bcats2010

biomedical computation at stanford

11th annual symposium

abstracts

Welcome

to the eleventh annual symposium on
Biomedical Computation at Stanford (BCATS)

This student-run one-day symposium provides an interdisciplinary forum for students and post-docs to discuss their latest work in computational biology and medicine with their peers at Stanford and other local universities. Since its inception in 1999, BCATS has seen growth and change in the field of biomedical computation and has evolved in concert. This year's schedule features cutting-edge research from one of the most diverse pools of participants in its 11 year history.

We thank our keynote speakers, student presenters, judges, sponsors, and all 2010 attendees.

The BCATS 2010 organizing committee

Konrad Karczewski, Biomedical Informatics

Rob Tirrell, Biomedical Informatics

Keyan Salari, Medical Scientist Training Program

Matt DeMers, Bioengineering

Jessica Faruque, Electrical Engineering

Amir Ghazvinian, Biomedical Informatics

MAKING A MEASURABLE DIFFERENCE.

As the world's premier measurement company, Agilent Technologies works with engineers, scientists and researchers around the globe to meet the challenges of today and tomorrow.

From home entertainment to forensics, from food safety to network reliability and from wireless communications to discovering the genetic basis of disease, Agilent provides the measurement capabilities that make our world more productive, safer, healthier and more enjoyable.

Agilent is committed to being an economic, intellectual and social asset to each and every nation and community where we operate right down to the neighborhoods where we work and live.

It's all part of making a measurable difference. For everyone.

WEB-BASED NEXT-GENERATION SEQUENCE DATA ANALYSIS, VISUALIZATION AND STORAGE

DNANEXUS FOR THE RESEARCHER

100 whole human genomes in one day

Whether you are an academic researcher with a single lane of data or a life science enterprise research organization with 1,000 whole genomes, DNAnexus offers a complete solution for your next-gen sequence data analysis needs.

- *Upload your sequence data in seconds –no additional hardware required*
- *Choose from a set of fully automated sequence analyses*
- *Retrieve analysis results and summary statistics*
- *Visualize your data in a fully integrated and easy-to-navigate browser*
- *Export all of your data for downstream analyses in other tools*

DNANEXUS FOR THE NEXT-GEN SEQUENCING CENTER

Future-proof your sequencing operations

DNAnexus is a complete solution for next-gen sequencing providers, enabling instant expansion of your sequencing capacity and streamlining of your data management so your users get closer to their results, faster. By using cloud computing, DNAnexus can support academic sequencing cores and commercial sequencing providers of any size, from a single sequencing instrument to 100 instruments.

SEQUENCING PLATFORMS SUPPORTED

- *Illumina Genome Analyzer, GAII/x/e, HiSeq 2000™*
- *All Life Technologies ABI SOLiD™ instruments*
- *Complete Genomics whole-genome sequencing*

APPLICATIONS

- *Genome/transcriptome mapping*
- *RNA-seq: expression quantification, alternative splicing, 3'-end quantification and discovery*
- *ChIP-seq: Identification of TF binding sites and broad regional interactions (e.g. histone modifications)*
- *Tag-based enrichment: general discovery and quantification of enrichment*
- *HpaII/MspI Methyl-seq: enzyme digest site quantification*
- *Nucleotide-level variation analysis*
- *Cancer variation analysis: tumor/normal sample comparisons to subtract out germline variants*
- *Small insertion and deletion detection*

DNANEXUS PROVIDES

- On-demand compute and storage infrastructure
- Access through web browser
- Long term data storage
- Instant sharing and collaboration

Sign up for a FREE Trial account! **www.dnanexus.com** ***Completely free for up to 3 personal samples.***

DNAnexus

SNL-Livermore, California

Sandia National Laboratories' California site was established in 1956 to provide systems engineering support for its neighboring weapons design facility, Lawrence Livermore National Laboratory. Today Sandia-California employs about 900 people, who occupy about 60 buildings on 413 acres on the outskirts of Livermore, a center of viniculture, ranching, and light industry on the eastern border of the San Francisco Bay Area.

...employs about 900 people, who occupy about 60 buildings on 413 acres...

The professional staff includes more than 400 people with advanced degrees — split almost evenly between PhDs and Masters. Engineering disciplines dominate the technical staff — principally mechanical and electrical engineers. The remainder represents a variety of scientific disciplines, including chemistry, computer science, and physics.

