Comprehensive Profiling of the Lung Transcriptome in Emphysema and IPF Using RNA-SEQ

John Brothers II*[1], Rebecca Kusko*[2], Jennifer Beane[3], Gang Liu[3], Lingqi Luo[3], Brenda Juan Guardela[4], John Tedrow[4], Yuriy Aleksyev[1,5], Ivana V. Yang[6], Mick Correll[7], Mark Geraci[8], John Quackenbush[7], Frank Sciurba[4], Marc Lenburg[1,2,3,5], David A. Schwartz[6], Naftali Kaminski[4], Avrum Spira[1,2,3]

[1] Bioinformatics Graduate Program, Boston University, Boston, MA
[2] Genetics and Genomics Graduate Program, Boston University School of Medicine, Boston, MA
[3] Section of Computational Biomedicine, Department of Medicine, Boston University School of Medicine, Boston, MA
[4] Simmons Center for Interstitial Lung Disease and Department of Medicine, University of Pittsburgh Medical Center, Pittsburgh, PA
[5] Department of Pathology and Laboratory Medicine, Boston University School of Medicine, Boston, MA
[6] Center for Genes, Environment and Health and Department of Medicine, National Jewish Health, Denver, CO
[7] Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA
[8] Department of Medicine, University of Colorado School of Medicine, Aurora, CO
*These authors contributed equally to this research

RNA-Seq is an approach for quantifying gene expression that provides improved detection of the transcriptional landscape. As part of the Lung Genomics Research Consortium, we sought to characterize transcriptomic changes underlying the molecular pathogenesis of emphysema and idiopathic pulmonary fibrosis (IPF) using mRNA-Seq and to integrate mRNA-seq and microRNA microarray expression data from the same samples. 89 lung tissue samples were sequenced on the Illumina GAIIx with 75nt pairedend reads resulting in ~30-40 million reads per sample. Using gapped aligner Tophat, 85% of reads aligned to hg19. Genes were annotated using Ensembl59. A mixture model was used to identify genes with detectable or absent expression across all samples. A well-defined subset of 58 samples from subjects with IPF (n=19), Emphysema (n=19), or control samples (n=20) differential expression was determined using a t-test. Using miRconnX, Agilent Human miRNA Microarray(V3) data and mRNA-Seq data were integrated with a prior network of miRNA-mRNA interactions. A total of 6359 genes were detected by mRNA-Seq, 9538 genes were not detected, and 8397 genes were variably detected across all samples. The expression levels of 1770 genes differed between IPF and Control, and 220 genes between Emphysema and control ($p<0.001$). Using MirConnX, we identified several miRNA, such as the miR-96/182/183 family, that co-varied with genes differentially expressed in IPF and emphysema. Genes whose expression was significantly altered in emphysema or IPF and which were predicted targets of miRNA were enriched for wound-healing (Emphysema, $p<0.01$) and immune response (IPF, $p<0.02$). Using the reads that aligned across known splice junctions, we identified 5 Emphysema-associated alternatively spliced isoforms and 19 IPF-associated spliced isoforms ($p<0.01$). Our data suggests that a sizable portion of the lung transcriptome is altered in these diseases and that these changes may include alterations in miRNA-based regulation of gene expression and alternatively spliced transcripts. These findings could ultimately lead to novel biomarkers for chronic lung disease and potential novel therapeutic targets.

Tracking Antiviral Responses Following Infection with Lassa Fever Virus

Ignacio Sanchez Caballero, Gracia Bonilla, Jae Yoon Chung
Alex Watson, Judy Y. Yen, John H. Connor

Bionformatics Graduate Program, Boston University, Boston, MA

Lassa fever is severe hemorrhagic fever caused by the Lassa virus. It is estimated that Lassa virus infects more than 300,000 people per year in Western Africa and that it causes more than 3,000 deaths annually. Its symptoms are often confused with those of other infections, such as Influenza, which can delay proper care. The case-fatality rate for hospital-admitted patients is approximately 15%, but in some cases has been reported to be greater than 50%. The confusing presentation of individuals infected with Lassa has driven considerable interest in developing better diagnostic tests. We were interested in determining if analyzing the circulating immune system would allow us to identify "markers" of Lassa virus infection. To test this hypothesis we investigated the gene expression patterns of cells that were extracted from Lassa infected animals over the course of infection. We used Agilent microarrays to examine the gene expression levels in peripheral blood mononuclear cells (PBMCs) extracted from non-human primates at different stages of infection. PBMCs perform an important role in the immune response to infection and are easily extracted from human blood. This makes them suitable for use as a diagnostic tool, even in remote areas where modern medical facilities are scarce. We used a linear regression model to estimate the significance of the fold-changes between the pre-infection samples and those obtained during early, middle and late stages of infection. This gene set was further filtered to include only those genes whose expression appeared sustained during the intermediate stages of infection, discarding those genes that had small changes in expression. From this analysis we have identified a set of genes whose expression profiles in blood cells appear to be specific to Lassa infection. We are currently carrying out experimental validation of their suitability as biomarkers of early-infection.

*Work carried out in the laboratory of Professor John H. Connor in the Department of Microbiology (Boston University School of Medicine)

Kindase-specific Phosphorylation Sites Predicted with Linear Scoring Functions

Özgür Demir

Freie Universität Berlin, Macromolecular Modelling Group

Phosphorylation is an important post-translational modification and used to control diverse signal-transduction pathways. Experimental detection of phosphorylation sites can be a time-consuming and laborious task. Hence, there is need for fast and accurate *in silico* methods to predict phosphorylation sites from the amino acid sequence or 3D structure. *In silico* models can be classified into kinase-unspecific and kinase-specific predictors. The former ones try to predict phosphorylation sites regardless of the involved kinase. The latter ones on the other hand try to predict sites, that may be modified by a specific kinase. The second type of phosphorylation site predictors may therefore be used to map the experimentally verified phosphorylation sites to a kinase enzyme. Recently, Que *et al*. [1] compared the most popular prediction approaches in an independent benchmark. The group has shown that current phosphorylation site predictors perform quite poorly, which makes none of the currently available predictors effective for practice use. Hence, there is still a substantial need for the kinase-specific phosphorylation site prediction. One possible reason for the low predictive power might be that most *in silico* approaches only consider a short sequence range surrounding the phosphorylation site. The whole protein itself is most often ignored. Another aspect is that even through some key motifs for kinase-specific substrate binding are known, the binding process is no yet fully understood in detail.

References

[1] Que, S., *et al.*, Evaluation of protein phosphorylation site predictors. *Protein Pept Lett.* **17**(1): pp. 64-9.

Gene expression profiles in nasal epithelium as a minimally invasive biomarker for the early detection of lung cancer

Joseph P. Gerrein*[1,2], Christina Anderlind*[1], Yuriy Alekseyev[3], Mark Kon[2,5]
Stefano Monti[1], Jerome Brody[4], Marc E. Lenburg[1,2], Avrum Spira[1,2]

[1] Section of Computational Biomedicine, Department of Medicine, Boston University
[2] Graduate Program in Bioinformatics, Boston University
[3] Department of Pathology and Laboratory Medicine, Boston University
[4] Pulmonary Center, Boston University School of Medicine
[5] Department of Mathematics and Statistics, Boston University
**These authors contributed equally*

Objective: As our lab has previously demonstrated, gene-expression differences in the cytologically normal bronchial airway can distinguish patients with and without lung cancer we now seek to determine whether or not nasal epithelial gene expression also varies between smokers with and without cancer since this could serve as the basis for a less invasively collected biomarker for lung cancer diagnosis.

