Abstracts of papers presented
at the 9th IEEE International Workshop on

GENOMIC SIGNAL PROCESSING AND STATISTICS (GENSIPS)

November 10–November 12, 2010

Cold Spring Harbor Laboratory
Cold Spring Harbor, New York

Abstracts of papers presented
at the 9th IEEE International Workshop on

GENOMIC SIGNAL PROCESSING AND STATISTICS (GENSIPS)

November 10–November 12, 2010

Arranged by

General Chair: Nevenka Dimitrova, *Philips Research, Briarcliff Manor*
Program Chair: Gurinder Singh "Mickey" Atwal, *Cold Spring Harbor Laboratory*
Program Chair: Haris Vikalo, *University of Texas, Austin*
Proceedings Chair: Byung-Jun Yoon, *Texas A&M University, College Station*

Technical Program Committee Members

Gurinder Atwal
Ulisses Braga-Neto
Xiaodong Cai
Aniruddha Datta
Nevenka Dimitrova
Edward Dougherty
John Goutsias
Yu-fei Huang
Ivan Ivanov
Seungchan, Kim
Judith Klein-Seetharaman
Rui Kuang
Doheon Lee
Ranadip Pal
Xiaoning Qian
Dan Schonfeld
Chao Sima
Haris Vikalo
Xiaodong Wang
Byung-Jun Yoon

Cold Spring Harbor Laboratory
Cold Spring Harbor, New York

Contributions from the following companies provide core support for the Cold Spring Harbor meetings program.

Corporate Sponsors

Agilent Technologies
AstraZeneca
BioVentures, Inc.
Bristol-Myers Squibb Company
Genentech, Inc.
GlaxoSmithKline
Life Technologies (Invitrogen & Applied Biosystems)
New England BioLabs, Inc.
OSI Pharmaceuticals, Inc.
Sanofi-Aventis

Plant Corporate Associates

Monsanto Company
Pioneer Hi-Bred International, Inc.

Foundations

Hudson-Alpha Institute for Biotechnology

Cover: Double-helix sculpture, Charles Jencks. Photograph by Constance Brukin.

9th IEEE International Workshop on
GENOMIC SIGNAL PROCESSING AND STATISTICS (GENSIPS)
Wednesday, November 10 – Friday, November 12, 2010

Wednesday	2:00 pm	**1** Tutorials
Wednesday	5:00 pm	Wine and Cheese Party
Wednesday	7:30 pm	**2** Network Biology and Pathway Analysis
Wednesday	9:00 pm	**3** Error Estimation in Genomic Data
Thursday	9:00 am	Keynote Speaker
Thursday	10:15 am	**4** Gene Expression Studies in Next Gen Seq
Thursday	1:00 pm	Keynote Speaker
Thursday	2:15 pm	**5** Dynamic Modeling and Regulatory Networks
Thursday	3:30 pm	**6** Poster Session
Thursday	6:00 pm	Cocktails / Banquet
Friday	9:00 am	Keynote Speaker
Friday	10:15 am	**7** Data Challenges in Next Generation Sequencing
Friday	1:30 pm	**8** Classification and Statistical Learning
Friday	2:45 pm	**9** Network Based Methods in Computational Biology

Mealtimes at Blackford Hall are as follows:
Breakfast 7:30 am-9:00 am
Lunch 11:30 am-1:30 pm
Dinner 5:30 pm-7:00 pm
Bar is open from 5:00 pm until late

Printed on 100% recycled paper.

PROGRAM

WEDNESDAY, November 10—2:00 PM

SESSION 1 TUTORIALS

Chairperson: **M. Atwal,** Cold Spring Harbor Laboratory, New York

Genome copy number measurements from hybridization and sequence read data
Michael Wigler, Boris Yamrom, Jud Kendell, Yoon-Ha Lee, Nick Navin, Kenny Ye.
Presenter affiliation: Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. 1

Genetics and population genomics
Lucia Hindorff.
Presenter affiliation: NHGRI, National Institutes of Health, Bethesda, Maryland.

WEDNESDAY, November 10—5:00 PM

Wine and Cheese Party

WEDNESDAY, November 10—7:30 PM

SESSION 2 NETWORK BIOLOGY AND PATHWAY ANALYSIS

Chairperson: **N. Bouaynaya,** University of Arkansas at Little Rock

Pathway and network analysis probing epigenetic influences on chemosensitivity in ovarian cancer
Nilanjana Banerjee, Angel Janevski, Sitharthan Kamalakaran, Vinay Varadan, Robert Lucito, Nevenka Dimitrova.
Presenter affiliation: Philips Research, Briarcliff Manor, New York. 2

WEDNESDAY, November 10—9:00 PM

SESSION 3 ERROR ESTIMATION IN GENOMIC DATA

Chairperson: **D. Schonfeld,** University of Illinois, Chicago

THURSDAY, November 11—9:00 AM

Introduction by: **Nevenka Dimitrova**

KEYNOTE SPEAKER

Andrea Califano
Columbia University

"Elucidating master integrators of tumor-related phenotypes"

THURSDAY, November 11—10:15 AM

SESSION 4 GENE EXPRESSION STUDIES IN NEXT GEN SEQ

Chairperson: **P. Beyerlein,** University of Applied Sciences Wildau, Germany

Studying gene expression and regulation in cancer using Next Generation Sequencing
Gary Schroth.
Presenter affiliation: Illumina, Inc., Hayward, California.

Methylome-transcriptome relationship in the rat peripheral nervous system in health and chronic pain
Andreas Beutler.
Presenter affiliation: Mayo Clinic, Rochester, Minnesota.

Automatic learning from RNA-seq data—Unbiased transcriptome discovery using the Wildau In-silicO Sequence Analysis (WIOS) framework
Peter Beyerlein.
Presenter affiliation: University of Applied Sciences Wildau, Germany.

THURSDAY, November 11—1:00 PM

Introduction by: **Haris Vikalo**

KEYNOTE SPEAKER

Edward R. Dougherty
Texas A&M University

"Intervention in gene regulatory networks" 6

THURSDAY, November 11—2:15 PM

SESSION 5 DYNAMIC MODELING AND REGULATORY NETWORKS

Chairperson: **J. Goutsias,** Johns Hopkins University, Baltimore, Maryland

THURSDAY, November 11—3:30 PM

SESSION 6 POSTER SESSION

THURSDAY, November 11

BANQUET

Cocktails 6:00 PM Dinner 6:45 PM

FRIDAY, November 12—9:00 AM

Introduction by: **Byung-Jun Yoon**

KEYNOTE SPEAKER

W. Richard McCombie
Cold Spring Harbor Laboratory

"Finding a needle in a haystack"

FRIDAY, November 12—10:15 AM

SESSION 7 DATA CHALLENGES IN NEXT GENERATION SEQUENCING

Chairperson: **N. Dimitrova,** Philips Research, Briarcliff Manor, New York

FRIDAY, November 12—1:30 PM

SESSION 8 CLASSIFICATION AND STATISTICAL LEARNING

Chairperson: **A. Tewfik,** University of Minnesota, Minneapolis

FRIDAY, November 12—2:45 PM

SESSION 9 NETWORK-BASED METHODS IN COMPUTATIONAL BIOLOGY

Chairperson: **X. Qian,** University of South Florida, Tampa

GENOME COPY NUMBER MEASUREMENTS FROM HYBRIDIZATION AND SEQUENCE READ DATA

Mike Wigler[1], Boris Yamrom[1], Jud Kendell[1], Yoon-Ha Lee[1], Nick Navin[1], Kenny Ye[2]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724, [2]Albert Einstein College of Medicine, Bronx, NY, 10461

Copy number variation is a rich source of genetic variation. Copy number mutation underlies many inherited and spontaneous genetic disorders in humans as well as being a driving force for progression in cancer.

Measuring copy number variation from either hybridization or sequence data presents challenging problems for quantitative biologists. We will discuss various solutions to the problems.

PATHWAY AND NETWORK ANALYSIS PROBING EPIGENETIC INFLUENCES ON CHEMOSENSITIVITY IN OVARIAN CANCER

Nilanjana Banerjee[1], Angel Janevski[1], Sitharthan Kamalakaran[1], Vinay Varadan[1], Robert Lucito[2], Nevenka Dimitrova[1]

[1]Philips Research, Ultrasound, Photonics, and Bioinformatics, Briarcliff Manor, NY, 10510, [2]Cold Spring Harbor Laboratory, Lucito Lab, Cold Spring Hbr, NY, 12724

Ovarian cancer is the leading cause of death in gynecological cancers. Carboplatinum-based therapy is the standard treatment choice for ovarian cancer. However, a majority of the patients develop resistance to carboplatinum fairly rapidly hence there is a clinical need for early predictors of carboplatinum resistance. While there are a few indicative gene markers, they have poor sensitivity and specificity in predicting response accurately. It is essential that multiple high throughput molecular profiling modalities are integrated and investigated to provide a full picture of the ongoing processes. Here, we propose a methodology to identify central players in platinum resistance from a list of stratifying genes using a data-driven approach. We have used correlation of DNA methylation and gene expression data and applied network based features to identify the influence of DNA methylation on gene expression. This provides interpretive analysis and is complementary to the biological pathway-enrichment approaches. We suggest that our method, based on network structure properties, adds a useful layer to multi-modal evidence integration to focus on the key processes and interactions in resistance mechanisms.

