicted to alter gene function for indels with frequency > 20%). We also found that a substantial portion of common FS indels are predicted to alter gene function, this suggests that common FS indels may have phenotypic relevance and it may be naïve to filter on an allele frequency filter alone. The SIFT indel prediction algorithm is available at http://sift-dna.org/www/SIFT_indels2.html. The server takes around 15-20 minutes to make predictions for 1000 indels. Researchers are encouraged to submit variants from sequenced human genomes to obtain SIFT predictions for both single nucleotide variants and indels.

Computational tools for analysis next generation sequencing data

Victor Solovyev

Department of Computer Science, Royal Holloway, University of London, UK

Igor Seledtsov, Denis Vorobyev, Petr Kosarev, Vladimir Molodsov, Nicolay Okhalin, Alexandr Bachinsky

Softberry Inc., 116 Radio Circle, Suite 400, Mount Kisco, NY 10549, USA

Next-generation sequencing (NGS) transforms today's biology research by providing fast sequencing of new genomes, genome-wide association studies (GWAS), sequencing personal genomes, variant discovery by resequencing targeted regions or whole genomes, de novo assemblies of bacterial and eukaryotic genomes, annotating the transcriptomes of cells, tissues and organisms (RNAseq), and gene discovery by metagenomics studies. To analyze next-generation sequencing data we advanced further our OligiZip assembler, SNP-Annotator and Transomics pipelines that provide solutions to the following tasks: 1) de novo reconstruction of genomic sequence; 2) reconstruction of sequence with a reference genome; 3) mapping RNA-Seq data to a reference genome and identification of alternative transcripts with quantification of their abundance; 4) SNPs discovery and estimation of SNP impact on a gene's function in genome-wide association studies

To test reconstruction of bacterial sequences we assembled genomic sequence of Methanopyrus kandleri TAG11 and Methanopyrus kandleri AV19. Solexa reads, about 6 million each for AV19 and TAG 11, were produced by sequencing lab of Harvard PCPGM. AV19 genome itself has been assembled perfectly in one contig. Time of AV19 assembling is ~ 1 min on one Linux node. OligiZip also has been successfully applied to assembly a model eukaryotic genome in Assemblathon 1 collaborative efforts where many research groups presented genome assembles to estimate the accuracy of various assembling software

For RNAseq data the Transomics pipeline initially maps reads to the genomic sequence and identifies spliced and non-spliced reads coordinates. This information used by our FGENESH gene prediction program that includes an iterative procedure for predicting alternative splicing gene variants. We have developed a module to compute a relative abundance of predicted alternative transcripts solving a system of linear equations. The initial variant of Transomics pipeline has been successfully applied to Human, C.elegans and Drosophila data of the RGASP project. We also developed a powerful Sequence assembling Viewer to work with the reads data and assembling results interactively. As an example of Transomics application for identification of disease specific genes we analyzed RNAseq data for non-tumorigenic epithelial cell line and epithelial cells from infiltrating ductal carcinoma of the breast. Comparative analysis shows a set of genes that have different alternative splicing forms in these cell lines

OligoZip, SNP_Annotator and Transomics pipeline components and other software programs are available to run 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.

TEXT AND KNOWLEDGE MINING FOR PHARMACOGENOMICS

LINKING PHARMGKB TO PHENOTYPE STUDIES AND ANIMAL MODELS OF DISEASE FOR DRUG REPURPOSING

Robert Hoehndorf 1, Anika Oellrich 2, Dietrich Rebholz-Schuhmann 2,

Paul N. Schofield 3, Georgios V. Gkoutos 1

1 Department of Genetics, University of Cambridge
Downing Street, Cambridge, CB2 3EH, UK

2 European Bioinformatics Institute
Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

3 Department of Physiology, Development and Neuroscience
University of Cambridge, Downing Street, Cambridge CB2 3EG, UK, and The Jackson
Laboratory, 600, Main Street Bar Harbor ME 04609-1500, USA