Methods: RNA from nasal epithelial brushings from smokers with suspect lung cancer was profiled on Affymetrix Human gene 1.0 ST arrays. Data from 44 current and former smokers (27 with cancer, 17 with benign disease) was used to identify differentially expressed genes using ANOVA. Gene Set Enrichment Analysis was used to identify enriched pathways. Classifiers to predict cancer diagnosis were constructed using Random Forest and support vector machine (SVM) methods with performance evaluated by leave-one-out cross validation.

Results: 31 differentially expressed genes were identified at an FDR<0.25. Enrichment was observed for the MAPK, FAS, p53 signaling, RAS and additional pathways (FDR <0.05). Biomarkers with accuracy greater than 0.8 were identified using SVM.

Conclusion: Our data suggests that gene-expression profiling in nasal epithelium might serve as a non-invasive approach for the early detection of lung cancer among smokers.

Decoding ChIPseq with multiple binding events provides site detection with high resolution and allows estimation of binding cooperativity.

Antonio Gomes[1], Matthew Peterson[2], Anna Lyubetskaya[1], Thomas Abeel[3]
Luís Carvalho[4], James Galagan[1,2,3]

[1] Bioinformatics Graduate Program, Boston University, Boston, MA
[2] Department of Biomedical Engineering, Boston University, Boston, MA
[3] Broad Institute of MIT and Harvard, Cambridge, MA
[4] Dept. of Mathematics and Statistics, Boston University, Boston, MA

Gene expression depends on the presence of regulatory proteins. Those regulatory proteins recognize and bind specific DNA regions, helping to decide when a gene should be expressed. Understanding how binding occurs *in vivo* is an essential step to understand gene regulation. A global comprehension of the binding network can be obtained experimentally using chromatin immuno-precipitation (ChIP) followed by sequencing (seq). Currently, most analyse of ChIP-seq data focus on how to identify enriched regions corresponding to true binding sites. However, little has been explored on the binding comprehension inside each region. In this research, we developed an approach to decode the signal inside each region, with special interest in cases with multiple binding sites. We started our analysis developing a parametric model to represent the binding signal. We assume coverage can be explained by two possible signals, the first considers the DNA fragments containing only a single binding event and the second the fragments containing two. Our model is applied into a pipeline that uses sequence motifs to constraint search space and refine the resolution of binding-site detection. The method can detect multiple sites inside the same region and outperforms the resolution of other currently used methods. It also shows high-reproducibility in both *M. tuberculosis* and human data. The use of a double-binding signal increases the sensitivity and specificity of site detection when compared to the case with only a single binding signal. Finally, we explore the secondary terms provided by double binding signals to estimate cooperative binding from ChIP-seq.

Genomic Signatures of Carcinogenicity

Daniel Gusenleitner, Harold Gomez, Tisha Melia

Bioinformatics Graduate Program, Boston University, Boston, MA

There are around 80,000 chemical compounds used in industry, many of which are suspected carcinogens. Standard approaches to carcinogen testing are costly and time-consuming and, as a result, only approximately 1,500 of the chemicals currently in commercial use have been tested. Additionally, some chemicals can have synergistic effects, making the characterization of carcinogenic compounds even more difficult as combinations have to be considered.

The goal of this project is to develop of computational models of carcinogenicity based on gene expression profiles to classify, with unprecedented speed, the carcinogenic potential of individual or complex mixtures of environmental pollutants and/or therapeutics, and to study their mechanisms of action. To this end, we analyzed a collection of 3610 gene expression microarray profiles from rats treated with 188 well-characterized chemicals, including genotoxic and non-genotoxic carcinogens, as well as non-carcinogens. Five different tissues types were profiled: liver, kidney, heart, thigh muscle and cell-cultured hepatocytes, at multiple doses and times of exposure.

The analysis aimed at identifying gene-expression signatures that could distinguish benign from carcinogenic substances, as well as genotoxic from non-genotoxic carcinogens. First, we used exploratory data analysis and clustering techniques to assess data quality and identify the dominant strata in the dataset. Then we derived differentially expressed genes and differentially regulated pathways in order to characterize the biological response to the chemical stressors. Finally, we built an ensemble classifier that uses gene expression-based and chemical structure-based classification models, and takes into account dataset substructure, to accurately predict compound carcinogenicity and genotoxicity. Integration of the toxicogenomic-based model with structure-based models leads to increased carcinogenicity prediction accuracy. Additional preliminary results confirm and expand upon previous studies implicating oxidative stress, DNA damage, and alterations in metabolic pathways in the response to exposure. Furthermore, prediction of carcinogenicity is tissue-dependent, underlining the importance of differences in catabolism between different cell types.

Identifying Neighborhoods of Coordinated Gene Expression and Metabolite Profiles

Timothy Hancock[1], Nicolas Wicker[2], Ichigaku Takigawa[1], Hiroshi Mamitsuka[1]

[1] Bioinformatics Center, Kyoto University, Japan.
[2] Laboratoire de Bioinformatique et Génomique Intégratives, Institut de Génétique et de Biologie Moléculaire et Cellulaire, Université de Strasbourg, France.

We evaluate the extent of the transcript-metabolite correlation with a two stage investigation correlating metabolic network gene expression and metabolites for *Escherichia coli K-12*. In the first stage we tested the hypothesis that transcript-metabolite correlations are sustained at long distances away from the target metabolite. Our results show that overall the transcript-metabolite correlations are sustained over a long network distance of about 3 to 7 genes away from the target metabolites. This suggests that a few hub reactions are controlling the coordinated structure within metabolic networks. We then propose a method to identify these hub reactions by searching for commonly traversed genes over all maximally coordinated expression paths connecting all metabolite pairs. We then use minimum set cover approach to identify hub reactions and metabolites. The result of this analysis revealed that a surprisingly small number of metabo-lites, approximately 18 to 25, are sufficient to summarize the highly correlated metabolic network structure. Additionally, we found that strong correlations exist between these hub metabolites and reactions and are related to a network regulatory signature.

Finding Extensions of Pharmacogenomic Pathways based on Structural Variations by Data Assimilation

Takanori Hasegawa[1], Masao Nagasaki[2], Rui Yamaguchi[3]
Seiya Imoto[3], Satoru Miyano[3]

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan
[2] Department of Integrative Genomics, Tohoku Medical Megabank Organization, To-hoku University, Japan
[3] Human Genome Center, Institute of Medical Science, University of Tokyo, Japan

Notably in the field of pharmacogenomics, from the view point of developing innovative medical treatments, elucidating intracellular pathways of drug responses is attracting a great deal of attention. These pathways are often constructed by combing molecular reactions from reliable biological experiments stored in literature. However, the constructed pathways often inadequately reproduce intracellular reactions because there possibly exist missing information and wrong relations in the literature. Therefore, methods to complement and expand literature-based pathways by using observed biological data are strongly demanded. In our previous work, we proposed a simulation-based method to achieve this by evaluating many candidate pathways that are partly changed from the original pathways evaluating comprehensively. [1] The method found many alternative pathways having higher predictive power for observed data than the original ones and indicated that there is still room for improvements in the literature. On the other hand, the method has a problem that evaluating a large amount of candidates comprehensively is computationally intensive. In this work, we propose a new method with high ability to evaluate candidates thought to have higher likelihood selectively. Namely, by selecting evaluation candidates based on similarities between candidates, the proposed method reduces the time required for evaluating implausible candidates. Also, parameters are estimated more efficiently by using estimated parameters of similar candidates based on the data-assimilation technique. The time-course microarray data of rat liver cells treated with corticosteroid are used to show the effectiveness of the proposed method in comparison to the previous method. Consequently, the proposed method accomplished more than 85% correct rate within approximately 15% time. In addition, we merge the obtained results and draw biological networks with a focus on drug responses.