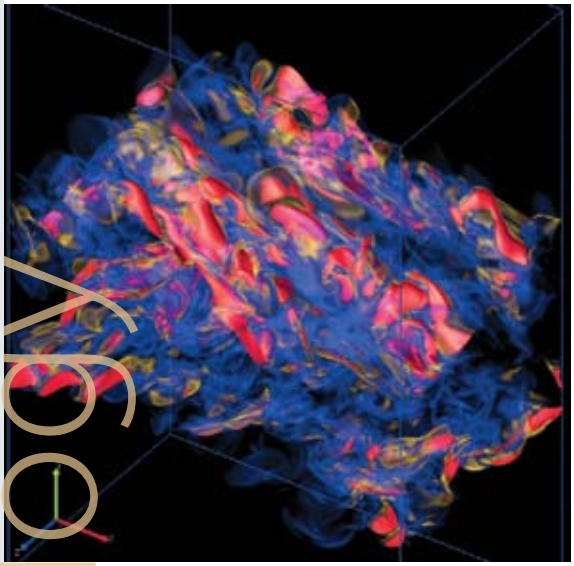
Innovative Technologies

Sandia-California has a particular focus on microelectronics, materials synthesis and processing, materials characterization, process simulation, engineering (theory and design), fabrication and demonstration technologies. Advanced computing, networking research, and modeling and simulation undergird many of our research programs.

Sandia has systems engineering responsibilities for about half of the weapons in the U.S. nuclear stockpile. Approximately 55 percent of the work at the California site concerns nuclear weapons. Equipping Cold War-era weapons to meet future deterrence needs poses greater engineering challenges than ever before. Expertise needed includes digital and analog designs, mechanical design and fabrication, data instrumentation, sensors, telemetry, embedded software, microsystems, systems engineering, and analysis.

Other researchers analyze threats to the nation's security. Sandians work on counterterrorism, nuclear material safeguards, and non-proliferation of weapons of mass destruction. Homeland defense against threats, including chemical or biological agents, has been an area of active research since the 1990s.

Systems analysis, chemical and biological detection capabilities, and basic research into cellular biology, molecular biology and biochemistry support that work.



The $37.9 million Distributed Information Systems Laboratory came on line at Sandia-California in 2005. It will enable development and deployment of distributed information systems technologies for the nuclear weapons complex.

At Sandia-California's Combustion Research Facility, two researchers recently performed the world's largest 3-D direct combustion simulation of a turbulent CO/hydrogen flame. The simulation will be used to understand turbulence-chemistry interactions to design more fuel-efficient, cleaner-burning engines.

innovative technology

additional SNL-CA

next-generation

The newest facility at Sandia-California is dedicated to biosystems research. At the center of the site, the Distributed Information Systems Laboratory has collaborative space for networking research partnerships with industry and academia, as well as separate areas for work to support nuclear weapons simulations or subsystems analysis.

Another facility gathers research applicable to small, integrated electromechanical systems — the Micro and Nano Technologies Laboratory. Here, next-generation chipmaking is being perfected with industry partners, while materials research supports thin-film and surface science studies, microfabrication advances, and energy storage solutions.

The site is also home to the unique and world-renowned Combustion Research Facility, which celebrated its 25th anniversary in 2005. The facility brings scientific tools to bear to increase fundamental understanding of combustion processes that provide the bulk of the world's energy for power and transportation uses.

Research Partnerships and University Relations

Sandia conducts collaborative research and development as a means of protecting national security interests and creating lasting value to the taxpayer. Our collaborations with industry and universities have resulted in new and enhanced technologies that have both commercial and national security benefits.

Sandia also offers programs for students ages 16 and up, through graduate school, research institutes for students to work on site, structured summer internships, and outreach to campuses. By supporting interest in science and technology careers, Sandia helps to reinforce advanced skills among the emerging U.S. workforce.

Operated by Lockheed Martin Corp. for the U.S. Department of Energy, Sandia maintains an environment in which every employee has the opportunity to achieve personal success. Employees have access to training courses and degree programs, as well as corporate programs to enhance work-life balance.

To find out more about our techncial activities at Sandia-California, visit

www.ca.sandia.gov

We're looking for people to help us **change the world.**

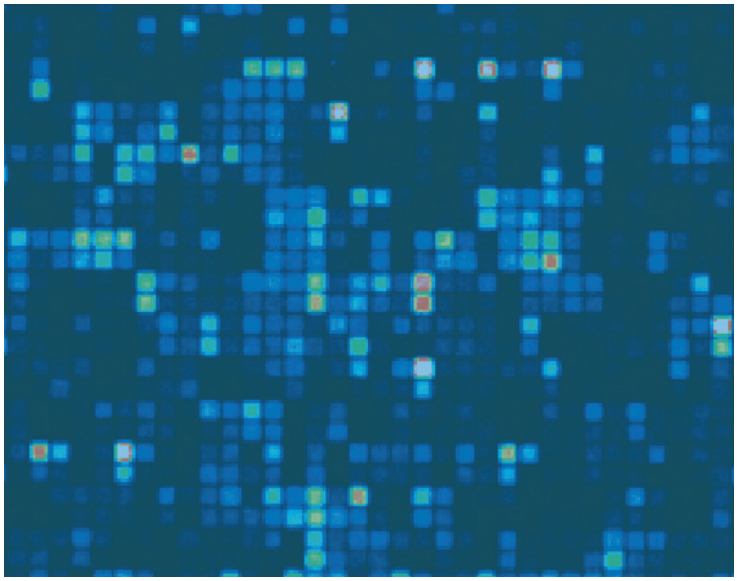
To get additional information and for current job opportunities, visit www.sandia.gov/employment

Sandia National Laboratories

LOCKHEED MARTIN

"Securing a peaceful and free world through technology."