OPTIMAL PERTURBATION CONTROL OF GENE REGULATORY NETWORKS

Nidhal Bouaynaya[1], Roman Shterenberg[2], Dan Schonfeld[3]

[1]University of Arkansas at Little Rock, Systems Engineering, Little Rock, AR, 72204, [2]University of
Alabama at Birmingham, Mathematics, Birmingham, AL, 35294,
[3]University of Illinois at Chicago, Electrical and Computer Engineering, Chicago, IL, 60607

We formulate the control problem in gene regulatory networks as an inverse perturbation problem, which provides the feasible set of perturbations that force the network to transition from an undesirable steady-state distribution to a desirable one. We derive a general characterization of such perturbations in an appropriate basis representation. We subsequently consider the optimal perturbation, which minimizes the overall energy of change between the original and controlled (perturbed) networks. The ``energy'' of
change is characterized by the Euclidean-norm of the perturbation matrix. We cast the optimal control problem as a semi-definite programming (SDP) problem, thus providing a globally optimal solution which can be efficiently computed using standard SDP solvers. We apply the proposed control to the Human melanoma gene regulatory network and show that the steady-state probability mass is shifted from the undesirable high metastatic states to the chosen steady-state probability mass.

APPROXIMATE EXPRESSIONS FOR THE VARIANCES OF NON-RANDOMIZED ERROR ESTIMATORS AND COD ESTIMATORS FOR THE DISCRETE HISTOGRAM RULE

Ting Chen, Ulisses Braga-Neto

Texas A&M University, Department of Electrical and Computer Engineering, College Station, TX, 77840

Estimation of the classification error and of the coefficient of determination (CoD) is a fundamental issue in discrete prediction problems. Analytical expressions for exact performance metrics of non-randomized error estimators and CoD estimators have been derived in previous publications by the authors. However, computation of these expressions becomes problematic as the sample size or predictor complexity increases, particularly in the case of second moments. Thus, fast and accurate approximations are desirable. In this paper, we make approximations to the variances of resubstitution and leave-one-out error estimators and CoD estimators. Our results show that these approximations not only are quite accurate but also reduce computation time tremendously.

BAYESIAN MMSE ESTIMATION OF CLASSIFICATION ERROR AND PERFORMANCE ON REAL GENOMIC DATA

Lori A Dalton[1], Edward R Dougherty[1,2,3]

[1]Texas A&M University, Dept. of Electrical and Computer Engineering, College Station, TX, 77843 , [2]Translational Genomics Research Institute, Computational Biology Division, Phoenix, AZ, 85004, [3]University of Texas M. D. Anderson Cancer Center, Dept. of Bioinformatics and Computational Biology, Houston, TX, 77030

Small sample classifier design has become a major issue in the biological and medical communities, owing to the recent development of high-throughput genomic and proteomic technologies. And as the problem of estimating classifier error is already handicapped by limited available information, it is further compounded by the necessity of reusing training-data for error estimation. Due to the difficulty of error estimation, all currently popular techniques have been heuristically devised, rather than rigorously designed based on statistical inference and optimization. However, a recently proposed error estimator has placed the problem into an optimal mean-square error (MSE) signal estimation framework in the presence of uncertainty. This results in a Bayesian approach to error estimation based on a parameterized family of feature-label distributions. These Bayesian error estimators are optimal when averaged over a given family of distributions, unbiased when averaged over a given family and all samples, and analytically address a trade-off between robustness (modeling assumptions) and accuracy (minimum mean-square error). Closed form solutions have been provided for two important examples: the discrete classification problem and linear classification of Gaussian distributions. Here we discuss the Bayesian minimum mean-square error (MMSE) error estimator and demonstrate performance on real biological data under Gaussian modeling assumptions.

INTERVENTION IN GENE REGULATORY NETWORKS

Edward R Dougherty[1,2,3]

[1]Texas A&M University, Department of Electrical and Computer Engineering, College Station, TX, 77843-3128, [2]Translational Genomics Research Institute, Computational Biology Division, Phoenix, AZ, 85004, [3]University of Texas M. D. Anderson Cancer Center, Department of Bioinformatics and Computational Biology, Houston, TX, 77030

A major reason for constructing gene regulatory networks is to use them as models for determining therapeutic intervention strategies by deriving ways of altering their long-run dynamics. Two paradigms have been taken for gene network intervention: (1) stationary external control is based on optimally altering the status of a control gene (or genes) over time to drive network dynamics; and (2) structural intervention involves an optimal one-time change of the network structure to beneficially alter the long-run behavior of the network. For the most part, these interventions have been studied in the framework probabilistic Boolean networks (PBNs); however, since they mainly depend on the Markov chain associated with a PBN, they are applicable to a wide class of regulatory networks. This talk reviews both types of intervention, including key issues such as complexity reduction and robustness.

AN ITERATED CONDITIONAL MODE SOLUTION FOR BAYESIAN FACTOR MODELING OF TRANSCRIPTIONAL REGULATORY NETWORKS

Jia Meng[1], Jianqiu Zhang[1], Yidong Chen[2], Yufei Huang[1]

[1]University of Texas at San Antonio, Electrical and Computer Engineering, San Antonio, TX, 78249, [2]University of Texas Health Science Center at San Antonio, Department of Epidemiology and Biostatistics, San Antonio, TX, 78229

Abstract—The problem of uncovering transcriptional regulation by transcription factors (TFs) based on microarray data is considered. A novel Bayesian sparse correlated rectified factor model (BSCRFM) coupled with its ICM solution is proposed. BSCRFM models the unknown TF protein level activity, the correlated regulations between TFs, and the sparse nature of TF regulated genes and it admits prior knowledge from existing database regarding TF regulated target genes. An efficient ICM algorithm is developed and a context-specific transcriptional regulatory network specific to the experimental condition of the microarray data can be obtained. The proposed model and the ICM algorithm are evaluated on the simulated systems and results demonstrated the validity and effectiveness of the proposed approach. The proposed model is also applied to the breast cancer microarray data and a TF regulated network regarding ER status is obtained.

A SCREENING METHOD FOR DIMENSIONALITY REDUCTION IN BIOCHEMICAL REACTION SYSTEM CALIBRATION.

W. Garrett Jenkinson, John Goutsias

The Johns Hopkins University, Whitaker Biomedical Engineering Institute, Baltimore, MD, 21218

Estimating the rate constants of a biochemical reaction model of cellular function is an important, albeit computationally intensive, problem in systems biology. In this paper, a variance-based sensitivity analysis approach is proposed, which can be used, as a pre-screening step, to identify parameters in a biochemical reaction system that do not appreciably influence the cost of estimation and, therefore, whose values cannot be precisely determined by parameter estimation. By only estimating the remaining parameters, appreciable qualitative and quantitative improvements can be achieved. A subset of a well-known biochemical reaction model of the EGF/ERK signaling pathway is used to illustrate the benefits achieved by the proposed method.

COOPERATIVE MIRNA TARGET PREDICTION ALGORITHM

Claudia Coronnello[1], Panayiotis V. Benos[1,2]

[1]University of Pittsburgh, Department of Computational and Systems Biology, Pittsburgh, PA, 15260, [2]University of Pittsburgh, Department of Biomedical Informatics, Pittsburgh, PA, 15260

MicroRNAs (miRNA) are a class of short (18-25 nucleotide) non-coding RNAs that regulate gene expression post-transcriptionally. Their regulatory activity depends heavily on the recognition of binding sites located mainly on the 3'-untranslated regions (3'UTRs) of target messenger RNA (mRNA). It is known that one miRNA can regulate many genes and that one gene can be regulated by more than one miRNAs. There are already many algorithms that computationally predict the miRNA targets, by only considering the site-specific factors of targeting. These algorithms are tested against the few available experimentally validated miRNA-target pairs and are able to predict only a portion of them. However, system-level factors also influence the efficiency of the miRNAs regulation, like the miRNAs concentration or the miRNAs targets abundance. We developed a new algorithm that can predict whether the expression of a given gene can be influenced by a set of miRNAs, given detailed cell conditions. This algorithm combines predictions by four existing algorithms, RNAHybrid, PITA, miRanda and TargetScan. The algorithm is trained with published experimental immunoprecipitation (IP) data of the miRISC proteins together with mRNA and miRNA expression data. These experiments provide a picture of the immunoprecipitated miRISC complexes, giving global information about the miRNAs and their mRNA targets. We trained the algorithm with IP enriched genes found in D. melanogaster (Dme) S2 cells and we tested it in predicting the targets of the same dataset (self-test) and on a similar dataset from C. elegans. Results show that our algorithm can predict a significant portion of genes that are up-regulated after miRNAs depletion. We also tested the algorithm on experimentally validated miRNA targets and housekeeping genes in Dme. Results show that our algorithm presents a significant improvement over standard miRNA prediction algorithms. To the best of our knowledge, this is the first algorithm that truly tries to address the problem of combinatorial binding of multiple miRNA genes to a given target.

USING MULTIRESOLUTION TRANSFORMATIONS FOR PREDICTING CLINICAL OUTCOMES FROM GENOME-WIDE DATA

Pablo Hennings-Yeomans, Gregory F Cooper

University of Pittsburgh, Biomedical Informatics, Pittsburgh, PA, 15260

The prediction of clinical outcomes, such as Alzheimer's and cardiac disease, from genome-wide data has often focused on finding one or a few good predictors (e.g., SNPs). In recent years, approaches that aim to harness multivariate interactions have been proposed. In this work, we show how multiresolution transformations can be adapted and applied to genomic data to extract cues of multivariate interactions, and effectively improve automatic recognition performance of clinical outcomes.