References

[1] Hasegawa, T., Nagasaki, M., Yamaguchi, R., Imoto S., Miyano, S.: Comprehensive pharma-cogenomic pathway screening by data assimilation, *Proc. 7th International Symposium on Bioinformatics Research and Applications* 2011, Heidelberg, LNCS, 6674, 160-171.

Protein Complex Prediction Via Improved Verification Methods Using Constrained Domain-domain Matching

Yang Zhao[1], Morihiro Hayashida[1], Jose C. Nacher[2]
Hiroshi Nagamochi[3], Tatsuya Akutsu[1]

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan
[2] Department of Complex and Intelligent Systems, Future University-Hakodate, Japan
[3] Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University

Keywords: protein complex, domain-domain interaction, integer linear programming

Introduction: Identification of protein complexes is important for understanding of gene regulatory networks and cellular mechanisms. Most methods for predicting protein complexes from protein-protein interaction (PPI) networks have been developed based on network topology. These methods often detect dense subgraphs in PPI networks as protein complexes because most proteins in complexes often interact with each other. However, almost all of the existing methods do not have paid any attention to the structural constraint of proteins in PPI networks, which resulted in low precision. The method proposed by Ozawa et al. has verified and reconstructed the topology of domain-domain interactions in PPI networks. This method makes use of the concept that proteins in candidates each of whose domains participates only in a single interaction can form a valid protein complex. In terms of this concept, this approach seeks for optimal combinations of domain-domain interactions (DDIs) in the complex candidates predicted from other existing methods. Although this approach has achieved a relatively high precision, it still outputs a number of false positives. In this poster, we propose a novel formulation of integer linear programming based on the idea that a candidate complex should not be divided into many small complexes, and improve the method by Ozawa et al. for verifying candidate complexes predicted by graph clustering methods. In addition, we use maximal components and extreme sets that are defined based on edge connectivity in graph theory. Since the internal proteins of a maximal component are connected more strongly with each other than with any other external proteins as well as an extreme set, they are expected to be useful to further increase the precision.

Method and Results: We propose a novel formulation of integer linear programming (IPc) based on the idea that a candidate complex should not be divided into many small complexes [1], and improve their original method (IPo). Since the problem of maximizing the size of a connected component as well as that of maximizing the number of protein-protein interactions can be proved as NP-hard, we use integer linear programming for solving the problem. However, we use an approximate reduction method because it is difficult to compactly formulate the problem as an integer linear program. Let P and D be a set of candidate proteins for constituting a complex, and a set of domains included in the proteins of P, respectively, where each domain $i:k \ 2 \ D$ is distinguished by the protein i that the domain k belongs to. Let IP, ID, and $IDi;j$ be a set of potentially interacting protein pairs, a set of potentially interacting domain pairs, and a set of potentially interacting domain pairs between proteins i and j, respectively. Then, we approximate the problem of maximizing the size of a connected component of proteins into that of maximizing the number of connected components with size three. This approximated problem can be simply transformed into the integer linear program shown in Fig. 1. In the inequalities, each variable of $xi;j;k$, $pi;j$, and $di:k;j:l$ takes 0 or 1. $xi;j;k = 1$ if and only if proteins i, j, and k are connected. $pi;j = 1$ if and only if proteins i and j interact. $di:k;j:l = 1$ if and only if domains $i:k$ and $j:l$ interact. For validating the performances of verification methods, we used WI-PHI as data of protein protein interactions, iPfam database (version 21.0) as data of potential domain-domain interactions, and a comprehensive catalog of yeast protein complexes CYC2008 as a true set of protein complexes. For obtaining candidate protein complexes, we applied MCL with varying the inflation parameter from 1.5 to 2.5 to WI-PHI. Fig. 2 shows the results of the precision by IPo, IPc, maximal components, extreme sets, maximal+IPc, and extreme+IPc, where the precision was calculated as $jfc \ 2 \ Cj9k \ 2 \ K \ jc \setminus kj=\sqrt{jcj \_ jkj \_ 0:5gj=jCj}$ for sets of predicted and known protein complexes, C and K. The results suggest that our proposed IP-based methods, especially extreme+IPc, considerably outperform the original IP-based method.

Conclusion: We have addressed the problem of verification of candidate protein complexes, and proposed an improved integer linear programming (IP)-based method by introducing the size of a connected component. The results of computational experiments suggest that our proposed methods outperform the existing IP-based method. However, as a future work, it remains to find a compact formulation of the problem of maximizing the size of a connected component because we solved this problem approximately.

References

[1] Zhao, Y., Hayashida, M., Nacher, J., Nagamochi, H. and Akutsu, T., Protein complex prediction via improved verification methods using constrained domain-domain matching, *Proc. 10th Asia-Pacific Bioinformatics Conference*, 394–406, 2012

Comparative Analysis of Antigenic Variant Gene Families of *Plasmodium* Species

Kazushi Hiranuka[1], Diego Diez[2], Nicolas Joannin[1], Susumu Goto[1]

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan
[2] Bioinformatics and Genomics Laboratory, WPI Immunology Frontier Research Center, Osaka University, Japan

Incidences of human and livestock pathogens (viruses, bacteria, and protozoan parasites) have been a serious problem all over the world. Some of them cause untreatable diseases responsible for a massive number of casualties. Antigenic variation is a key phenomenon for these pathogens to evade host-immune responses or vaccines. Among these pathogens, the protozoan parasites have been found to possess relatively large number of antigenic variant multigenes within a single genome and thought to expand the size of variant pool via complex molecular mechanisms including subtelomeric recombination, segmental gene conversion, and duplicative transposition, leading to population-wide antigenic polymorphism. A well-studied example of antigenic variant gene family, PfEMP1(var) of *Plasmodium falciparum*, causative agent of severe malaria in humans, has been shown to have hyper-variable regions that consist of polymorphic domains with different numbers of repeats. Besides such variable regions, PfEMP1 members also share highly conserved regions at the C-terminus separated by transmembrane regions. Non-coding flanking regions are also relatively conserved, and interestingly they have some cluster property which is thought to reflect hybridization affinity upon recombination or gene conversion. Since such molecular mechanisms are expected as a major force of antigenic polymorphism, importance of the cluster property in the antigenic variant gene families has been recognized. Unfortunately, other antigenic variant gene families have significantly different sequence architectures from each other, and even within a family there is high heterogeneity not only at the sequence level but also at the domain-structure level, both of which make them unsuitable to perform multiple alignment and clustering. As a result, extraction of informative regions is difficult without detailed knowledge of sequence anatomy, and thus has not been comprehensively conducted in many antigenic variant gene families. In this study, we have performed comparative analysis on different antigenic variant gene families of *Plasmodium*: var, rifin, vir, kir, cir, bir, and yir. For each family, heterogeneity of sequence conservation was investigated based on non-synonymous and synonymous substitution rate (*dN* and *dS*). Biased distribution of base substitution rates along a coding sequence was observed, indicating of acute selective pressure acting on specific regions. In addition, prevalence of hyper-conserved regions where both *dN* and *dS* are significantly low was confirmed, promoting a new insight on the evolutionary pathway of these families.