5/2006



BECAUSE

Our CAUSE is Milosh and his cancer.

For more than 30 years, Genentech has been at the forefront of the biotechnology industry, using human genetic information to develop novel medicines for serious and life-threatening diseases. Today, Genentech is among the world's leading biotech companies, with multiple therapies on the market for cancer and other unmet medical needs.

Our founders believed that hiring talented, enthusiastic people would make Genentech a success. Today, we still believe our employees are our most important asset.

Genentech's research organization features world-renowned scientists who are some of the most prolific in their fields and in the industry. Our more than 1,100 scientists and postdocs have consistently published important papers in prestigious journals and have secured approximately 7,400 patents worldwide (with another 6,250 pending). Genentech's research organization combines the best of the academic and corporate worlds, allowing researchers not only to pursue important scientific questions, but also to watch an idea move from the laboratory into development and out into the clinic. We are proud of our long history of groundbreaking science leading to first-in-class therapies, and we hope you will consider us as we continue the tradition.

Genentech is a proud sponsor of the
2010 Biomedical Computation at Stanford (BCATS) Conference & Symposium.

To learn more about our current opportunities, please visit careers.gene.com. Genentech is an equal opportunity employer.

In January 2010, Genentech was named to FORTUNE's list of the "100 Best Companies to Work For" for the twelfth consecutive year.

Genentech
A Member of the Roche Group

NIH Center for Biomedical Computation

enabling groundbreaking research in physics-based simulations of biological structures

Interested in how biocomputation is changing biology and medicine?

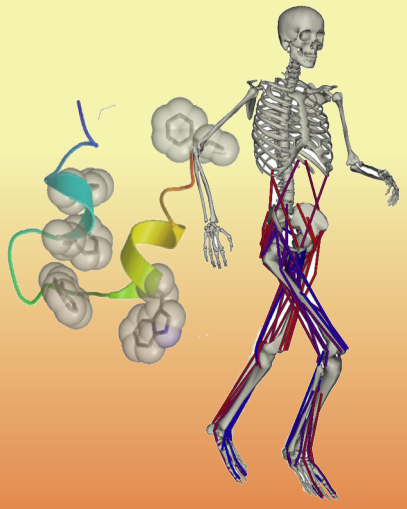
Sign up for a free subscription at:
www.BiomedicalComputationReview.org

Want to develop, share or find biosimulation software or data?

Explore the biosimulation repository and development environment at: www.simtk.org

Looking for high performance tools for solving equations?

Download the open-source SimTK Core libraries at: www.simtk.org/home/simtkcore



Interested in collaborating on important computational biological problems?

Visit us at: http://simbios.stanford.edu

Running image courtesy of Sam Hamner

NIH Roadmap Grant U54 GM072970

BIO-X

Bio-X is a Stanford University program supporting interdisciplinary work related to biology and medicine. The program is a joint effort by the Schools of Humanities and Sciences, Engineering, Earth Sciences, and Medicine.

The Bio-X Program encourages interdisciplinary work through various programs, including a seed grant program (the Interdisciplinary Initiatives Program), Bio-X Graduate Student Fellowships and Bio-X Stanford Interdisciplinary Graduate Fellowships in Human Health, Bio-X Undergraduate Research Awards, and Bio-X Travel Awards.

The Bio-X Program, which reaches across the university to nearly 450 faculty members in over 60 departments, is facilitated by the James H. Clark Center, which was completed in 2003 thanks to the enormous generosity of Jim Clark and Atlantic Philanthropies.

The Clark Center comprises the equipment, resources and utilities required to conduct breakthrough research at the cutting edge of engineering, science and medicine.

For more information on Bio-X,
visit our website at:

http://biox.stanford.edu

Stanford Biomedical Informatics Training Program

bmi.stanford.edu

Stanford: leading research and training in biomedical informatics since 1982

•Special expertise in key informatics research areas:

- •Translational Bioinformatics
- •Ontologies and the Semantic Web
- •Temporal Reasoning
- •Data Integration
- •Physics-based Simulation
- •Imaging Informatics

•Established research collaborations with over 40 labs across campus
•A unified Bioinformatics and Clinical Informatics program
•Engineering, Life Sciences and the Medical School are all located on one campus

Full-Time On-Campus Programs

- •PhD
- •PhD Minor for Stanford PhD students
- •Research-based Masters
- •Co-terminal Masters for Stanford undergraduates

Rigorous training programs including courses in informatics, computer science, probability and statistics, decision science, life sciences and ethics.