Multiresolution transformations are signal processing tools that expand the original signal representation into a collection of multiple subspace signals with specific properties that can be optimized for the application at hand. In the last two decades, multiresolution methods have been applied to denoising, compression and pattern recognition. To our knowledge, in genomics these methods have not been used for automatic classification of clinical outcomes. The work presented here is the first attempt to design genome-wide multiresolution methods with the purpose of clinical classification.

In this work, we exploit the fact that multiresolution coefficients represent functions of local interactions of genomic data, specifically of neighboring single nucleotide polymorphisms (SNPs), and that multiresolution packet transforms can be adapted to include signal spaces that are better for recognition. This adaptive property has been used similarly in biometrics, as for example in fingerprint, face, and iris recognition.

We show that standard classification algorithms, such as logistic regression and support vector machines, can be used on top of multiresolution methods to improve clinical outcome predictions over what would be obtained as if using only the original SNP data. For example, we have found that for an Alzheimer's data set, standard algorithms obtain an ROC area (AUC) between 0.72 and 0.73, while with multiresolution methods presented here the AUC improves to 0.76. As this is ongoing work, we will discuss insights we have found in the design of such multiresolution methods for genome-wide data.

EXON-LENGTH DISTRIBUTION DYNAMICS IN GENOME EVOLUTION

Brian S Hilbush[1], Jayson T Durham[2]

[1]Real Time Genomics, Inc., Genomics, San Francisco, CA, 94105, [2]AgeIO, LLC, Computer Science, Lakeside, CA, 92040

Models of evolutionary mechanisms that are proposed to underlie genome evolution must account for a broad spectrum evolutionary influences that include strictly random processes (e.g. de novo mutations, gene duplication and loss), selective pressure, and fundamental molecular constraints. This range of interdependencies impacts the final outcome of the evolution of natural organisms. Next-generation sequencing (NGS) technologies will continue to expand the number of fully sequenced genomes that are available for analysis and processing. This unprecedented growing resource of genomics data provides an opportunity to analyze the evolutionary trends of statistical properties of gene features. For example, the evolutionary changes in exon size distributions can be characterized by tracking changes in the fundamental shape characteristics of the respective exon distributions across clades or even within genomes of a particular species.

The distributional properties of exons are of keen interest for a number of reasons. For example, the formation of new exons can be explained by small or large-scale genomic duplication events or via a process requiring the splitting of an existing exon. Thus, the analysis of whole genome sequence data affords the opportunity to gain insights into gene structural evolution across the phylogenetic spectrum. Recently, results reported by Ryabov and Gribskov (2008) suggested that a trend toward shorter exon lengths is the result of a Kolmogoroff fractioning process. In this previous study, the phylogenetic evolution of the distribution of exon lengths was described as a non-Gaussian distribution that was fitted using a mixture of two Gaussian distributions.

Here, we describe an approach that uses a mixture of two four-parameter nongaussian distributions (e.g. Johnson Family) to improve exon-length distribution fitting capabilities. The results reported herein support earlier results but further suggest that for better tracking distributional shape dynamics, the additional degrees of freedom from using two Johnson distributions may help to better characterize the observed distributional changes. With the recent availability of additional genomes, a larger number of genomes are analyzed with demonstration of improved distribution fitting capabilities.

APPLYING A GENE REGULATORY MODEL TO INVESTIGATE THE EFFECT OF COPY NUMBER VARIATIONS ON GENE EXPRESSION VALUES

Fang-Han Hsu[1], Erchin Serpedin[1], Yidong Chen[3], Edward R Dougherty[1,2]

[1]Texas A&M University, Department of Electrical and Computer Engineering, College Station, TX, 77843, [2]Translational Genomics Research Institute, Computational Biology Division, Phoenix, AZ, 85004, [3]The University of Texas Health Science Center at San Antonio, Greehey Children's Cancer Research Institute, San Antonio, TX, 78229

DNA abnormalities in the form of copy number variations (CNVs) are major causes of genetic diseases including cancer. Over the past few years, array Comparative Genomic Hybridization (aCGH), a technique that provides consistent results for detecting aberration regions, has been widely applied for genome wide analysis of CNVs.

Experiments have been done for evaluating the correlations between CNVs and gene expression values. A breast cancer study by (Pollack, *et al*. 2002) revealed that 62% of highly amplified genes have moderately or highly elevated expression. They found that at least 12% of all the variation in gene expression among the breast cancer tumors is directly attributable to CNVs. It is sufficient to say that CNVs do affect gene expression values. However, how CNVs affect gene expression values and what transcriptional mechanism makes the differences are still quite unclear.

To get deeper into the transcriptional mechanism, we realized that a mathematical model is needed. Based on Endo-16, a well defined cis-regulatory system of sea urchin studied by (Yuh, *et al*. 1998), we applied queuing theory to model the random process that transcription factors bind onto DNA strings. Sea urchins are close kin to humans in terms of genomics, thus Endo-16 is a good blueprint for us to start.

We successfully built a model and mathematically evaluated the effect of CNVs upon gene expression profiles. The relationships between gene expression and DNA copy number would be linear only when all TFs are plentiful. Otherwise, they are nonlinear and not easily detectable. Moreover, by mutual information we revealed that a TF with minimum activation probability can have most effect on gene transcription. We also show through simulation that the Expectation-Maximization algorithm can estimate TF arrival rate/degradation rate ratios (α) when the relationship between gene expression values and aCGH data are non-linear.

Properties of CNVs unknown for years now become much more clear, and further applications such as classification, estimation, and detection algorithms can be expected.

FROM BIOLOGICAL PATHWAYS TO REGULATORY NETWORKS

Ritwik K Layek[1], Aniruddha Datta[1], Edward R Dougherty[1,2]

[1]Texas A&M University, Electrical and Computer Engineering, College Station, TX, 77843-3128, [2]Translational Genomics Research Institute, Computational Biology Division , Phoenix, AZ, 85004

This work presents a general theoretical framework for generating Boolean networks whose state transitions realize a set of given biological pathways or minor variations thereof. This ill-posed inverse problem, which is of crucial importance across practically all areas of biology, is solved by using Karnaugh maps which are classical tools for digital system design. It is shown that the incorporation of prior knowledge, presented in the form of biological pathways, can bring about a dramatic reduction in the cardinality of the network search space. Constraining the connectivity of the network, the number and relative importance of the attractors, and concordance with observed time-course data are additional factors that can be used to further reduce the cardinality of the search space. The networks produced by the approaches developed here should facilitate the understanding of multivariate biological phenomena and the subsequent design of intervention approaches that are more likely to be successful in practice. As an example, the results of this research are applied to the widely studied ATM-p53-Mdm2-Wip1 pathway and it is shown that the resulting network exhibits dynamic behavior consistent with experimental observations from the published literature.

CANCER THERAPY DESIGN BASED ON PATHWAY LOGIC

Ritwik K Layek[1], Aniruddha Datta[1], Edward R Dougherty[1,2]

[1]Texas A&M University, Electrical and Computer Engineering, College Station, TX, 77843-3128, [2]Translational Genomics Research Institute, Computational Biology Division, Phoenix, AZ, 85004

Cancer is an umbrella term for a large number of diseases that are associated with loss of cell-cycle control, leading to uncontrolled cell proliferation and/or reduced apoptosis. This loss of cell-cycle control is usually caused by malfunction(s) in the cellular signaling pathways. These malfunctions can occur in many different ways and at many different locations in a particular pathway. As a result, a proper design of cancer therapy should first attempt to identify the location and type of malfunction in the pathway and then arrive at a drug or drug combination that is particularly well suited for it. Unfortunately, the current approach to cancer therapy does not follow such a systematic procedure. Consequently, the only kinds of cancers for which a high rate of therapeutic success has been achieved, namely chronic myelogenous leukemia (CML) and acute promyelocytic leukemia (APML), are the ones for which the pathway malfunctioning usually occurs at only one location in the pathways and, that too, in a very predictable fashion. Thus, for the vast majority of cancers, there is a critical need for precisely identifying (as best as we can) the failure point(s) in the pathway, hopefully leading to a more targeted therapy with a better likelihood of success.

In this work, we focus on the growth factor (GF) initiated signal transduction pathways which are widely studied pathways in the context of cancer. The interaction between different components of these pathways, as currently understood by biologists, are first modeled using Boolean logic gates (Boolean Network). Thereafter, all the possible single malfunctions in the resulting Boolean Network are enumerated and the response of the different malfunctioning circuits to a specific 'test' input are used to group together the malfunctions into a number of classes. Here malfunctions producing the same response to the test input are grouped into the same class. The effect of different therapeutic drugs, which target different parts of the Boolean circuit, are taken into account in deciding which particular drug or set of drugs should be used, based on the response of the malfunctioning pathway to the test input. In this way, we are able to map each possible malfunction to its appropriate drug or set of drugs. If the theoretical results in this work for the Growth Factor initiated signal transduction pathways withstand the test of experimental verification, it is very likely that the approach developed here would initiate a paradigm shift in designing targeted therapies for cancer.