Glycolysis in hepatocytes and hepatomas

Alexandra Iovkova

Max-Delbrück-Center for Molecular Medicine Berlin, Germany

Hepatocellular carcinoma (hepatoma) is the most common type of primary liver cancer. Malignant cells proliferate at high rates, showing an excessive glycolysis rate compared to hepatocytes. Pyruvate, the end product of glycolysis, is subsequently reduced to lactate without entering the Krebs cycle. Together, the high rate of aerobic glycolysis and the lactate fermentation in cancer cells are referred to as the Warburg effect. By contrast, hepatocytes, which make up around 70-80% of the liver's mass, are a very altruistic cell type of central importance for the glucose homeostasis in the body. Under aerobic conditions healthy liver cells engage in the Krebs cycle (with Acetyl-CoA derived mainly from fatty acids) as their main source of ATP. In order to maintain glucose homeostasis, hepatocytes take up surplus blood glucose and store it in glycogen. In case of plasma glucose depletion, new glucose can be retrieved by using up the glycogen storage or by generating glucose from other substrates in the process of gluconeogenesis. The current focus of our work is to better understand the Warburg effect on a molecular level. In kinetic models for the metabolic strategies of hepatoma and hepatocytes, we will investigate and compare the influence of enzyme expression profiles and differing isoforms on the behavior. Ultimately, the regulation by signaling pathways will be added to the models.

Prediction of Protein Residue-Residue Contacts Using Conditional Random Field Based on Residue Coevolution

Mayumi Kamada[1], Morihiro Hayashida[2], Jiangning Song[2] Tatsuya Akutsu[1]

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan
[2] Department of Biochemistry and Molecular Biology, Monash University, Australia

Protein function depends on its tertiary structure and interactions between protein residues.
Compensatory mutations arising from functional, structural and folding constraints during the evolutionary process lead to residue coevolution at important sites for interaction. In fact, statistical analysis of Lahn and coworkers [1] revealed that interacting residues tend to coevolve. Many investigations have been conducted to identify coevolving residues and several studies have shown that mutual information (MI) between residues, which is calculated from the distribution of amino acids in multiple sequence alignments for homologous proteins, is useful for predicting interacting residues. In our previous work, we proposed a sequence-based prediction method for protein residue-residue contacts based on this idea [2]. It is a combination of MI and the discriminative random field (DRF), which can recognize specific characteristic regions in an image, and is a special type of conditional random field (CRF) [3]. Although our model showed its usefulness in predicting protein residue contacts, it had a problem with the potential function in DRF. If the potential function includes interaction potentials representing relationships between neighbor residue pairs, a parameter estimation method did not converge. It is thought to be due to strong associations between images and DRF.
In this work, we improve our model using a 2-dimensional CRF instead of DRF to deal with interaction potentials. Moreover, we use additional features representing amino acid properties as inputs. For the purpose of validating the improved model, we perform computational experiments using datasets from PDB and compare with DRF model and other methods. The results show the usefulness of the model with 2-dimensional CRF and additional features for residue contacts prediction.

References

[1] SS. Chi, W. Li, BT. Lahn, Robust signals of coevolution of interacting residues in mammalian proteomes identified by phylogeny-aided structural analysis, *Nature Genetics*, vol. 37, pp. 1367-1371, 2005.
[2] M. Kamada, M. Hayashida, J. Song, and T. Akutsu, Discriminative random field approach to prediction of protein residue contacts, *in Proc. 2011 IEEE international Conference on System Biology*, pp. 285-291, 2011.
[3] S. Kumar and M. Hebert, Discriminative random fields, *International Journal of Computer Vision*, vol. 68, no. 2, pp. 179-201, 2006.

Supporting scientific workflows - a lightweight approach

Maciej M. Kańduła, Paweł P. Łabaj, David P. Kreil

Boku University Vienna, Austria

With the increasing complexity and the rapidly growing amounts of data in the modern life sciences, the supervision and quality control of individual computational analysis steps is becoming a considerable challenge in the field of bioinformatics and related disciplines. While software systems have been developed to support data analysis pipelines, they typically form a useful platform for the automation of routine data analysis steps as common in industrial or facility settings. The integration of new tools and data, however, is often complicated and the generation of new workflows can be very time consuming, and also limited by graphical user interfaces. With little support for rapid prototyping and version control, most systems lack the flexibility required in scientific research. We here study three modern lightweight workflow management tools – MOA, Bpipe, and Ruffus. We present a comparison using the bioinformatics use case of building, executing, and extending a sequence analysis pipeline. Special consideration is given to the question of how policy and rule based handling of constraint enforcement, assertions, and exceptions could be incorporated in these systems.

Label Propagation through Graph-based Feature Reconstruction

Masayuki Karasuyama, Hiroshi Mamitsuka

Bioinformatics Center, Kyoto University, Japan.

We propose a new method for estimating node labels from given instances (nodes) with a graph, particularly focusing on the estimation of the graph edge weights. We estimate edge weights through hyper-parameter optimization of a *harmonic Gaussian field* model for feature vectors, which we call *feature vector propagation* (FVP). FVP defines edge weights as a parameterized similarity function and optimizes edge hyper-parameters by cross-validation over feature vectors of all nodes. That is, the optimization is independent of labeled instances, leading to several important advantages, such as the robustness against sparsely labeled graphs and the applicability to multi-class problems. FVP can also capture the local structure of data by the objective function which shares the same form as the local reconstruction error in *locally linear embedding*. Experimental results demonstrated the effectiveness of FVP both in synthetic and real datasets.

Detecting Modulating Factors Causing Gene Network Alterations by Structural Equation Models

Yuto Kataoka

University of Tokyo, Japan

Recently, for estimating gene networks from RNA expression data, various kinds of computational algorithms have been intensively investigated and proposed with successful applications. Most gene network prediction algorithms estimate a gene network from a gene expression dataset; it is considered that this gene network is common to all samples. On the other hand, an algorithm called NetworkProfiler can predict sample-specific gene regulatory networks from gene expression data of multiple types of samples by using the information from a modulator that shows a sample characteristic. To use this algorithm, we have to prepare one modulator *a priori*, it is not, however, guaranteed that we could always set appropriate modulator. We thus consider preparing a number of candidate modulators, and from them we propose a novel computational method for choosing modulators that have the potential to affect the target gene network. We newly define a score that measures the amount of network changes for evaluating the effect of modulator. The modulators that significantly contribute the network changes are possibly master-regulators, which activate or inhibit target gene networks; this information would be used to control the gene networks. We applied this method to RNA expression dataset of 1511 breast cancer patients in Gene Expression Omnibus and found significant modulators.