Part-Time Distance Learning Programs

- •AMIA 10 x 10 classes in Informatics offered in partnership with AMIA
- •Professional Masters Program (Honors Cooperative Program)
- •Certificates in Bioinformatics and Clinical Informatics
- •Individual Graduate Classes – the Non-Degree Option

For more information, contact:

Student Services Officer
Stanford Biomedical Informatics Training Program
Medical School Office Building, Room X-215
251 Campus Drive, Mail Code: 5479
Stanford, CA 94305-5479

Phone: (650) 723-1398
Fax: (650) 725-7944
email: bmi-contact@lists.stanford.edu
bmi.stanford.edu

Nov. 2010

Additional Sponsors

23andMe, Inc.

Invest in Yourself at:
http://www.23andme.com

Butte Lab

Post-doc positions available:
http://buttelab.stanford.edu

BCATS 2010 SCHEDULE

8.00	On-Site Registration & Breakfast	Clark Atrium
8.30	Opening Remarks	Clark Auditorium
8.45	Keynote Address: TBA	
9.30	**Christina Pop** (pg 21) *Analysis of the regulatory landscape affecting translational efficiency*	
9.45	**Aaron Diaz** (pg 22) *Signal combining chip-seq data to determine oncomir promoters*	
10.00	**Jonathan Karr** (pg 23) *Toward a Whole Cell Model of Mycoplasma genitalium*	
10.15	**Michael Kertesz** (pg 24) *Genome-wide Measurement of RNA Secondary Structure*	
10.30	Spotlight Presentations	
10.45	Poster Session I (odd-numbered posters)	Clark Atrium
11.45	**Veena Thomas** (pg 25) *Computer-Aided Drug Repurposing Against Malaria*	
12.00	**Marina Sirota** (pg 26) *Discovery and Validation of Novel Drug Interactions Using Compendia of Public Gene Expression data*	
12.15	**Nick Tatonetti** (pg 27) *Mining the Adverse Event Reporting System for Novel Drug-Drug Interaction*	
12.30	Lunch	Clark Atrium
1.15	Keynote Address: **Benjamin J. Fregly** (pg 18) *Simulation-Based Treatment Design for Knee Osteoarthritiis*	Clark Auditorium
2.15	**Vikash Gilja** (pg 28) *A robust high performance cortically-controlled motor prosthesis*	
2.30	**Paul Nuyujukian** (pg 29) *Generalization and robustness of a high performance cortically-controlled motor prosthetic*	
2.45	**Greg Bowman** (pg 30) *Atomistic models of protein-ligand interactions reveal a role for bth conformational selection and the induced fit mechanism*	
3.00	**Amirali Kia** (pg 31) *Computer simulation of the structural changes of the antimicrobial peptide cecropin near a bacterial cell inner membrane*	
3.15	Spotlight Presentations	
3.30	Poster Session II (even-numbered posters)	Clark Atrium
4.30	**Alex Morgan** (pg 32) *Medical risk interpretation of a whole genome*	Clark Auditorium
4.45	**Olivier Gevaert, Jiajing Xu** (pg 33) *Integrating medical images and transcriptomic data in non-small cell lung cancer*	
5.00	**Suchi Saria** (pg 34) *Discovering informative representations of clinical temporal data*	
5.15	Break	
5.30	Awards & Closing Remarks	
5.45	Reception	