The investigation of phenotypes in model organisms has the potential to reveal the molecular mechanisms underlying disease. The large-scale comparative analysis of phenotypes across species can reveal novel associations between genotypes and diseases. We use the PhenomeNET network of phenotypic similarity to suggest genotype-disease association, combine them with drug-gene associations available from the PharmGKB database, and infer novel associations between drugs and diseases. We evaluate and quantify our results based on our method's capability to reproduce known drug-disease associations. We find and discuss evidence that levonorgestrel, tretinoin and estradiolare associated with cystic fibrosis (p < 2.65 x 10^-6, p < 0:002 and p < 0:031, Wilcoxon signed-rank test, Bonferroni correction) and that ibuprofen may be active in chronic lymphocytic leukemia (p < 2.63 x 10^-23, Wilcoxon signed-rank test, Bonferroni correction). To enable access to our results, we implement a web server and make our raw data freely available. Our results are the first steps in implementing an integrated system for the analysis and prediction of drug-disease associations for rare and orphan diseases for which the molecular basis is not known.

Infer Novel Genes and Pathways that Modulate Sphingolipid Pathway from a Novel Yeast Gene Network Derived from Ontology Fingerprints

Tingting Qin, Lam C. Tsoi1
Bioinformatics Graduate Program, Dept of Biochem & Mol Biology, Medical University of South Carolina, Charleston, SC

Nabil Matmaty
Department of Biochem and Mol Biology, Medical University of South Carolina, Charleston, SC

Bidyut K. Mohanty
Department of Biochem and Mol Biology, Medical University of South Carolina, Charleston, SC

Andrew B. Lawson
Biostatistics and Epidemiology Division, Department of Medicine, Medical University of South Carolina, Charleston, SC

Yusuf Hannun, W. Jim Zheng
Department of Biochem and Mol Biology, Medical University of South Carolina, Charleston, SC

We integrated biomedical literature, ontology, network analysis and experimental investigation to infer novel genes that can modulate yeast sphingolipid pathway. Such modulations may play important roles in regulating the cellular functions of bioactive lipids and pharmacogenomics. We first constructed novel gene networks by performing pairwise comparison of all yeast genes' Ontology Fingerprints—a set of ontology terms overrepresented in the PubMed abstracts linked to a gene along with their corresponding enrichment p-value. The comparison generated a weighted undirected gene network where genes are nodes and the similarity scores between genes are weighted edges. The network was further refined by applying a Bayesian hierarchical model to distinguish biologically relevant connections from those that are not. To infer novel genes potentially involved in sphingolipid pathway modulation, we identified a subnetwork of well known yeast sphingolipid genes together with non-sphingolipid genes that have no "sphingo" prefix or "ceramide" in PubMed abstracts and descriptions associated with these gene. These non-sphingolipid genes are considered as candidate genes that can modulate sphingolipid pathway. The candidate genes were further prioritized by "Total Score"—the sum of similarity scores between a candidate gene and all sphingolipid genes. The larger the score is, the more likely that the gene is relevant to sphingolipid pathway. We tested the Myriocin sensitivity of the deletion strains of the top ranked candidate genes, followed by lipidomic analysis. As a control, bottom ranked genes are randomly selected for Myriocin screening. The results show that the proportion of top ranked candidates genes whose deletion showing altered sphingolipid pathway activity (Myriocin sensitivity) is significantly higher than that among lower ranked ones, and the lipidomic profiles of these deletion strains are significantly different from that of wide type. Our novel network analysis provides a powerful tool to study pathway modulation and can be applied to study human disease related pathways.