CySBML: a Cytoscape Plugin for SBML

Matthias König, Andreas Dräger, Hermann-Georg Holzhütter

Institute of Biochemistry, University Medicine Charité Berlin, Germany

SBML is a free and open interchange format for computer models of biological processes currently supported by over 230 software tools. SBML is used to represent models for a wide range of cell biology, including cell signaling, metabolism and gene regulation. SBML provides a common standard of interoperability and exchange allowing several researchers to work with diverse tools on building, curation, annotation, simulation, analysis, and visualization of the same model. Cytoscape, a widely used open-source platform for complex network analysis and visualization, currently only provides rudimentary SBML capabilities. We present CySBML, a Cytoscape SBML plugin supporting all versions and levels of SBML, handling models in SBML and the SBML Qualitative Model format, including validation of imported SBML files, providing a navigation menu based on SBML structure and easy access to BioModels via web services. Special focus was put on making annotation information and the semantic layer accessible to the user and linking these data to additional web resources. CySBML integrates seamlessly with the Cytoscape ecosystem making features from a multitude of plugins accessible for SBML models, like for instance, search of network motifs (NetMatch), analysis of topological parameters (NetworkAnalyzer) or visualization of flux distributions (FluxViz).

http://www.charite.de/sysbio/people/koenig/software/cysbml

A tool for improving RNA-Seq precision

Paweł P. Łabaj, David P. Kreil

Chair of Bioinformatics, Boku University Vienna, Austria

With currently available RNA-Seq pipelines, expression estimates for most genes are very noisy. We here introduce MapAl, a tool for fast and straightforward expression profiling by RNA-Seq that builds on the existing tools. In the post-processing of RNA-Seq reads, MapAl incorporates gene models already at the stage of read alignment, increasing the number of reliably measured known transcripts consistently by 50%. Adding genes identified de-novo then allows a reliable assessment of double the total number of transcripts compared to other available pipelines.
This substantial improvement is of general relevance: Measurement precision determines the power of any analysis to reliably identify significant signals, such as in screens for differential expression, independent of whether the experimental design incorporates replicates or not.

MapAl supports both users and further development by giving a free choice of combining alternative steps at different stages of the process. In particular, a wide range of read mappers supporting the standard SAM format can be employed. With the new release we have also improved the handling of exon junctions, especially when reads spanning multiple splice junctions. These reads are particularly powerful in the discrimination of specific spliceforms and with read lengths of modern platforms ever increasing, reads spanning multiple splice junctions are becoming a more frequently observed issue.

Phylogenetic Analysis of Eicosanoid Biosynthesis Enzymes

Sayaka Mizutani, Toshiaki Tokimatsu, Susumu Goto

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan

Eicosanoids are a major class of lipid mediators (lipids that act as signaling molecules). They are the oxygenated products of membrane fatty acids such as arachidonic acid. They play important roles in the mammalian physiological functions, including inflammation, immune function and reproduction. Eicosanoid biosynthesis pathways are well-studied in mammals. Upon its release from the membrane phospholipids, arachidonic acid (AA) is modified by distinct enzymatic reaction pathways including the cyclooxygenase (COX) pathway, where AA is oxygenated by COX, then further converted to several different forms of prostanoids by the actions of different enzymes (Table 1). They are the members of several distinct protein families and well-defined by Pfam database [1]. COX products have been biochemically identified in organisms of several invertebrate phyla. Several lines of evidence suggest the requirement of COX activities in these organisms' physiology (See [2], for example). However, invertebrate eicosanoid biosynthetic enzymes remain mostly unidentified with a few exceptions in Corals and Crustaceans. Surprisingly, there are no mammalian COX homologs identified in terrestrial Arthropods. We performed a phylogenetic analysis to investigate the genomic repertoires of enzymes involved in the COX pathway in eukaryotic organisms with complete and nearly complete genomes stored in the KEGG database [3]. Since eicosanoid biosynthetic enzymes are often the members of the large protein families, phylogenetic analyses need a careful enzyme family analysis. This process often requires an identification of sequence motifs that can differentiate homologs of a specific protein from other family members. In this workshop, we present a recent result of our family analysis on Animal heme peroxidase family, a large enzyme family which contains COX and several other peroxidases. Resulted motifs were found to be specific for enzymes with lipid substrates, and are expected to distinguish COX homologs from other peroxidases.

Table 1: Enzyme families / domains defined by Pfam

Enzyme	Description	Enzyme/domain family defined by Pfam
COX	Cyclooxygenase	Animal heme peroxidase
mPGES1	Prostaglandin E synthase 1	Membrane-Associated Proteins in Eicosanoid and Glutathione metabolism (MAPEG)
mPGES2	Prostaglandin E synthase 2	Glutatione S-transferase C term
LPGDS	Lipocalin-type prostaglandin D2 synthase	Lipocalin
HPGDS	Hematopoietic prostaglandin D synthase	Glutatione S-transferase N/C terms
PGFS	Prostaglandin-F synthase	Aldo/keto reductase
PGE2 9-red	Prostaglandin 9-ketoreductase	NADP-dependent dehydrogenases/reductases
TBXAS1	Thromboxane-A synthase	Cytochrome P450 (5A1)
PTGIS	Prostacyclin synthase	Cytochrome P450 (8A1)

References

[1] Punta, M., *et al.*, The Pfam protein families database, *Nucleic Acids Res*, **40**:D290-D301, 2012.

[2] Rowley, F.R., Claire, L.V., Taylor, G.W., and Clare, A.S., Prostaglandins in non-insectan invertebrates: recent insights and unsolved problems, *The Journal of Experimental Biology*, **208**:3-14, 2005.

[3] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M., KEGG for integration and interpretation of large-scale molecular datasets, *Nucleic Acids Res*, **40**, D109-D114, 2012.

A Clique-Based Method Using Dynamic Programming for Unordered Tree Edit-Distance Problem

Tomoya Mori[1], Takeyuki Tamura[1], Daiji Fukagawa[2], Atsuhiro Takasu[3], Etsuji Tomita[4], Tatsuya Akutsu[1]

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan
[2] Faculty of Culture and Information Science, Doshisha University, Japan
[3] National Institute of Informatics, Tokyo, Japan
[4] University of Electro-Communications, Tokyo, Japan

Many kinds of tree-structured data such as RNA secondary structures have become available due to the progress of techniques in the field of molecular biology. In order to analyze the tree-structured data, various measures for computing the similarity between them have been developed. Above all, tree edit distance is one of the most widely used measures. The tree edit distance is defined as the minimum cost sequence of edit operations (deletion, insertion, substitution) that transform one tree into the other. However, the tree edit-distance problem for unordered trees is NP-hard [1]. Therefore, it is required to develop efficient algorithms for the problem. Recently, Fukagawa *et al*. have proposed a clique-based method [2] for computing the tree edit distance between unordered trees in which each instance of the tree edit distance problem is transformed into an instance of the maximum vertex-weighted clique problem and then an existing clique algorithm is applied. The method was applied to comparison of similar glycan structures and shown to be efficient for moderate-size tree structures. However, it was not fast enough for large glycan structures. In this work, we present an improved clique-based method for the tree edit-distance problem for unordered trees. The improved method is obtained by introducing a dynamic programming scheme and heuristic techniques to the previous clique-based method (details are given in [3]). For evaluating the efficiency of the improved method, we also applied the method to comparison of glycan structures. As a result, for large glycan structures, the improved method is much faster than the previous method. In particular, for hard instances, the improved method achieved more than 100 times speed-up.