BCATS 2010 Posters

	poster		page
SPOTLIGHT PRESENTATIONS	S1 Aaron M. Wenger	Functional Interpretation of GWAS and Epigenetic Markers using GREAT	36
	S2 Imran S. Haque	Bigger, Longer, and Uncut: Chemical Informatics at Scale	37
	S3 Yael Garten	Teaching computers to read the pharmacogenomics literature... so you don't have to.	38
	S4 Tiffany J. Chen	An Automated Workflow for Characterising Potential Chemotherapeutics	39
	S5 Joanna Lankester	Methods for Adjusting Dietary Recall Data	40
	S6 Debashis Sahoo	Computational Prediction of Human Cell Cycle Genes	41
	S7 Vladimir Jojic	A Model of Regulatory Program Differentiation in Immune Cell Development	42
	S8 Eric Van Nostrand	Highly Occupied Target (HOT) Regions in the C. elegans genome	43
	S9 Linda Liu	Transmission distortion in Crohn's disease risk gene ATG16L1 leads to sex difference in disease association	44
	S10 Elsa Birch	Host Genetic Interactions During Viral Infection: A Systems Virology Approach to λ Phage Infection of E. Coli	45
	1 Joel Dudley	Matching Cancer Genomes to Established Cell Lines for Personalized Oncology	47
	2 Tiffany Ting Liu	The Imaging Biomarker Ontology: ontology-based support for imaging biomarker research	48
	3 Pablo Cordero	Robustness to mutations as a key property of functional RNAs	49
	4 Jenelle Bray	Comparison of Cartesian and torsional normal modes for describing conformational changes in proteins	50
	5 Alex Pankov	Using Computational Algebraic Topology to Characterize Chromosomal Instability in Cancer	51
	6 Henry C. Hunter	Using confirmed dicistronic gene structures from several Drosophilid species to understand eukaryotic translation diseases	52
	7 Gurgen Tumanyan	Work in Progress: Identification of Protein Functional Sites using Machine Learning with Support Vector Machines	53
	8 Joseph W. Foley	UniPeak: Quantification of parallel sequencing experiments	54
	9 Emidio Capriotti	Defining the relationship between sequence and three-dimensional structure conservation in RNA	55
	10 Kai J. Kolhoff	An open-source library of data clustering algorithms for GPUs	56
	11 Jinesh Lalan	Toward the Use of cloud computing for bioinformatics: case study of Amazon Elastic Compute Cloud EC2	57
	12 Daniel Y. Li	The Beginnings of the Phenologue: An Ontology of the Genetic Causes of Autism Spectrum Disorders and Their Resulting Phenotypes and Endophenotypes	58
	13 Trevor Blackstone	Development of FEATURE and WebFEATURE 2.0 Software using Modern Software Engineering	59
	14 Eric Levy	Adaptive genetic variation in Quercus lobata	60
	15 Hyatt Moore IV	The Stanford EEG Viewer: A High-Throughput Platform for the Visualization and Analyis of Sleep Data	61
	16 Chuan-Sheng Foo	Deconvolving Individual Gene Expression Data into Cell-Type Specific Profiles	62
	17 Jessica Faruque	Developing a Similarity Reference Standard for CBIR Using Matrix Completion	63
	18 Diego F. Munoz	Simulation Modeling of the impac of prophylactic surgery and screening on life expectancy in BRCA1/2 mutation carriers	64
	19 Laleh Jalilian	Assessment of Diffusion Parameters at Pre-, Mid-, and Post-Radiation Therapy in Glioblastoma Patients Receiving Bevacizumab Therapy	65

BCATS 2010 Posters

poster			page
20	Joline Fan, Paul Nuyujukian	Bitrate optimization of a continuous control neural cursor applied to discrete selectioni tasks	66
21	Michael Zhou	IRAP: A Knowledge Based Discriminatory Function for Protein Structure	67
22	Robert Bruggner	Reconstruction of Signaling Transduction Networks from Mass Cytometry Data	68
23	Chiyere Nwabugwu	Quantitative Unmixing of Surface-Enhanced Raman Scattering Spectra	69
24	David R	Automatic video monitoring system to detect behavioral patterns in neurodegenerative disorders	70
25	Narmadan A. Kumarasamy	Differences in Antipsychotic Adverse Events among Adult, Pediatric, and Geriatric Populations - An Analysis of the FDA Adverse Events Reporting System	71
26	Kris Weber	Identifying Tumor-Specific Mutations in Acute Myelogenous Leukemia	72
27	Jia-Ren Lin	Reversed ORFeome: an evolutionary crossroad between recombination and gene regulation	73
28	Hyunggu Jung	A Model for Reasoning about Interaction with Users in Hospital Decision Scenarios	74
29	Roy Navon	Novel statistics reveal cancer universal microRNA activity	75
30	Guy Haskin Fernald	A Systems Biology Method for Predicting Drug-Gene Interactions in Saccharomyces cerevisiae	76
31	Hao Xiong	A flexible estimating equations approach for mapping function-valued traits	77
32	Tianyun Liu	Detecting ligand-binding sites similarity: A geometric-constraint free method using multiple FEATURE microenvironments (mFEATURE)	78
33	Jinjian Zhai	Cooling Design and Simulation of the Frontend for a 1mm^3 Resolution Breast Cancer Dedicated PET Camera	79

Keynote Speakers

BCATS Keynote Speaker

Benjamin J. Fregly, Ph.D.

University of Florida

Associate Professor, Department of Mechanical and Aerospace Engineering

B.J. Fregly is an Associate Professor in Mechanical and Aerospace Engineering at the University of Florida. He also holds a joint appointment in the Department of Biomedical Engineering, as well as a courtesy appointment in the Department of Orthopaedics and Rehablitation. He recieved a B.S. from Princeton University, followed by a M.S. and Ph.D in Mechanical Engineering from Stanford University. Afterward, he held a research position at the University of Lyon's Center for Mechanics. B.J. also gained industry experience in R&D and Software Engineering at Rasna Corporation and Parametric Technology Corporation. B.J. maintains a vision that future treatment decisions for musculoskeletal disease could be based on highly personalized computational models of patients instead of the population studies used today. His current research focuses on developing 3-D musculoskeletal modeling, simulation, and optimization techniques for studying clinical problems related to the human knee joint. This includes surgical simulation and predeiction of surgical outcomes, virtual prototyping of artificial knee designs, and surrogate modeling of joint contact mechanics for effficient calculation in large-scale dynamic simulations. B.J. also proves that musculoskeletal computation is clinically applicable through his work with simulation-based gait retraining and surgical planning aimed at slowing the porgression of osteoarthritis.