INFERENCE OF GENE PREDICTOR SET USING BOOLEAN SATISFIABILITY

Pey-Chang K Lin, Sunil P Khatri

Texas A&M University, Electrical & Computer Engineering, College Station, TX, 77840

The inference of gene predictors in the gene regulatory network (GRN) has become an important research area in the genomics and medical disciplines. Accurate predicators are necessary for constructing the GRN model and to enable targeted biological experiments that attempt to validate or control the regulation process. In this paper, we implement a SAT- based algorithm to determine the gene predictor set from steady state gene expression data (attractor states). Using the attractor states as input, the states are ordered into attractor cycles. For each attractor cycle ordering, all possible predictors are enumerated and a conjunctive normal form (CNF) expression is generated which encodes these predictors and their biological constraints. Each CNF is solved using a SAT solver to find candidate predictor sets. Statistical analysis of the resulting predictor sets selects the most likely predictor set of the GRN, corresponding to the attractor data. We demonstrate our algorithm on attractor state data from a melanoma study and present our predictor set results.

FASTCALLER, A NEW BASE CALLER FOR DNA RE-SEQUENCING 1 BUNGTOWN ROAD

Fabian Menges[1], Bud Mishra[1,2]

[1]New York University, Courant Institute, Bioinformatics Group, New York, NY, 10003, [2]Cold Spring Harbor Laboratory, Quantitative Biology Center, Cold Spring Habor, NY, 11724

Short read sequencing technologies, such as 454 Life Sciences, SOLiD and Illumina sequencing, have become more common in the last few years. While providing high throughput rates at low cost in comparison to traditional sequencing technologies, short reads increase the complexity of DNA re-sequencing and de novo DNA sequence-assembly because of their small length (ranging from 30 to 500 base pairs). High coverage is required to obtain usable sequence reads and improved SNP detection.

FastCaller is a new base caller for DNA re-sequencing which is capable of significantly increasing the read quality in comparison to all other available base callers. FastCaller combines the knowledge from raw sequencing data (e.g. Illumina intensity data) with information from a reference genome utilizing a branch and bound algorithm to call a base.

For this purpose we developed a base by base alignment algorithm, built upon the Burrows Wheeler transform, which serves as a feedback-control for a simple base caller relying just on raw sequencing data. Multiple high quality reads are concurrently extended during the base calling process using a branch and bound algorithm for recovering a single read described by the raw sequencing data.

In addition to a software implementation, an FPGA implementation of FastCaller is presented, which potentially allows a speedup by a factor of eight in comparison to the software implementation running on state of the art CPUs.

While FastCaller was developed for Illumina intensity data, its framework is general enough to be applied to all other short read technologies. In our work we compare FastCaller to several well-known Illumina base callers, such as Bustard (Illumina), Ibis (Max Planck) and BayesCall (UC Berkeley).
Since FastCaller implicitly performs alignments and therefore depends on the structure of a reference genome its performance is evaluated for several genomes from different organisms.
We show that FastCaller reduces base calling error rates without sacrificing its accuracy in SNP detection, when compared against other existing base callers.

SUTTA: SCORING-AND-UNFOLDING TRIMMED TREE ASSEMBLER

Giuseppe Narzisi

New York University, Computer Science, New York, NY, 10003

We have developed a novel branch-and-bound based algorithmic framework for whole genome sequence assembly. In the most general setting, this problem addresses reconstruction of a DNA sequence from a collection of randomly sampled fragments (similarly to a large jigsaw puzzle) complicated by the presence of haplotypic ambiguities, sequencing errors and repetitive sections. The most successful approaches so far in the literature have been to recast the problem in graph-theoretic terms as one of finding a collection of paths in the graph satisfying certain specific properties. Contingent upon how the overlap relation is represented in these graphs, two dominant assembly paradigms had dominated so far: *overlap-layout-consensus* and *sequencing-by-hybridization*. SUTTA is a new sequence assembly algorithm that, in contrast to graph based assemblers, provides a flexible framework in which each contig is assembled independently and dynamically one after another using the *branch-and-bound* strategy.

At a high level, SUTTA views the assembly problem simply as a constrained optimization. It relies on a rather simple and easily verifiable definition of feasible solutions as consistent layouts. It generates potentially all possible consistent layouts, organizing them as paths in a double-tree structure, rooted at a randomly selected seed read (see methods). A path is progressively evaluated in terms of an optimality criteria, encoded by a set of score functions based on the set of overlaps along the layout. In contrast to previous assemblers where mate-pair constraints and validation were performed in a post-processing step, SUTTA concurrently verifies the validity of the layouts (with respect to various long-range information) through well-chosen penalty functions that combines different structural properties (e.g., transitivity, coverage, mated pairs, physical maps, etc). Facilitated by this new technology, we have assembled DNA sequence-read data both from old Sanger chemistry and next generations sequencing technologies (e.g., Illumina). In our experimental comparison we have found that SUTTA is competitive against the state-of-the-art assemblers in contig size and out-competes them in quality as shown by the Feature-Response curve analysis.

This new class of technology - aiming to enable *long-range haplotypic sequence assembly* - promises to open up the possibility of genomic approaches and algorithms that remain both agnostic to underlying chemistries as well as immune to the errors intrinsic to the underlying platforms. This is joint work with Bud Mishra of Courant Institute, NYU.

EFFICIENT DESIGNS FOR MULTIPLE GENE KNOCKDOWN EXPERIMENTS

Bobak Nazer, Robert D Nowak

University of Wisconsin - Madison, Electrical and Computer Engineering, Madison, WI, 53706

This paper develops theoretical bounds on the number of required experiments to infer which genes are active in a particular biological process. The standard approach is to perform many experiments, each with a single gene suppressed or knocked down. However, certain effects are not revealed by single-gene knockouts and are only observed when two or more genes are suppressed simultaneously. Here, we propose a framework for identifying such interactions without resorting to an exhaustive pairwise search. We exploit the inherent sparsity of the problem that stems from the fact that very few gene pairs are likely to be active. We model the biological process by a multilinear function with unknown coefficients and develop a compressed sensing framework for inferring the coefficients. Our main result is that if at most S gene or gene pairs are active out of N total then approximately S^2 log N measurements suffice to identify the significant active components.

CONTROL OF STOCHASTIC MASTER EQUATION MODELS OF GENETIC REGULATORY NETWORKS BY APPROXIMATING THEIR AVERAGE BEHAVIOR

Ranadip Pal[1], Mehmet U Caglar[2]

[1]Texas Tech University, Electrical and Computer Engineering, Lubbock, TX, 79409, [2]Texas Tech University, Physics, Lubbock, TX, 79409

Stochastic master equation (SME) models can provide detailed representation of genetic regulatory system but their use is restricted by the large data requirements for parameter inference and inherent computational complexity involved in its simulation. In this paper, we approximate the expected value of the output distribution of the SME by the output of a deterministic Differential Equation (DE) model. The mapping provides a technique to simulate the average behavior of the system in a computationally inexpensive manner and enables us to use existing tools for DE models to control the system. The effectiveness of the mapping and the subsequent intervention policy design was evaluated through a biological example.

PICXAA-R: PROBABILISTIC STRUCTURAL ALIGNMENT OF MULTIPLE RNA SEQUENCES USING A GREEDY APPROACH

Sayed M Sahraeian, Byung-Jun Yoon

Texas A&M University, ECE Dept., College Station, TX, 77843

Increasing number of newly discovered non-coding RNAs with huge functional variety have revealed the substantial roles they play in living organisms. The function of these RNAs is largely ascribed to their folding structure, which is often better conserved than their primary sequence. Thus, it is important to consider this structural aspect in the comparative analysis of RNAs through structural alignment algorithms.

In this abstract, we introduce PicXAA-R (probabilistic maximum accuracy alignment of RNA sequences), a novel non-progressive approach that can efficiently find the structural alignment of multiple RNA sequences with maximum expected accuracy. PicXAA-R greedily builds up the structural alignment from sequence regions with high local similarities and high base pairing probability. Thus, it avoids the propagation of early stage alignment errors, typically observed in progressive alignment schemes.

To simultaneously grasp local similarities among sequences and take the advantage of their conserved structure, we incorporate three types of probabilistic consistency transformations. These transformations modify both the pairwise base alignment probabilities and the base pairing probabilities using the information from other sequences in the alignment.

For a fast and accurate construction of the alignment, we propose an efficient two-step graph-based alignment scheme. In the first step, we greedily insert the most probable alignments of base-pairs with high base pairing probability. Hence, we build up the skeleton of the alignment using the secondary structure information. Next, we successively insert the most probable pairwise base alignments into the multiple structural alignment, as in PicXAA [1], a multiple protein sequence alignment scheme that we have recently proposed. This step can effectively grasp the local sequences similarities. Finally, we use a discriminative refinement step to improve the overall alignment quality in sequence regions with low alignment probability.

Extensive experiments on several local alignment benchmarks clearly show that PicXAA-R is one of the fastest algorithms for structural alignment of multiple RNAs and it consistently yields accurate results. We compared PicXAA-R to several well-known structural RNA alignment schemes, including MXSCARNA and CentroidAlign. On average, PicXAA-R shows 5-6% improvement in terms of SP score and 7-8% improvement in terms of SCI sore over MXSCARNA. It also outperforms CentroidAlign by 1-2% in both scores.

[1] S. M. E. Sahraeian and B. J. Yoon, "PicXAA: Greedy Probabilistic Construction of Maximum Expected Accuracy Alignment ", Nucleic Acids Research, 2010.