References

[1] Zhang, K., Statman, R., Shasha, D.: On the editing distance between unordered labeled trees, *Information Processing Letters*, **42**:133-139, 1992.
[2] Fukagawa, D., Tamura, T., Takasu, A., Tomita, E., Akutsu, T.: A clique-based method for the edit distance between unordered trees and its application to analysis of glycan structures, *BMC Bioinformatics*, Suppl. for APBC 2011.
[3] Mori, T., Tamura, T., Takasu, A., Tomita, E., Akutsu, T.: A clique-based method using dynamic programming for computing the edit distance between unordered trees, *Journal of Computational Biology*, in press.

FTMAP: Extended Protein Mapping with User-Selected Probe Molecules*

Scott E. Mottarella[1], Chi Ho Ngan[2], Tanggis Bohnuud[1], Dmitri Beglov[2], Elizabeth Villar[2], David R. Hall[2], Dima Kozakov[2], Sandor Vajda[2]

Bioinformatics Graduate Program, Boston University, Boston, MA

Protein-binding hot spots, or regions with high binding affinity, are usually identified by X-ray crystallography or NMR methods that attempt to capture the position and pose of bound ligands. This process involves screening proteins against small libraries of organic molecules in an expensive and time-consuming manner. FTMAP (http://ftmap.bu.edu/param) is an *in silico* method that mimics these experiments and allows for fast and simple hot spot discovery. FTMAP samples the surface of a target protein using small organic molecules as probes, finds favorable positions, clusters the conformations, and ranks the clusters on the basis of the average energy. Recent additions to FTMAP allow for small molecules of the user's choosing to serve as probes, resulting in accurate depictions of the bound orientation and conformation of the ligand. Bound poses of these ligands may be considered relevant if they overlap the hot spots detected by FTMAP. For each ligand, multiple conformations that are both low in energy and minimally different from other conformations are considered. This allows for a better representation of physical space where ligands may take on a less common conformation to bind a protein. This approach helps predict bound poses that can serve to better describe a protein or be used to develop larger ligands.

*Work performed in the laboratory of Professor Sandor Vajda in the Department of Biomedical Engineering

Extraction of Reaction Sequence Motifs from Metabolic Pathways

Ai Muto, Masaaki Kotera, Toshiaki Tokimatsu
Zenichi Nakagawa, Susumu Goto, Minoru Kanehisa

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan

Metabolic pathway is one of the most important biological networks. In nature, most biological products are synthesized through reaction sequences in which various enzymes contribute. When we look into reaction sequences in metabolic pathways, we can find reaction sequences sharing the same sequence of transformation patterns. Previous studies suggested the existence of functional units (or modules) of reactions in metabolic pathways [1, 2]. Our motivation of this study is to detect such pathway modules for the use of pathway prediction, and also for figuring out the principle of pathway evolution. There have been many attempts to find these modules. Most of them are either sequence-based or enzyme-based approaches. Sequence-based approaches define the reaction-similarity based on the evolutionary or domain-sharing relationships between enzymes, although there are many cases where similar reactions are catalyzed by different enzymes with diverse sequence similarity, and furthermore, this type of approaches is not applicable for the enzymes whose sequences are not identified. On the other hand, most enzyme-based approaches are dependent on the IUBMB's Enzyme List (sometimes referred to as the EC numbers), which needs manual assignment and consensual decision making [3]. Therefore this type of approaches cannot deal with the reactions that are only predicted based on the metabolomic-scale studies. Accordingly, many metabolic pathways that are specific for an organism or a group of organisms, sometimes referred to as secondary metabolisms, have not been subjected to the pathway module analyses. We have been developing computational methods to analyze metabolic reactions focusing on the chemical transformation patterns of metabolic compounds in KEGG [4]. Here we show an approach to identify the functional modules of metabolic pathways, by means of a reaction-comparison method based on the structural transformation of metabolic compounds. For this purpose, we used the KEGG RCLASS database (May 24, 2012), which contains the Reaction Classes classifying 2,481 conformational changes around reaction center atoms of 7,875 reactant pairs of 7,978 enzymatic reactions. We grouped Reaction Classes into 1,566 groups including 1,190 singletons. Then, we searched reaction sequences sharing the same sequence of grouped Reaction Classes (Reaction sequence pattern) in the metabolic pathways using the KEGG PATHWAY database. As the result, we obtained 2,444 reaction sequence patterns that consist of up to eight reactions. Subsequently, we manually organized reaction sequence motifs from all the reaction sequence patterns. For example, we obtained different reaction sequence motifs between aerobic and anaerobic degradation of aromatic rings. Furthermore, we found aerobic degradation motifs are classified into four types. We would like to report the result of characterization of the reaction sequence motifs that give insights on some noteworthy patterns from biochemical aspects.

References

[1]Oh M et al., J Chem Inf Model. 2007 Jul-Aug;47(4):1702-12.
[2]Tohsato Y et al., Proc Int Conf Intell Syst Mol Biol. 2000;8:376-83.
[3]Tipton K et al., Bioinformatics. 2000 Jan;16(1):34-40.
[4]Kanehisa M et al., Nucleic Acids Res. 2006 Jan 1;34(Database issue):D354-7.

Comparative and functional analysis of intragenic miRNAs in metazoan genomes

Yosuke Nishimura, Masaaki Kotera, Toshiaki Tokimatsu, Susumu Goto,

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan

MicroRNAs (miRNAs) are small (~22 nt) non-coding RNAs present in diverse organisms and regulate the expression of thousands of genes via the RNA silencing machinery. miRNAs are key regulators in gene expression networks and implicated in control of signal transduction, development, cellular differentiation, metabolism, and adaptation to the environment. They also have an important influence on pathologic states in multiple diseases. Since miRNAs are processed from one or two arms of the hairpin precursor (pre-miRNA), two different miRNAs are occasionally generated from the same precursor [1]. Hundreds or more than a thousand of miRNAs have been discovered in various metazoan genomes including *C. elegans*, *D. melanogaster*, and many vertebrate species. A large part of these sequence is discovered by recent deep-sequencing technology. Thus, functions of most miRNAs have remained unclear. It is known that around 50% of metazoan miRNAs are overlapping with introns, exons, or UTR of annotated genes (These miRNAs and genes are called respectively as intragenic miRNAs and host genes) and they are largely transcribed by RNA polymerase II as the same part of transcription unit of the host gene. Suppose intronic miRNAs have specific functions and are important as key regulator in some cellular condition, it would not be surprising that each intronic miRNA has a functional association with its host gene through regulating the host gene itself or a set of genes which is in the same functional module or complex of the host gene. We call them "direct regulation model" and "functional regulation model", respectively. Though a previous study [2] suggested that targets of human intragenic miRNAs are enriched in the pathway of the host genes, their analysis of functional association of intragenic miRNAs with their host genes is restricted to human and also restricted to find which pathway is remarkably associated on, because the statistical tests are not focused on each intragenic miRNA, but on each pathway. To shed light on functional association of each intragenic miRNAs of various metazoan species with their host genes, we collected a list of intragenic miRNAs and their host genes from 10 species (*H. sapiens*, *P. troglodytes*, *C. familiaris*, *B. taurus*, *M. musculus*, *R. norvegicus*, *G. gallus*, *D. rerio*, *D. melanogaster*, and *C. elegans*). miRNAs are from miRBase v16 and genes are from RefSeq rel.46. miRNA Target prediction has been performed on our calculation server by using four published prediction implementations: miRanda, TargetScan, PITA, TragetSpy. We regarded genes predicted by three or four implementations as positive targets of each miRNA. We designed statistical tests on each intragenic miRNA and evaluated its functional association with the host gene in the light of both the direct regulation model and the functional regulation model. To analyze the functional regulation model, target enrichment tests were performed on KEGG PATHWAY and BRITE. Host gene orthology fetched from NCBI HomoloGene was used to compare these functional associations between species. We are currently checking the details.