Talk Abstract:

Imagine a world where orthopedic surgeries and rehabilitation procedures are custom tailored to the patient, similar to how suits can be custom tailored to the business executive. In addition to using subjective clinical experience and simple anatomic measurements, clinicians use patient-specific computer models to design customized treatments. These models are created from pre-treatment movement and imaging data and are utilized within state-of-the-art simulation environments. The simulations allow clinicians to develop objective predictions of post-treatment function for various treatment designs under consideration. The end result is millions of patients whose quality of life is greatly improved through these technologies.

The Computational Biomechanics Lab at the University of Florida is seeking to make this futuristic scenario a reality. The lab's current research focus is on clinical problems related to knee osteoarthritis, with existing projects involving 1) computational simulation of knee osteoarthritis development, 2) computational design of a rehabilitation treatment for knee osteoarthritis, and 3) computational estimation of knee muscle and contact forces during walking. Each of these three areas will be covered in this keynote address. With further development, these computational technologies could also be applied to clinical problems related to stroke rehabilitation, cerebral palsy surgery, and knee ligament repair surgery.

bcats2010

Talk Abstracts

talk no.

1

Analysis of the regulatory landscape affecting translational efficiency

Cristina Pop
Stanford University

Nicholas Ingolia
Jonathan Weissman
Daphne Koller

Purpose

Gene expression is an essential cellular process that must be carefully controlled to ensure an organism's survival. Whereas transcriptional regulation has been studied intensively, factors contributing to translation of mRNA transcript into protein have proven more difficult to understand. In this work, we present a model that improves on measures of protein synthesis and describe a way to calculate efficiency of translation initiation based on various endogenous properties of a transcript.

Materials and Methods

The recent work of Ingolia et al. [1] utilizes deep sequencing of ribosome-protected fragments (footprints) to extract the density of translating ribosomes per codon along genes in Saccharomyces cerevisiae. Although the average footprint count per gene roughly represents translation rate, the counts are inflated/deflated at slow/fast positions in a transcript. Assuming a constant ribosome flow per transcript, we present a sophisticated computational model that accounts for these differential translocation rates.

Results

Our model elucidates two factors governing ribosomal pausing: the codon being translated and slow downstream positions leading to stacking. Correcting for these provides higher-quality estimates for protein synthesis rates, compared to previous experimental measures of protein abundance. Using our improved translational efficiency measures and the mRNA levels of budding yeast with a different ploidy, we can predict protein abundance in the different ploidy more accurately than using mRNA abundance.

We find that high translational initiation efficiency is correlated to: high folding energy of the window around the start codon (loose fold), presence of the Kozac site initiation motif, high folding energy of the 5' UTR, usage of preferential codons along the gene, and presence/absence of RNA binding proteins known to help/inhibit translation. Using linear regression on these features, we can accurately predict the translational initiation efficiency of a test set of genes not seen in model training, indicating we have captured a large degree of the regulatory landscape affecting translation. Finally, TAI correlates more strongly with protein abundance than translational efficiency, suggesting that higher ribosome usage creates evolutionary pressure for faster codons.

Conclusions

We have developed a computational model for correcting ribosomal profiling data for differential translocation rates. The codon being translated and ribosomal stacking are found to play determining roles in elongation efficiency whereas initiation efficiency is also affected by structural elements near the 5′ end of the transcript.

References

[1] N.T. Ingolia et al., (2009) Science 324:218-223

talk no.

2 Signal combining chip-seq data to determine oncomir promoters

Aaron Diaz
UCSF

Jun Song

Purpose:
Oncomirs are small non-coding RNAs which promote tumorgenesis by inhibiting the translation of genes associated with tumor suppression, apoptosis, and senescence. Determining the promoter regions of oncomirs can be addressed computationally by studying the co-localization of certain histone modification marks and transcription factors. Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) is a tool for mapping these loci. Computational methods are needed to address the synthesis of multiple ChIP-seq data sets. Moreover, significant computational challenges exist in modeling the biases intrinsic to high-throughput sequencing.

Methods:
We synthesize ChIP-seq data ensembles, down weighting noisy data sets and up weighting less noisy data sets, to form a consensus. Given ChIP alignment density signals s_1,...,s_k and associated input signals (control experiments) n_1,...,n_k we construct a consensus by computing the weighted sums s=w_1*s_1+...+w_k*s_k, n=w_1*n_1+...+w_k*s_k. The weights w_1,...,w_k are chosen, via an active set method, to maximize the signal-to-noise ratio. The consensus alignment densities can be used to annotate the promoters of oncomirs, providing a powerful method of discovering DNA elements that regulate microRNA biogenesis.