IDENTIFICATION OF GENES INVOLVED IN OVARIAN CANCER PLATINUM SENSITIVITY THROUGH MULTI-MODAL COX REGRESSION

Vinay Varadan, Sitharthan Kamalakaran, Nilanjana Banerjee, Angel Janevski, Nevenka Dimitrova

Philips Research North America, Ultrasound, Photonics and Bioinformatics, Briarcliff Manor, NY, 10510

Ovarian cancer is the leading cause of death in gynecological cancers. Carboplatinum-based therapy is the standard choice for adjuvant treatment of ovarian cancer but only a subset of patients benefit from the therapy. The identification of ovarian cancer patients who would benefit from carboplatinum therapy remains a significant clinical challenge. Although previous studies have identified genes involved in ovarian cancer chemosensitivity, the mechanisms underlying chemosensitivity are far from fully known. The Cancer Genome Atlas has created a unique opportunity to identify genomic interactions and pathways that are clinically relevant by characterizing the same patient samples using multiple genomic modalities.

This work addresses the problem of identifying functional interactions amongst genes involved in ovarian cancer platinum sensitivity by the integrated analysis of whole-genome copy number variations and DNA methylation. Several techniques have been previously proposed for inferring gene-gene interactions from one set of genomic data based on pairwise mutual information, Bayesian networks and graphical Gaussian models. Our method is fundamentally different from these in that it identifies only those interactions that are related to the phenotype and any interactions representing general biological functions that are unrelated to the phenotype are ignored.

We first deduce the platinum free interval (PFI) for each patient included in the TCGA ovarian cancer dataset using the clinical data on adjuvant chemotherapy, time to progression or recurrence. We then use Cox regression analysis to identify DNA methylation loci that can significantly stratify patients based on PFI. Independently, we also determine copy number variations that significantly stratify patients based on PFI. For each significant DNA methylation locus, we then estimate the Cox coefficients of the product terms with all the copy number variation loci. Product terms with significant Cox regression coefficients suggest a statistically significant interaction of the two variables with respect to the platinum free interval. Significance levels of the Cox coefficients are estimated using permutation analysis and subsequent multiple testing correction.

We show that the resulting graph of interactions amongst DNA methylation loci and copy number changes specific to platinum free interval capture pathways that mediate chemosensitivity in ovarian cancer.

MAKING A COMPARATIVE ASSEMBLER A PSEUDO DE-NOVO ASSEMBLER USING MINIMUM DESCRIPTION LENGTH

Bilal Wajid[1,2], Erchin Serpedin[2]

[1]University of Enginerring and Technology, Electrical Department, Lahore, 54890, Pakistan, [2]Texas A & M University, Electrical and Computer Engineering, College Station, TX, 77843-3128

Rissanen's Minimum Description length (MDL) [1] has had a profound impact on all branches of technology where models are needed to be predicted and inferences made, including bioinformatics. MDL itself has evolved from the Two-part MDL to Sophisticated MDL and towards the Minimax Regret [2]. Genome assembly, broadly divided into comparative and de-novo assembly, itself is an inference problem where a set of reads are used to infer the genome. In inference problems, it only makes sense to use as much data as possible for making inferences. Using a reference sequence for inferring novel genome is no different. Therefore, comparative assemblers do use a reference sequence, which therefore, limits their potential use to either re-sequencing previously sequenced genomes or sequencing their mutants. However, De-novo assemblers have the ability to sequence genomes which were not previously sequenced making them very powerful methods. Using MDL this paper identifies the means whereby which every comparative assembler can also sequence those genomes that were not previously sequenced thereby enhancing their capabilities. In other words, comparative assemblers now have all the capabilities of a de-novo assembler and hence the name of the title "***Making a Comparative Assembler a Pseudo De-Novo Assembler Using MDL***". In order to sequence genomes that are not previously sequenced all we need is a reference sequence so that comparative assemblers too can be employed. Using MDL we search for the reference sequence that best describes the data, reads, and use it for comparative assembly.

Using simulated data our initial results shows that Two-part MDL can be effectively used for searching for a reference sequence that closely resembles the genome, to be sequenced. The results are promising enough to use this technique for real data and to measure the quality of the end product, the sequenced genome.

References

[1] Jorma Rissanen (2007), *Information and Complexity in Statistical Modeling*. Springer. NewYork.
[2] Teemu Ross. Information Theoretic modelling, Lecture 9: MDL Principle.
http://www.cs.helsinki./group/cosco/Teaching/Information/2009/lectures/lecture5a.pdf

A MATLAB IMPLEMENTATION OF GLM AND PLM FOR QTL ANALYSIS

Liya Wang[1], Matthew W Vaughn[1], Peter J Bradbury[2], Lincoln D Stein[1,3]

[1]Cold Spring Harbor Laboratory, iPlant Collaborative, Cold Spring Harbor, NY, 11724, [2]USDA-ARS-NAA , Robert W. Holley Center, Ithaca, NY, 14853, [3]Ontario Institute for Cancer Research, Bio-Computing, Toront, M5G0A3, Canada

The bottle-neck of testing the significance of millions of SNPs (Single Nucleotide Polymorphism) and especially their combinations is speed. One solution might be executing a highly optimized SNP testing 'kernel' on a computer cluster. Here we implement the optimized GLM (General Linear Model) and PLM (partitioned linear model) in Matlab for testing the statistical significance of SNPs in the NAM (Nested Association Mapping) line. In this dataset, the common parent, B73, was crossed to the other 25 founders, followed by selfing, to generate 25 segregating F2 populations. Out of each F2 population, 200 RILs were derived through single-seed descent with selfing to the F6 generation. There are around 1.6 million SNPs for each of the 26 founder lines but only 1106 markers for each of the 5000 RILs distributing across 10 chromosomes. Thus, the first implementation is the projection of SNP values from the combination of marker data and each of the 25 founder lines. Phenotype data are the residuals from a linear model containing joint linkage QTL for all but the chromosome targeted. The data used is the days to anthesis but they have been transformed into residuals. The first optimization of GLM is achieved by reducing duplicate linear algebra operations through introducing intermediate terms. The second optimization of GLM is by improvement in computing XTX when looping through SNPs. X is incidence matrix or so called design matrix. The next improvement, also the most significant improvement, is utilizing PLM (partitioned linear model). PLM achieves better computational efficiency over GLM by dividing the 'higher' level linear model to 'lower' level linear model plus one. In other words, PLM only computes the difference between two models. This is very efficient for forward model regression (gradually adding SNPs), stepwise model regression, as well as all strategies for constructing final models with combination of multiple SNPs. With GLM and PLM, routines, such as forward model regression and stepwise model regression, can be easily implemented. The second one differs from the first one by re-evaluating the entire model after pushing a new SNP into the model. These are two common strategies for building a model that can address the majority of the variance within the phenotype data.

DYNAMICS, STABILITY AND CONSISTENCY IN REPRESENTATION OF GENOMIC SEQUENCES

Liming Wang, Dan Schonfeld

University of Illinois at Chicago, Electrical and Computer Engineering, Chicago, IL, 60607

Processing of biological data sequences represented by mapping into numerical signals is a commonly used technique. The operators such as de-noising filter, smoothing filter and certain algorithm could be used iteratively. Little is unknown about the consistency of analysis results with different mapping strategies in this situation. Meanwhile, due to the errors and noises in acquisition of data, the stability of analysis results should never be neglected. In this paper, we provide a method for analyzing the consistency between different mappings under iterations of operator. We define different concepts of mapping equivalence. We show the necessary and sufficient condition for consistency under iteration of affine operator. We present a few theoretical results on the equivalent mappings on the concept of Fatou and Julia Set. We give the definition of stability under iteration of operator and show the stability issue can be viewed as a special case of mapping equivalence. We also establish the connection of stability to Fatou and Julia set. Finally, we present experimental results on human gene AD169 sequence and rhodopsin gene sequence under one of the widely used mappings and illustrate the equivalent mapping for a smoothing filter.

CONDITIONING-BASED MODEL FOR THE REGULATORY ACTIVITIES OF MICRORNAS IN SPECIFIC DIETARY CONTEXTS

Chen Zhao[1], Ivan Ivanov[2], Manasvi Shah[3], Laurie A Davidson[3], Robert S Chapkin[3], Edward R Dougherty[1,4]

[1]Texas A&M University, Dept. of Electrical and Computer Engineering, College Station, TX, 77843, [2]Texas A&M University, Dept. of Veterinary Physiology and Pharmacology, College Station, TX, 77843, [3]Texas A&M University, Program in Integrative Nutrition & Complex Diseases, College Station, TX, 77843, [4]Translational Genomics Research Institute, Computational Biology Division, Phoenix, AZ, 85004

For the first time, we studied the applicability of a conditioning-based model to a heterogeneous data set composed of expression values for microRNA, total mRNA and polysomal mRNA resulting from experiments about two dietary contexts. The results suggest that some of the microRNAs are likely to be involved in the regulation of a large set of genes and not just their putative targets. Furthermore, the regulatory activities of intestinal microRNA appear to be dependent on the sub-cellular location of mRNA within the cell.