References

[1] Griffiths-Jones, S., Hui, J.H.L., Marco, A., and Ronshaugen, M., MicroRNA evolution by arm switching, *EMBO Rep.*, 12(2):172-177, 2011

[2] Hinske, L.C.G., Galante, P.A.F., Kuo, W.P., and Ohno-Machado, L., A potential role for intragenic miRNAs on their hosts' interactome, *BMC Genomics*, 11:533, 2010

Challenges in Benchmarking Joint miRNA - Gene Expression Analysis

Smriti Shridhar, David P. Kreil
Boku University Vienna, Austria

MicroRNAs are small non-coding RNAs that play an important role in gene regulation, binding to specific target mRNAs. The regulatory networks of interest can be complex because one miRNA may target many different genes, and one gene may be targeted by many miRNAs. With the easy availability of genome scale experiments, there have been increasing efforts in untangling these molecular interaction networks systematically. Considerable improvements are expected from the integration of data from different experiments. A number of computational approaches have been proposed that predict potential miRNA–gene interactions by combining data from multiple sources. We have observed, however, that there is nearly no overlap between miRNA–gene interactions predicted by the different methods. This motivated a systematic comparison of prediction efficiencies of the latest algorithms using different benchmark sets and approaches.

Current challenges in benchmarking integrative approaches for inferring regulatory interactions between miRNA and genes are discussed. In particular, we apply alternative benchmarks for the assessment of miRNA–gene associations as predicted from the joint analysis of miRNA and gene expression profiles. Benchmarks include comparisons to independent predictions from sequence analysis, genomic miRNA clusters, experimentally validated miRNA–gene interactions, as well as dataset specific enrichment, such as cancer associated miRNAs and genes. Randomize data served as null model. Intriguingly, most methods struggle to show performance above background levels. It turns out that only the experimentally validated miRNA–gene interactions provide consistent results. Extensions of this benchmark set, especially by known negatives, will be crucial for further method development.

DNase-Seq analysis reveals dynamic effects of growth hormone pulses activating STAT5 on local chromatin structure in male mouse liver.

George Steinhardt[1], Andy Rampersaud[1], Aarathi Sugathan[1], Guoyu Ling[2]
Jeanette Connerney[2], David J. Waxman[1,2]

[1] Bioinformatics Graduate Program, Boston University
[2] Department of Biology, Boston University, Boston, MA

Sex differences in gene expression characterize >1,000 genes in mouse liver, impacting many liver functions including lipid metabolism and drug detoxification. This dimorphism regulates Sex-dependent patterns of pituitary growth hormone (GH) secretion. The transcription factor STAT5 is activated by GH intermittently in male liver but persistently in female liver and is essential for liver sex differences. We investigated whether or not the intermittent pulses of STAT5 activity dynamically alter chromatin structure in male mouse liver. Livers collected from individual male mice were assayed for GH-activated STAT5 activity (STAT5-high vs. STAT5-low livers) by EMSA analysis. Open chromatin regions (DNase hypersensitive sites, DHS) were identified by high-throughput sequencing of DNase-released fragments (DNase-seq). Cumulative digital footprint analysis was used to validate the STAT5 binding status of individual livers. DHS peaks present in two validated STAT5-high and three validated STAT5-low livers were compared quantitatively using the peak normalization algorithm MAnorm. We thus identified 1,492 DHS sites >2-fold more accessible in STAT5-high livers and 300 DHS sites more accessible in STAT5-low livers. The 1,492 STAT5-high DHS sites were strongly enriched (6.8-fold) in the set of male-biased DHS sites and were 2-fold depleted at sex-independent DHS sites, indicating that a subset comprised of 440 male-biased DHS is dynamically opened in male liver with each plasma GH pulse. The 1,492 DHS sites were also strongly enriched (6.2-fold) in male-enriched STAT5 binding sites discovered by ChIP-seq, linking the dynamic effects of plasma GH pulses to chromatin opening and male-enriched STAT5 binding. Cumulative digital footprint analysis revealed a unique signature of STAT5 binding to liver chromatin, characterized by peaks and troughs in DNase accessibility at specific nucleotide positions relative to the STAT5 consensus motif TTC-NNN-GAA. In particular, STAT5 binding altered DNase hypersensitivity and chromatin structure in an asymmetric manner over a region encompassing the core STAT5 binding motif and 6 additional nucleotides on each flank. We conclude that male plasma GH pulses dynamically alter mouse liver chromatin structure at male-biased DHS sites, where STAT5 binding is male-enriched. DNase-seq can thus reveal effects of STAT5 binding on chromatin structure, and may be generally applicable to other context-dependent factors and their interactions with chromatin remodeling proteins. *Supported in part by NIH grant DK33765.*

Comparative Study on Frequent Itemset Mining-based Biclustering Methods

Kei-ichiro Takahashi[1], Ichigaku Takigawa[2], Hiroshi Mamitsuka[1]

[1] Bioinformatics Center, Kyoto University, Japan
[2] Creative Research Institute, Hokkaido University, Japan

Gene expression data have been accumulating in public repositories such as GEO or ArrayExpress, and the demand to discover biologically significant information from those data has been increasing. Standard analysis on an expression matrix involves clustering such as hierarchical clustering or k-means, which divides the whole expression matrix into multiple groups so that similar vectors are grouped together in a row or column direction. However, those standard methods often have difficulty in extracting functional genes under specific experimental conditions, because genes rarely show similar expression across a wide range of conditions. Biclustering methods address this problem. Biclustering finds biclusters in an expression matrix which are defined as certain types of clusters with only part of both rows and columns. We have developed a new biclustering method based on maximally frequent itemset mining by converting an expression matrix into a fine-grained transaction database- a typical database structure used in frequent itemset mining. We compared our method with other frequent itemset mining-based methods, confirming the performance advantage of our method over the other competing methods in terms of GO enrichment analysis.

Stability and Restoration in Canadian Lynx and Snowshoe hare Population Cycle

Lisa Uechi, Tatsuya Akutsu

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan

The concept of the stability is important in order to understand natural phenomena in biological and engineering systems. We study the relation between the conservation law and the stability of a 2*n*-dimensional competitive system that contains competitive interactions, self-interactions and mixing interactions. The system consists of 2*n*-dimensional nonlinear differential equations required from Noether's theorem [1]. The 2*n*-dimensional nonlinear ordinary differential equations for the competitive system constructed to satisfy the conservation law have properties such as the addition law, which is empirically interpreted as recovery from injuries of skin and tissues in biological bodies. The interaction between Canadian lynx and snowshoe hare has a well-known characteristic property as a prey-predator system. The Canadian lynx and snowshoe hare have a synchronous ten-year cycle in population numbers [2]. The structure of this cycle is related to the several factors such as nutrient, predation and social interactions [3], but the fundamental mechanisms have not been clarified. We emphasize that the properties of the system which has the conserved quantity are a key to understand the unanswered question for the existence of this cycle. In this work, we discuss a system of interactions generalizing Lotka-Volterra type nonlinear competitive interactions and explain that the conservation law is important to understand complex phenomena even in biological and ecological systems. The binary-coupled form system has several properties, and in this study, by simulating external perturbations, we explicitly and numerically discuss the properties of the conserved stable 2-variable nonlinear interacting system, its indications and possible applications to nonlinear interacting systems. Furthermore, we show that 2-variable ND model which was advocated in previous work has the properties of restoration and recovery from external perturbations.