Results:
One usually combines multiple data sets by taking unions and intersections of peaks called individually from members of the ensemble. This method is not robust against noise, because peaks not strictly in the intersection can be lost and taking a union admits peaks called by only one member of the collection. Signal combining, before peak calling, allows each data set in the ensemble to contribute incrementally to the consensus.

We have found that p-values computed by model-based peak callers such as MACS scale with the total number of reads and are therefore unreliable. If one increases the total number of reads while keeping the relative distribution the same, then the p-values associated with putative peaks will significantly decrease. Consequently, the number of peaks called at a given p-value can dramatically increase.

Conclusion:
The underlying problem with currently available methods is the choice of null model, i.e. the baseline assumption of how reads are distributed. The widely used Poisson null model, used in MACS for example, scales the local mean with the total number reads. Our work in signal combining underscores the need to develop an empirical null model which takes into account biases in both sequencing and alignment.

talk no.

3

Toward a Whole Cell Model of Mycoplasma genitalium

Jonathan R Karr
Stanford University

Jayodita C. Sanghvi
Jared M. Jacobs
Markus W. Covert

A central challenge in biology is to understand how cellular life emerges from individual biochemical interactions. To address this challenge we have developed a novel computational framework which facilitates the integration of multiple disparate biochemical networks and data into a single unified model. Using this framework we have developed a detailed computational model of the complete life cycle of the smallest known freely-living organism, Mycoplasma genitalium. The model describes the life cycle of a single cell including DNA, RNA, and protein synthesis, metabolism, and cell division. The model accounts for the specific function of every annotated gene product, and simulates the dynamics of every molecular species. We have validated the model using several publically available experimental datasets. Currently we are using the model to gain insight into the control and regulation of cellular growth by exploring the effects of genomic and environmental perturbations on cellular behavior. In addition, we are developing an open-source, open-access web-based platform to facilitate broad application of our model.

URL: http://covertlab.stanford.edu/subpages/research.php#WholeCell

talk no.

4 Genome-wide Measurement of RNA Secondary Structure

Michael Kertesz
Stanford University

Yue Wan
Elad Mazor
John L. Rinn
Robert C. Nutter
Howard Y. Chang
Eran Segal

The structures of RNA molecules are often important for their function and regu-lation, yet there are no experimental techniques for genome-scale measurement of RNA structure. We describe a novel strategy termed Parallel Analysis of RNA Structure (PARS), which is based on deep sequencing fragments of RNAs that were treated with structure-specific enzymes, thus providing simultaneous in-vitro profiling of the secondary structure of thousands of RNA species at single nucleotide resolution.

We apply PARS to profile the secondary structure of the mRNAs of the budding yeast S. cerevisiae and obtain structural profiles for over 3000 distinct transcripts. Analysis of these profiles reveals several RNA structural properties of yeast transcripts, including the existence of more secondary structure over coding regions compared to untranslated regions, a three-nucleotide periodicity of secondary structure across coding regions, and an anti-correlation between the efficiency with which an mRNA is translated and the structure over its translation start site. PARS is readily applicable to other organisms and to profiling RNA structure in diverse conditions, thus enabling studies of the dynamics of secondary structure at a genomic scale.

talk no.

5 Computer-Aided Drug Repurposing Against Malaria

Veena Thomas
Stanford University

Edgar Deu
Paul Novick
Matthew Bogyo
Vijay Pande

Purpose: Drug repurposing -- identifying new indications for approved drugs -- has several advantages; such compounds have already been proved to be safe for human use, thus reducing the time and cost of research needed to move these compounds into new clinical uses. This is especially important for research on treatments for neglected diseases, where there is little incentive for traditional drug discovery research in the pharmaceutical industry. We have developed a new protocol for computer-aided drug repurposing, using both chemical similarity and structure-based approaches, to identify existing drugs that are inhibitors of DPAP1, a malaria drug target. In addition, we are using this protocol to predict targets for recently publicly-released pharmaceutical phenotypic data against malaria, from GSK, Novartis, and St. Jude.

Materials and Methods: We use 3D chemical similarity methods to compare 3D chemical structures of known inhibitors against both the WONTKILL database we have developed of compounds demonstrated safe for human use, and the Hopkins database of approved drugs. Hitlists from 3D chemical similarity screens are viewed in the context of protein structural information, and compounds are experimentally tested against both the malaria target DPAP1, and against the malaria parasite.

Results: In preliminary results, by combining chemical similarity approaches with structural information, and subsequent experimental followup, we have identified a clinically-used antimalarial with an unknown target of action, that inhibits DPAP1 with a Ki value of 230 uM. In other preliminary work, we have identified 7 approved drugs that kill the parasite in a single-point assay at 100 uM; we are following this up with dose-response curves. We are using these preliminary hits as further queries to identify more inhibitors of DPAP1. In addition, we used this protocol to analyze publicly-released phenotypic data on malaria from Novartis and GSK, and have discovered that some compounds which kill malaria with unknown targets share 3D chemical similarity with known DPAP1 inhibitors; we will experimentally test these predictions.