RMS BOUNDS AND SAMPLE SIZE CONSIDERATIONS FOR ERROR ESTIMATION IN LINEAR DISCRIMINANT ANALYSIS

Amin Zollanvari[1], Ulisses M Braga-Neto[1], Edward R Dougherty[1,2,3]

[1]Texas A&M University, Electrical and Computer Engineering, College Station, TX, 77843, [2] Translational Genomics Research Institute, Computational Biology Division, Phoenix, AZ, 85004, [3]University of Texas M.D. Anderson Cancer Center, Department of Computational Biology and Bioinformatics, Houston, TX, 77030

The validity of a classifier depends on the precision of the error estimator used to estimate its true error. This paper considers the necessary sample size to achieve a given validity measure, namely RMS, for resubstitution and leave-one-out error estimators in the context of LDA. It provides bounds for the RMS between the true error and both the resubstitution and leave-one-out error estimators in terms of sample size and dimensionality. These bounds can be used to determine the minimum sample size in order to obtain a desired estimation accuracy, relative to RMS. To show how these results can be used in practice, a microarray classification problem is presented.

DE NOVO ASSEMBLY OF LARGE GENOMES USING CLOUD COMPUTING.

Michael C Schatz

Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY, 11724

The first step towards analyzing a previously unsequenced organism is to assemble the genome by merging together the sequencing reads into progressively longer contig sequences. New assemblers such as Velvet, Euler-USR, and SOAPdenovo attempt to reconstruct the genome by constructing, simplifying, and traversing the de Bruijn graph of the reads. These assemblers have successfully assembled small genomes from short reads, but have had limited success scaling to larger mammalian-sized genomes, mainly because they require memory and compute resources that are unobtainable for most users.

Addressing this limitation, we are developing a new assembly program Contrail (http://contrail-bio.sf.net), which uses the Hadoop/MapReduce distributed computing framework to enable de novo assembly of large genomes. MapReduce was developed by Google to simplify their large data processing needs by scaling computation across many computers, and the open-source version called Hadoop (http://hadoop.apache.org) is becoming a de facto standard for large data analysis, including in “cloud computing” environments where compute resources are rented on demand. For example, we have also successfully leveraged Hadoop and the Amazon Elastic Compute Cloud for Crossbow (http://bowtie-bio.sf.net/crossbow) to accelerate short read mapping and genotyping, allowing quick (< 4 hours), cheap (< $100), and accurate (> 99% accuracy) genotyping of an entire human genome from 38-fold short read coverage.

Similar to other leading short read assemblers, Contrail relies on the graph-theoretic framework of de Bruijn graphs. However, unlike these programs Contrail uses Hadoop to parallelize the assembly across many tens or hundreds of computers, effectively removing memory concerns and making assembly feasible for even the largest genomes. Our preliminary results show Contrail produces contigs of similar size and quality to those generated by other leading assemblers when applied to small (bacterial) genomes, but scales far better to large genomes. We are also developing extensions to Contrail to efficiently compute a traditional overlap-graph based assembly of large genomes within Hadoop, a strategy that will be especially valuable as read lengths increase to 100bp and beyond.

PRACTICAL NGS ANALYSIS ON THE CLOUD WITH GALAXY AMIS: UNCOVERING MITOCHONDRIAL VARIATION

Enis Afgan, Hiroki Goto, Ian Paul, Kateryna Makova, James Taylor, Anton Nekrutenko

galaxyproject.org, galaxyproject.org, University Park, PA, 16802

We have developed a solution that allows experimentalists to perform large-scale analysis using cloud-computing resources with nothing more than a web browser (http://usegalaxy.org/cloud). Using our solution, a user without computational expertise can instantiate an analysis environment on a cloud, and can add storage and compute resources to this environment as needed. Because the solution is built on the Galaxy framework, analyses using this solution are accessible, transparent, and reproducible. Popular tools and workflows for analyzing sequence data from various types of experiment are built-in and ready to run.
To demonstrate the utility of this analysis solution we have sequenced mitochondrial genomes from multiple related individuals from three independent families (a total of 48 Illumina datasets). Our goal was quantify and track heteroplasmy in mitochondrial DNA. Our analysis involved standard mapping steps as well as custom developed tools to identification of nucleotide substitutions and indels in mixtures. All analysis steps from data pre-processing to polymorphism calling were performed using a Galaxy instance instantiated on the cloud. Within Galaxy we use a variety of analysis tools to process this data and identified a number of somatic mutations and heteroplasmic sites. This is the first practical demonstration that cloud-computing resources can be made available to researchers with no computational infrastructure to successfully perform complex large-scale analyses. While performing the analyses we developed a series of Galaxy workflows that can used by anyone in the community to replicate our analyses exactly as they were performed initially. In addition we used Galaxy pages system to annotate and explain every step of each workflow as well as describe metadata associated with every of 48 illumina datasets used in this study.

HOW AND WHY DO RATES OF DIFFERENT MUTATION TYPES CO-VARY? MULTIVARIATE STATISTICAL ANALYSES IN THE CONTEXT OF LOCAL GENOMIC LANDSCAPE

Guruprasad Ananda, Anton Nekrutenko, Francesca Chiaromonte, Kateryna Makova

Penn State University, Center for Comparative Genomics and Bioinformatics, University Park, PA, 16802

The abundance of sequenced genomes has greatly accelerated studies of regional heterogeneity in rates of individual mutation types. However co-variation among rates of multiple mutation types remains largely unexplored, hindering a deeper understanding of mutagenesis and genome dynamics. Here, utilizing primate genomic alignments, we apply linear and non-linear versions of two multivariate analysis techniques (Principal Component and Canonical Correlation) to investigate the structure of rate co-variation in four mutation types, and simultaneously explore its associations with multiple genomic landscape features. We observe a consistent, largely linear co-variation among rates of nucleotide substitutions, small insertions, and small deletions (with some non-linearities present on chromosome X and near autosomal telomeres). This co-variation appears to be shaped by a common set of genomic features, both novel (nuclear lamina binding sites and methylated non-CpG sites) and studied previously (e.g., GC content and recombination rates). Strong non-linearities (especially near centromeric regions of large chromosomes) are evident among the genomic predictors of mutation rate co-variation. In contrast, microsatellite mutability does not co-vary with the rates of the other three mutation types investigated, and is elevated in unmethylated regions. Our results on the genomic determinants of mutation rate co-variation allow us to speculate about the role of different molecular mechanisms (e.g., replication, recombination, and repair), and of the local chromatin environment, in mutagenesis. The software tools developed for the present study are made available through Galaxy, an open-source genomics portal, which will greatly facilitate future applications of multivariate analysis techniques to other inquiries in genomics.

DETECTING RARE GENETIC VARIANTS IN THE LARGE-SCALE 1000 GENOMES EXOME RESEQUENCING PROJECT

Fuli Yu[1], Danny Challis[1], Jin Yu[1], Amit R Indap[2], Wen Fung Leong[2], Christopher L Hartl[3], Kiran V Garimella[3], Chris Tyler-Smith[4], Gabor T Marth[4], Richard A Gibbs[1], and the 1000 Genomes Project Exon Pilot Sequencing Subgroup

[1]Baylor College of Medicine, Human Genome Sequencing Center, Department of Human and Molecular Genetics, Houston, TX, 77030, [2]Boston College, Department of Biology, Chestnut Hill, MA, 02467, [3]Broad Institute of MIT and Harvard, Genome Sequencing and Analysis, Program in Medical and Population Genetics, Cambridge, MA, 02142, [4]Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, United Kingdom

The 1000 Genomes Exon Pilot Project generated high coverage sequence data primarily in the coding regions of approximately 1,000 genes from 697 individuals sampled from seven different populations. The data was collected using multiple DNA capture technologies combined with two different next generation sequencing platforms (Illumina and 454). Boston College (BC), Baylor College of Medicine-Human Genome Sequencing Center (BCM-HGSC) and the Broad Institute (BI) independently implemented informatics pipelines for data analysis that made SNP and indel discoveries with high confidence. The results have been released at the DCC website (www.1000genomes.org) for public download.

The median per-individual sequence coverage within the seven populations ranged from 30X to 67X. Analysis of all 697 samples by both BC and BI yielded ~13,000 total overlapping variant sites (SNPs). The Exon Pilot working group has carried out extensive experimental validations (>1200 sites) in five tiers using either Sequenom or PCR-Sanger pipelines, to understand the quality of the SNP call release. The validation experiment targeted various categories, including singleton-low frequency SNPs called by different groups, and functional SNPs, for example nonsynonymous SNPs, splice sites. Analysis based on comparison to the validation results indicates that the accuracy of the SNP calls is very high overall with validation rate at 95-98% for rare SNPs including singletons. The considerable sequence depth makes it possible to detect low-frequency variants with very high sensitivity, and therefore, ascertain the low-frequency end of the site frequency spectrum with much better accuracy than achievable in low-coverage sequence data. The ability to detect rare alleles in genomic regions of interest, and the modest cost compared to high-depth whole-genome sequencing, make capture-sequencing approaches attractive for medical re-sequencing studies. The 1000 Genomes Project aims to expand the exome sequencing program to the whole exome coverage in ~2500 individuals. This genetic variation resource enables detailed understanding of the variant frequency spectrum, particularly the percentage of the rare variants in the coding regions. These less common SNPs and their population-specific metrics (such as MAF, frequency spectrum and LD patterns) have not been well cataloged or characterized, which potentially alter gene functions and likely contribute to human disease risks.

MODEL-BASED SEQUENTIAL BASE CALLING FOR ILLUMINA SEQUENCING

Shreepriya Das, Haris Vikalo, Arjang Hassibi

The University of Texas at Austin, ECE, Austin, TX, 78712-0240

In this paper, we study the efficacy of a model-based base-calling approach for Illumina's sequencing platforms. In particular, we investigate Genome Analyzer I reads and provide a detailed biochemical model of the sequencing process, incorporating various non-idealities evident in such systems. Parameters of the model are estimated via a supervised learning based on the particle swarm optimization technique. A computationally efficient sequential decoding method is proposed for base-calling. It is demonstrated that the performance of the proposed approach is comparable to Illumina's base-calling method.