WORKSHOPS AND MEETINGS

Workshop: Structure and Function of Chromatin and Chromosomes

FastHASH: A new GPU-friendly algorithm for fast and comprehensive next-generation sequence mapping

Hongyi Xin
Computer Science Department, Carnegie Mellon University, Pittsburgh, PA

Donghyuk Lee
Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA

Farhad Hormozdiari
Department of Computer Science, University of California Los Angeles, CA

Can Alkan
Department of Genome Sciences, University of Washington, Seattle, WA

Onur Mutlu
Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA

With the introduction of next-generation sequencing (NGS) technologies, we are facing an exponential increase in genomic sequence data. NGS provides low-cost and high-throughput genome sequencing. Yet, it requires enormous computational resources. This is because 1) NGS maps many reads to the entire human genome and 2) each read is short, making it computationally expensive to search for all potentially-matching locations in the human reference genome, especially in the presence of polymorphisms. A key challenge in making NGS successful in medical and genetic applications is to design algorithms that can process and analyze the enormous amounts of sequence data very fast and energy-efficiently.

To address this ever increasing computational need, several solutions were proposed. Mapping algorithms based on Burrows-Wheeler transform and Ferragina-Manzini index such as BWA and SOAP2 are fast in searching for exact matches of a read in the human reference genome, but are not comprehensive, i.e. they do not search for all potentially-matching locations. On the other hand, hash table based seed-and-extend algorithms, like mrFAST and SHRiMP, are comprehensive but typically much slower. Some of these algorithms use single-instruction multiple-data (SIMD) operations available in modern processors to modestly improve performance. Several other algorithms utilize graphics processing units (GPUs), but they are not optimized to fully take advantage of the massively-parallel GPU architecture.

In this work, we propose a new algorithm, FastHASH, which drastically improves performance over mrFAST, while maintaining comprehensiveness. Our key observation is that mrFAST performs too many costly computations that can be avoided by intelligently restructuring the algorithm to take advantage of the full knowledge of the human reference genome. There are two sources of costly computations in MrFAST that our new algorithm reduces. First, for every read, mrFAST finds out all potentially-matching locations in the reference genome. Then, for each possible potentially-matching location, mrFAST performs an expensive string comparison to extract complete differential information between the input read and the reference genome, even if the location will not match. FastHASH uses two key ideas to reduce both types of expensive computations. First, it drastically reduces the number of potentially-matching locations considered for string comparison, while still preserving comprehensiveness, by making use of complete information of the reference genome. We call this method Cheap Key Selection. Second, it drastically reduces the number of string comparisons by rejecting obviously-incorrect locations in the early stages of mapping. We call this method Adjacent Filtering.

Our initial CPU implementation of FastHASH provides 40-fold speedup over mrFAST, while preserving the same comprehensiveness. Since FastHASH has little control-flow that significantly diverges, a property essential for efficient GPU implementation, FastHASH is also amenable to a GPU implementation.

Workshop: Systems Pharmacogenomics -- Bridging the Gap

A multi-level Bayesian approach for discovering predictive pharmacodynamic markers of kinase inhibitor efficacy using quantitative phosphoproteomics

Y. Ann Chen 1, Guolin Zhang 2, Irene Bai 2, Jiannong Li 2, Bin Fang, Kate, J. Fisher 1, Eric A. Welsh 3, Steven Eschrich 3, John M. Koomen 4, Eric B. Haura 2