References

[1] L. Uechi and T. Akutsu. Conservation laws and symmetries in competitive systems. *Progress of Theoretical Physics Supplement*, (194), 2012.
[2] N.C. Stenseth,W. Falck, K.S. Chan, O.N. Bjornstad, M. O'Donoghue, H. Tong, R. Boonstra, S. Boutin, C.J. Krebs, and N.G. Yoccoz. From patterns to processes: phase and density dependencies in the canadian lynx cycle. *Proceedings of the National Academy of Sciences*, **95**(26):15430, 1998.
[3] C.J. Krebs, R. Boonstra, S. Boutin, and A.R.E. Sinclair. What drives the 10-year cycle of snowshoe hares? *BioScience*, **51**(1):25—35, 2001.

Finding Mutations in Cancer by Bayesian EM with Haplotype Sequences

Naoto Usuyama

University of Tokyo, Japan

High-throughput DNA sequencing enables us to analyze cancer-specific genetic alterations in whole genomes. Simply, we can detect candidate mutations by comparing tumor and matched normal genomes of a patient. Previous methods basically align reads, make DNA pileup and check the differences between tumor and normal. However, there are some difficulties for alignment algorithms in regions with long indels and repetition. Especially, when there exist long indels, alignment algorithms often produce soft clippings that are unaligned subsequences of reads. Pileup-based methods cannot show high performance in these regions because of inaccurate alignment. To improve the performance on these regions, we propose a Bayesian model selection approach for mutation calling. First, we prepare a set of candidate haplotype sequences by using mapping and pileup algorithms. We then estimate the source haplotype of each read by mapping read to haplotype using profile HMM method. Finally, we infer haplotype frequencies by variational Bayesian EM algorithm in tumor and normal genomes and statistically evaluate their differences by the Bayes factor. By realigning reads and selecting a haplotype from which the reads come, our proposed method succeeded in classifying tumor and normal reads in the difficult regions described above. We apply our method to both simulated data and real data to illustrate its efficiency.

Mathematical modeling of lipid droplets dynamic in hepatocytes

Christin Wallstab

Institute of Biochemistry, University Medicine Charité Berlin

The liver is a major organ responsible for lipid and glucose metabolism. Defects in lipid metabolism can lead to obesity, insulin resistance and liver cirrhosis. Several drugs, alcohol, malnutrition and adiposity cause fatty liver diseases. Hepatocytes, amongst other cells, store excess lipids as lipid droplets (LD). LDs play a crucial role in the fatty liver disease, which is characterized by abnormal LD content in hepatocytes. Non-Alcoholic-Fatty-Liver-Disease (NAFLD) occurs in a quarter of the western population and can develop toward liver cirrhosis (10% of NAFLD). LDs are intracellular organelles, consisting of a neutral lipids core (i.e. triacylglycerol and cholesterol ester) surrounded by a phospholipid monolayer coating. In this shielding membrane different kinds of coating proteins are integrated serving as dynamic scaffold for the growth of the droplet and controlled lipolysis. The nascent lipid droplets, replete with neutral lipids formed in the membrane of the endoplasmic reticulum (ER), buds off as vesicles from the ER membrane. In the dynamic process of maturation LDs grow while exchanging different coating proteins, which control the process of assembly, enlargement and movement. For example, perilipin controls lipase activity to modulate the lipolystic process, making it a key player in understanding size distribution of LD population. We have developed a descriptive model to simulate the process of LD growth by fusion and degradation by lipolysis from current literature. In future work we will focus on building a mathematical model of the live cycle of LDs including the regulatory properties of different PAT proteins as well as external conditions like malnutrition and hormonal state on the accumulation and utilization of LDs to gain better insights in the processes leading to steatosis.

Transcriptomic Changes in the Oral Mucosal Epithelium Reflect the Host Response to Indoor Coal Smoke Exposure

Teresa Wang[1,2], Bozena Krystyna[2,3], Gang Liu PhD[2], Ji Xiao BA[2], Yuriy Alekseyev PhD[2,4]
Marc Lenburg PhD[1,2,3], Wei Hu[5], Nathaniel Rothmann[5], Qing Lan[5]
H. Dean Hosgood III[5], Avrum Spira MD MSc[1,2,3,4]

[1] Bioinformatics Graduate Program, Boston University, Boston, MA
[2] Section of Computational Biomedicine, Boston University School of Medicine, Boston, MA
[3] Section of Pulmonary, Critical Care and Allergy, Boston University School of Medicine, Boston, MA
[4] Department of Pathology and Laboratory Medicine, Boston University School of Medicine, Boston, MA
[5] Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, US

Rationale: Human exposure to indoor smoke emissions from the domestic burning of solid fuels is a major public health issue. There is a current unmet need for connecting external metrics, such as air quality assessments, to associated measures of the internal host response. Given that gene expression profiles in the upper airway epithelium can reflect tobacco smoke-related responses, we sought to determine whether transcriptomic changes in the oral mucosa epithelium may capture the physiologic host response to indoor coal smoke exposure.

Methods: Buccal mucosa epithelial cell scrapings were collected from healthy female subjects in rural Xuanwei, China. All subjects are exposed to varying levels of coal smoke from the indoor use of unvented firepits for heating and cooking. RNA from 24 samples was extracted, processed and hybridized to Affymetrix Human Gene 1.0 ST arrays. Personal 24-hr filters were provided to each subject to monitor indoor levels of fine particulate matter. The filters were later analyzed using chromatography-mass spectrometry to calculate benzo(a)pyrene concentration (ng/m3), an indicator of carcinogenic polycyclic aromatic hydrocarbons. Genes whose expressions were significantly associated with BAP concentration were identified using a linear model. Functional enrichment of these genes was determined using DAVID and GSEA.

Results: A signature of 270 genes ($p<.01$) was associated with increasing BAP from indoor smoky coal exposure. The 236 down-regulated genes were functionally enriched for oxidative phosphorylation, apoptosis and signal transduction, while the 34 up-regulated genes were enriched for microtubule nucleation and the Toll-like receptor signaling pathway (pDAVID<.001). Furthermore, GSEA results demonstrate that the BAP-associated gene expression is concordantly enriched for genes whose expression levels change in current smokers relative to never smokers. (FDR<0.01)

Conclusion: Buccal epithelial gene expression profiles reflect differences in household coal smoke exposure. These preliminary results are also concordant with independent bronchial epithelial gene expression profiles that distinguish current smokers from never smokers, which suggest that smoky coal emissions may share common biological mechanisms to cigarette smoke. The processing of additional samples from healthy and diseased subjects may help us further elucidate the molecular underpinnings of coal smoke exposure and susceptibility to lung disease.