EFFECTS OF PARTIAL REPORTING OF CLASSIFICATION RESULTS

Mohammadmahdi Rezaei Yousefi[1], Jianping Hua[2], Chao Sima[2], Edward R Dougherty[1,2,3]

[1]Texas A&M University, Department of Electrical and Computer Engineering, College Station, TX, 77843, [2]Translational Genomics Research Institute, Computational Biology Division, Phoenix, AZ, 85004, [3]University of Texas M. D. Anderson Cancer Center, Department of Bioinformatics and Computational Biology, Houston, TX, 77030

It is common practice in the bioinformatics literature for modelers to propose a new classification scheme, perhaps in the form of a classification rule or a feature selection method, and report its performance on data sets of interest, such as gene-expression microarrays. These data sets often include thousands of features but a small number of sample points, which increases variability in feature selection and error estimation, resulting in highly imprecise reported performances. This suggests that the reported performance of the proposed scheme would be less correlated with and highly biased from the actual performance if only the best results are demonstrated. This paper confirms this by showing the behavior of the joint distributions of the minimum reported estimated errors and corresponding true errors as functions of the number of samples tested in a large simulation study using both modeled and real data.

SUBTYPE SPECIFIC BREAST CANCER EVENT PREDICTION

Herman Sontrop[1], Wim Verhaegh[1], Rene van den Ham[1], Marcel Reinders[2], Perry Moerland[3]

[1]Philips Research, Molecular Diagnostics, Eindhoven, 5656 AE, Netherlands, [2]Delft University of Technology, Delft Bioinformatics Lab, Delft, 2628 CD, Netherlands, [3]Academic Medical Center, Bioinformatics Laboratory, Amsterdam, 1100 AZ, Netherlands

We investigate the potential to enhance breast cancer event predictors by exploiting subtype information. We do this with a two-stage approach that first determines a sample's subtype using a recent module-driven approach, and secondly constructs a subtype-specific predictor to predict a metastasis event within five years. Our methodology is validated on a large compendium of microarray breast cancer datasets, including 43 replicate array pairs for assessing subtyping stability. Note that stratifying by subtype strongly reduces the training set sizes available to construct the individual predictors, which may decrease performance. Besides sample size, other factors like unequal class distributions and differences in the number of samples per subtype, easily obscure a fair comparison between subtype-specific predictors constructed on different subtypes, but also between subtype specific and subtype a-specific predictors. Therefore, we constructed a completely balanced experimental design, in which none of the above factors play a role and show that subtype-specific event predictors clearly outperform predictors that do not take subtype information into account.

INFERENCE OF GENE-REGULATORY NETWORKS USING MESSAGE-PASSING ALGORITHMS

MANOHAR SHAMAIAH, SANG HYUN LEE, HARIS VIKALO

U T AUSTIN, Electrical & Computer Engineering, AUSTIN, TX, 78712-0240

We present an application of message-passing techniques to gene regulatory network inference. The network inference is posed as a constrained linear regression problem, and solved by a distributed computationally efficient message-passing algorithm. Performance of the proposed algorithm is tested on gold standard data sets and evaluated using metrics provided by the DREAM2 challenge. In particular, inference of networks that were the focus of INSILICO1 and INSILICO2 challenges is considered. Performance of the proposed algorithm is comparable to that of the techniques which yielded the best results in the DREAM2 challenge competition

NETWORK PROPAGATION MODELS FOR GENE SELECTION

Wei Zhang[1], Baryun Hwang[1], Baolin Wu[2], Rui Kuang[1]

[1]University of Minnesota, Department of Computer Science and Engineering, Minneapolis, MN, 55455, [2]University of Minnesota, Division of Biostatistics, School of Public Health, Minneapolis, MN, 55455

In this paper, we explore several network propagation methods for gene selection from microarray gene expression datasets. The network propagation methods capture gene co-expression and differential expression with unified machine learning frame- works. Large scale experiments on five breast cancer datasets validated that the network propagation methods are capable of selecting genes that are more biologically interpretable and more consistent across multiple datasets, compared with the existing approaches.

HIERARCHICAL ANALYSIS OF REGULATORY NETWORKS AND CROSS-DISCIPLINARY COMPARISON WITH THE LINUX CALL GRAPH

Koon-Kiu Yan, Mark Gerstein

Yale University, Molecular Biophysics and Biochemistry, New Haven, CT, 06511

The study of hierarchical organization of complex networks provides a more intuitive picture on the regulatory interactions in various complex systems, including both biological and technological systems. In the first part of the talk, I will introduce the integrated regulatory network based on the systematic integration of various high-throughput datasets from the modENCODE project. The network consists of three major types of regulation: TF->gene, TF->miRNA and miRNA->gene. I will examine the topological structures of the network, with emphasis on its hierarchical organization. In the second part of the talk, I will further present the hierarchical organization of the E. coli transcriptional regulatory network and the call graph of the Linux operating system. The effects on the robustness of the systems and insights from evolution will be discussed.

DYNAMIC AND STATIC ANALYSIS OF TRANSCRIPTIONAL REGULATORY NETWORKS IN A HIERARCHICAL CONTEXT

Nitin Bhardwaj[1], Mark Gerstein[1,2]

[1]Program in Computational Biology and Bioinformatics, Department of Molecular Biophysics and Biochemistry, New Haven, CT, 06511, [2]Program in Computational Biology and Bioinformatics, Department of Computer Science, New Haven, CT, 06511

Gene regulatory networks have been shown to share some common aspects with commonplace social governance structures such as hierarchies. Thus, we can get some intuition into their organization by arranging them into well-known hierarchical layouts. Here we study a wide range of regulatory networks (transcriptional, modification and phosphorylation) in a hierarchical context for five evolutionarily diverse species. We specify three levels of regulators -- top, middle and bottom -- which collectively regulate the non-regulator targets lying in the lowest fourth level, and we define quantities for nodes, levels and entire networks that measure their degree of collaboration and autocratic or democratic character. Overall we show that in all the networks studied, the middle level has the highest collaborative propensity and that co-regulatory partnerships occur most frequently amongst mid-level regulators, an observation that has parallels in efficient corporate settings where middle managers need to interact most to ensure organizational effectiveness. Then to study dynamic effects, superimpose the phenotypic effects of tampering with nodes and edges directly onto the hierarchies. We reconstruct modified hierarchies reflecting changes in the level of regulators within the hierarchy upon deletions or insertions of nodes or edges. Overall, we find that rewiring events that affect upper levels have a more dramatic effect on cell proliferation rate and survival than do those involving lower levels. We also investigate other features connected to the importance of upper-level regulators: expression divergence, back-up copies and expression level.

Participant List

Mr. Majid Alsagabi
University of Minnesota
alsa0054@umn.edu

Dr. Mickey Atwal
Cold Spring Harbor Laboratory
atwal@cshl.edu

Mr. Ferhat Ay
University of Florida
ferhatay@ufl.edu

Dr. Nilanjana Banerjee
Philips Research North America
angel.janevski@philips.com

Dr. Andreas Beutler
Mayo Clinic
beutler.andreas@mayo.edu

Prof. Peter Beyerlein
University of Applied Sciences Wildau
peter.beyerlein@tfh-wildau.de