1 Biostatistics, 2 Experimental Therapeutics, 3 Bioinformatics, 4 Molecular Oncology Moffitt Cancer Center

Dysregulation of signaling networks plays a critical role in cancer biology. Mass-spectrometry based phosphoproteomics profiling has identified and quantified signaling proteins, and has further enabled protein network mapping. However, association of expression of each level of marker (e.g., phosphopeptides, proteins, or pathways) with outcomes of interest (such as drug response) has been typically assessed only at the single marker level. For example, in traditional pathway analysis, association of a pathway with an outcome would be performed after summarizing the group behavior of genes or proteins within a pathway, but typically the effects of subsets of proteins, pathways, and their combinations are not accessed. We have developed a Bayesian approach to integrate experimental phosphoproteomics data with existing biological pathway and site-specific functional information. Pathway information was obtained from the KEGG pathway database and phosphorylation site-specific functional annotation was obtained from PhosphoSite maintained by the Cell Signaling Technology. The joint effects of selected proteins and phosphotyrosines (pYs) for predicting drug efficacy were first evaluated through structured stochastic search procedures in a linear model. Latent variables were introduced for the indication of inclusion/exclusion of the markers. Similarly, for the proteins or pYs with functional information, the joint effects of pathways (or pY-specific functional information) and pYs were evaluated in a similar fashion. We illustrate the approach with an application to quantitative phosphoproteomic data generated using anti-phosphotyrosine peptide affinity purification followed by liquid chromatography-mass spectrometry (LC-MS/MS). We treated a human sarcoma cell line A204 with multi-kinase inhibitors, dasatinib and imatinib, at three different concentrations. A total of 388 unique phosphotyrosines (pYs) from278 proteins passed the quality control criteria for inclusion in the analyses. Proteins and pYs were ranked by their marginal posterior probabilities in the protein-pY two-level model. Some of the highly ranked proteins, including FYN, YES, EPHA2, PGFRA, and GAB1, are known to be drug targets; while the roles for other highly ranked proteins, such as VIME and P85B (PI3K regulatory subunit beta), are less clear. Phosphorylation patterns were different for some of the pYs within the same protein supporting the need for multi-level analysis. For the model integrating biological functional information with the pY data, pathways with high posterior probabilities included NOD-like signaling pathway, tight junction, and focal adhesion. We also identified additional proteins and pYs, whose joint effects were predictive of drug efficacy, including MK01 (ERK2), MK03 (ERK1), FLNB, GNAS1, KAPCB, LAP2, and P55G (PI3K regulatory subunit gamma). We have validated the relative phosphorylation levels of MK01 and MK03 using quantitative western blotting. Our approach identified several key pathways and phosphoproteins responsible for kinase inhibitor efficacy.

CAFA/CAGI Challenges Discussion

CAGI: The Critical Assessment of Genome Interpretation a community experiment to evaluate phenotype prediction

Susanna Repo 1, John Moult 2, Steven E. Brenner 1

1 Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720
2 IBBR, University of Maryland, Rockville, MD 20850

The 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. At the first CAGI workshop in December 2010, Pauline Ng, Iddo Friedberg, Sean Tavtigian, Gad Getz, and Sean Mooney made detailed assessments of 108 prediction entries from 8 countries (USA, Belgium, China, Finland, France, Germany, Italy, and Singapore). The workshop gathered an enthusiastic crowd of 40 participants, as well as 7 off-site predictors who viewed a live feed. The CAGI 2010 datasets included rare variants identified from resequencing in cancer cases and controls; nonsynonymous point mutations within a human metabolic enzyme; clinical phenotypes associated with complete human genomes and exomes; cancer cell-line pharmacogenomics; effects of double-mutants in reactivation of p53; and mechanisms underlying GWAS disease associations. The meeting revealed the relative strengths of different prediction approaches, showing some that worked consistently well, while other classes worked only on special types of problems. Even with the simplest dataset, involving nonsynonymous mutations in a human metabolic enzyme, yielded great variability of the results: the best groups had a Spearman rank correlation of ~0.6 with the correct results, while predictions submitted by some groups were only as good as random or worse. Several predictions for the cancer case / control dataset significantly segregated the individuals into their respective cohorts, and some methods performed better than the method initially applied by the dataset author. Overall, CAGI 2010 highlighted the need for customized approaches for specific problems. We are currently completing further analysis of the prediction quality.

This presentation will focus on the CAGI 2011, whose prediction season ran through October, with assessment period in the late fall, and a meeting held at UCSF Mission Bay on 9-10 December 2011. In addition to the dataset classes of 2010, the CAGI 2011 experiment includes datasets on exome pharmacogenomics, genomes of identical twins with discordant disease, exomes of mice with defined disease phenotypes, classification of Crohn's disease patients and healthy individuals based on exome data and microbial fitness under stress conditions. Current information is available at the CAGI website at http://genomeinterpretation.org.

INDEX