Dr. Nitin Bhardwaj
Yale University
nitin.bhardwaj@yale.edu

Dr. Nidhal Bouaynaya
University of Arkansas at Little Rock
nxbouaynaya@ualr.edu

Mr. mehmet caglar
Texas Tech University
umut.caglar@gmail.com

Dr. Andrea Califano
Columbia University
califano@c2b2.columbia.edu

Ms. Ting Chen
Texas A&M University
chenting@tamu.edu

Dr. Claudia Coronnello
University of Pittsburgh
clc196@pitt.edu

Ms. Lori Dalton
Texas A&M University
ldalton@tamu.edu

Mr. Shreepriya Das
The University of Texas at Austin
shree@mail.utexas.edu

Dr. Suprakash Datta
York University
datta@cse.yorku.ca

Dr. Nevenka Dimitrova
Philips Research
Nevenka.dimitrova@philips.com

Ms. Laurita dos Santos
National Institute for Space Research
lauritas9@gmail.com

Dr. Edward Dougherty
Texas A & M University
e-dougherty@tamu.edu

Prof. John Goutsias
The Johns Hopkins University
goutsias@jhu.edu

Dr. Pablo Hennings-Yeomans
University of Pittsburgh
pablo@pitt.edu

Dr. Brian Hilbush
Real Time Genomics
brian@realtimegenomics.com

Dr. Lucia Hindorff
NIH
hindorffl@mail.nih.gov

Ms. Mu-Fen Hsieh
Texas A&M University
mufen@cse.tamu.edu

Mr. Fang-Han Hsu
Texas A&M University
aha0413@tamu.edu

Dr. Yufei Huang
University of Texas at San Antonio
yhuang@utsa.edu

Dr. Ivan Ivanov
Texas A&M University
iivanov@cvm.tamu.edu

Mr. William Jenkinson
The Johns Hopkins University
jenkinson@jhu.edu

Dr. Philipp Kapranov
ST. LAURENT INSTITUTE
pkapranov@helicosbio.com

Dr. Fedor Karginov
CSHL
karginov@cshl.edu

Dr. Rui Kuang
University of Minnesota
kuang@cs.umn.edu

Mr. Ritwik Layek
Texas A&M University
ritwik@neo.tamu.edu

Mr. Pey-Chang Lin
Texas A&M University
k1arte@neo.tamu.edu

Dr. Jingyu Liu
mind research network
jliu@mrn.org

Dr. Kateryna Makova
Penn State University
kdm16@psu.edu

Dr. Vibha Mane
Stony Brook University
vibhamane1@gmail.com

Dr. William McCombie
Cold Spring Harbor Laboratory
mccombie@cshl.edu

Mr. Fabian Menges
New York University
Fabian.Menges@gmail.com

Dr. Perry Moerland
Amsterdam Medical Center

Mr. Giuseppe Narzisi
New York University
narzisi@nyu.edu

Dr. Bobak Nazer
University of Wisconsin - Madison
bobak@ece.wisc.edu

Dr. Anton Nekrutenko
Penn State University
anton@bx.psu.edu

Ms. Hani Neuvirth
IBM Research - Haifa
hani@il.ibm.com

Dr. Albert Oliveras
Ass. Professor/UPC-BARCELONATECH
albert@tsc.upc.edu

Dr. Toshio Ota
Kyowa Hakko Kogyo Co., Ltd.
toshio.ota@kyowa-kirin.co.jp

Prof. Ranadip Pal
Texas Tech University
ranadip.pal@ttu.edu

Dr. Xiaoning Qian
USF
xqian@cse.usf.edu

Mr. Mohammadmahdi Rezaei Yousefi
Texas A&M University
m.rezaei@neo.tamu.edu

Mr. Sayed Sahraeian
Texas A&M University
msahraeian@tamu.edu

Dr. Michael Schatz
Cold Spring Harbor Laboratory
mschatz@cshl.edu

Dr. Gary Schroth
Illumina, Inc.
gschroth@illlumina.com

Dr. Scott Schwartz
Texas A&M University
scott@stat.tamu.edu

Mr. Manohar Shamaiah
U T AUSTIN
manohar.shamaiah@gmail.com

Mr. Jai Singh
IRDE
jpsingh1972@yahoo.co.in

Mr. Ali Sobhi Afshar
Johns Hopkins University
afshar@jhu.edu

Mr. Georges St. Laurent III
ST. LAURENT INSTITUTE
georgest98@yahoo.com

Dr. Sing-Hoi Sze
Texas A&M University
shsze@cse.tamu.edu

Dr. Vinay Varadan
Philips Research North America
vinay.varadan@philips.com

Dr. Wim Verhaegh
Philips Research
wim.verhaegh@philips.com

Dr. Haris Vikalo
University of Texas

Mr. Bilal Wajid
Texas A and M University
bilalwajidabbas@neo.tamu.edu

Mr. Liming Wang
University of Illinois at Chicago
lwang37@uic.edu

Dr. Liya Wang
Cold Spring Harbor Labs
wangli@cshl.edu

Dr. Michael Wigler
Cold Spring Harbor Laboratory
wigler@cshl.edu

Mr. Baolin Wu
Yale University
baolin.wu@yale.edu

Dr. Koon-Kiu Yan
Yale University
koon-kiu.yan@yale.edu

Prof. Byung-Jun Yoon
Texas A&M University
bjyoon@ece.tamu.edu

Dr. Fuli Yu
Baylor College of Medicine
fyu@bcm.edu

Mr. Chen Zhao
Texas A&M University
feihuo2003@neo.tamu.edu

Mr. Amin Zollanvari
Texas A&M University
amin_zoll@neo.tamu.edu

VISITOR INFORMATION

EMERGENCY	CSHL	BANBURY
Fire	**(9) 742-3300**	**(9) 692-4747**
Ambulance	**(9) 742-3300**	**(9) 692-4747**
Poison	**(9) 542-2323**	**(9) 542-2323**
Police	**(9) 911**	**(9) 549-8800**
Safety-Security	**Extension 8870**	

Emergency Room **Huntington Hospital** 270 Park Avenue, Huntington	**631-351-2300** **(1037)**
Dentists Dr. William Berg Dr. Robert Zeman	 **631-271-2310** **631-271-8090**
Doctor MediCenter 234 W. Jericho Tpke., Huntington Station	**631-423-5400** (**1034**)
Drugs - 24 hours, 7 days Rite-Aid 391 W. Main Street, Huntington	**631-549-9400** (**1039**)

Free Speed Dial
Dial the four numbers (********) from any **tan house phone** to place a free call.

GENERAL INFORMATION

Books, Gifts, Snacks, Clothing, Newspapers
BOOKSTORE 367-8837 (hours posted on door)
Located in Grace Auditorium, lower level.

Photocopiers, Journals, Periodicals, Books, Newspapers
Photocopying – Main Library
Hours: 8:00 a.m. – 9:00 p.m. Mon-Fri
10:00 a.m. – 6:00 p.m. Saturday
Helpful tips - Obtain PIN from Meetings & Courses Office to enter Library after hours. See Library staff for photocopier code.

Computers, E-mail, Internet access
Grace Auditorium
Upper level: E-mail only
Lower level: Word processing and printing.
STMP server address: mail.optonline.net
To access your E-mail, you must know the name of your home server.

Dining, Bar
Blackford Hall
Breakfast 7:30–9:00, Lunch 11:30–1:30, Dinner 5:30–7:00
Bar 5:00 p.m. until late
Helpful tip - If there is a line at the upper dining area, try the lower dining room

Messages, Mail, Faxes

Message Board, Grace, lower level

Swimming, Tennis, Jogging, Hiking

June–Sept. Lifeguard on duty at the beach. 12:00 noon–6:00 p.m. Two tennis courts open daily.

Russell Fitness Center

Dolan Hall, east wing, lower level
PIN#: **Press 64565 (then enter #)**

Concierge

On duty daily at Meetings & Courses Office.
After hours – From tan house phones, dial x8870 for assistance

Pay Phones, House Phones

Grace, lower level; Cabin Complex; Blackford Hall; Dolan Hall, foyer

CSHL's Green Campus

Cold Spring Harbor Laboratory is pledged to operate in an environmentally responsible fashion wherever possible. In the past, we have removed underground oil tanks, remediated asbestos in historic buildings, and taken substantial measures to ensure the pristine quality of the waters of the harbor. Water used for irrigation comes from natural springs and wells on the property itself. Lawns, trees, and planting beds are managed organically whenever possible. And trees are planted to replace those felled for construction projects.

Two areas in which the Laboratory has focused recent efforts have been those of waste management and energy conservation. The Laboratory currently recycles most waste. Scrap metal, electronics, construction debris, batteries, fluorescent light bulbs, toner cartridges, and waste oil are all recycled. For general waste, the Laboratory uses a "single stream waste management" system, removing recyclable materials and sending the remaining combustible trash to a cogeneration plant where it is burned to provide electricity, an approach considered among the most energy efficient, while providing a high yield of recyclable materials.

Equal attention has been paid to energy conservation. Most lighting fixtures have been replaced with high efficiency fluorescent fixtures, and thousands of incandescent bulbs throughout campus have been replaced with compact fluorescents. The Laboratory has also embarked on a project that will replace all building management systems on campus, reducing heating and cooling costs by as much as twenty-five per cent.

Cold Spring Harbor Laboratory continues to explore new ways in which we can reduce our environmental footprint, including encouraging our visitors and employees to use reusable containers, conserve energy, and suggest areas in which the Laboratory's efforts can be improved. This book, for example, is printed on recycled paper.

1-800 Access Numbers

AT&T	**9-1-800-321-0288**
MCI	**9-1-800-674-7000**

Local Interest

Fish Hatchery	631-692-6768
Sagamore Hill	516-922-4447
Whaling Museum	631-367-3418
Heckscher Museum	631-351-3250
CSHL DNA Learning Center	x 5170

New York City

Helpful tip -

Take Syosset Taxi to Syosset Train Station ($8.00 per person, 15 minute ride), then catch Long Island Railroad to Penn Station (33rd Street & 7th Avenue). Train ride about one hour.

TRANSPORTATION

Limo, Taxi

Syosset Limousine	516-364-9681 (**1031**)
Super Shuttle	800-957-4533 (**1033**)
To head west of CSHL - Syosset train station	
Syosset Taxi	516-921-2141 (**1030**)
To head east of CSHL - Huntington Village	
Orange & White Taxi	631-271-3600 (**1032**)
Executive Limo	631-696-8000 (**1047**)

Trains

Long Island Rail Road	822-LIRR
Schedules available from the Meetings & Courses Office.	
Amtrak	800-872-7245
MetroNorth	800-638-7646
New Jersey Transit	201-762-5100

Ferries

Bridgeport / Port Jefferson	631-473-0286 **(1036)**
Orient Point/ New London	631-323-2525 **(1038)**

Car Rentals

Avis	631-271-9300
Enterprise	631-424-8300
Hertz	631-427-6106

Airlines

American	800-433-7300
America West	800-237-9292
British Airways	800-247-9297
Continental	800-525-0280
Delta	800-221-1212
Japan Airlines	800-525-3663
Jet Blue	800-538-2583
KLM	800-374-7747
Lufthansa	800-645-3880
Northwest	800-225-2525
United	800-241-6522
US Airways	800-428-4322