Conclusions: Computer-aided drug repurposing is a useful, timely, and cheaper strategy against neglected diseases. We have developed a new protocol for computer-aided drug repurposing via incorporating protein structural information and 3D chemical similarity techniques, to use known inhibitors of DPAP1 as queries to search for other inhibitors. 3D chemical similarity has the potential to find other inhibitors that share 3D chemical similarity, while being dissimilar in a 2D sense (“scaffold-hopping”). Our protein structure-based 3D chemical similarity method can be used to predict targets of action of phenotypic datasets.This method has broad applicability to other diseases with known inhibitors and protein structural information.

talk no.

6

Discovery and Validation of Novel Drug Indications Using Compendia of Public Gene Expression Data

Marina Sirota
Stanford University

Joel T. Dudley
Mohan Shenoy
Reetesh Pai
Annie P. Chiang
Alex A. Morgan
Alejandro Sweet-Cordero

Drug repositioning, the application of established drug compounds to novel therapeutic indications, offers several advantages over traditional drug development, including reduced development costs and shorter paths to approval. Recent approaches to drug repositioning employ high-throughput experimental approaches to assess a compound's potential therapeutic qualities. Here we present a systematic computational method to predict novel therapeutic indications based on comprehensive testing of molecular signatures in drug-disease pairs. We integrated gene expression measurements from 100 diseases and gene expression measurements on 164 drug compounds yielding predicted therapeutic potentials for these drugs. While we demonstrate the ability to recover many known drug and disease relationships using computationally-derived therapeutic potentials, we also predict many new indications for these drugs. We tested our top prediction for Crohn's disease (CD) using the rat TNBS model of inflammatory bowel disease (IBD), and successfully validated the predicted efficacy of the anti-convulsant topiramate in attenuating inflammation and macroscopic damage. Continued application of this approach to the experimental measurements of other diseases and drugs may yield many novel applications for other established drugs.

talk no.

7 Mining the Adverse Event Reporting System for Novel Drug-Drug Interactions

Nicholas P. Tatonetti
Stanford University

Guy Haskin Fernald
Russ B Almtan

Adverse drug events are a major cause of morbidity and mortality in the United States. Adverse events caused by drug-drug interactions are common and may result from shared pathway of mechanism or metabolism, however, we very few of these interactions are known. This is particularly concerning in our aging population where many patients are taking a myriad of drugs. We have developed an data mining technique to identify these interactions, validate these putative interactions in electronic medical record systems, and finally predict the underlying genetics of these interactions. In this study we focus on the discovery of a novel drug-drug interaction between a commonly prescribed anti-depressent and a commonly prescribed statin. We predicted that when used in combination these drugs would cause an increase in blood glucose and subsequently validated this hypothesis in three independent electronic medical record systems. We also validated our methods against a bronze standard data set of drug-drug interactions to demonstrate systematic performance.

talk no.

8 A robust high performance cortically-controlled motor prosthesis

Vikash Gilja
Stanford University

Paul Nuyujukian
Cindy A. Chestek
John P. Cunningham
Byron M. Yu
Joline Fan
Steven I. Ryu
Krishna V. Shenoy

Neural prostheses, or brain-machine interfaces (BMIs), translate neural signals to guide an actuator or computer cursor. Although current demonstrations provide a compelling proof-of-concept, limited performance impedes clinical viability. BMIs can be viewed from a feedback control perspective (see Cunningham et al, SFN 2010): the brain is the controller of a new plant defined by the BMI. This perspective leads us to two advances that significantly improve qualitative and quantitative performance.

We modify the Kalman filter commonly used in BMI literature (e.g. Kim et al, 2008) to verify these advances. In the first advance, during training, we fit neural data to a presumption of the desired volitional control signal, instead of observed or instructed kinematics. In the second advance, we develop the feedback Kalman filter, whose observation model incorporates cursor position as feedback.

Performance was tested in closed loop with two rhesus macaques. On each trial, the monkey acquired and held a 2D target in an allotted time period with a cursor, controlled by the native contralateral limb or BMI. Neural data were recorded from a 96-electrode array (Blackrock) spanning PMd and M1. All experiments used spike counts found by -4.5×RMS threshold detection without spike sorting (see Chestek et al, SFN 2010). Such a system has clinical appeal, particularly for arrays with potentially decreased SNR (monkeys L and J were 19-27 and 4-8 months post-implantation, respectively).

During online tests, the new BMI appears more controllable, producing straighter reaches and crisper stops. Relative to a standard Kalman filter, mean time to target (and failure rate) is reduced from 1323 ± 686ms (20%) to 993 ± 364ms (1%) in the same center out and back session. Further bin width and model training optimizations reduce this time to 600-800ms for both L and J, approaching arm reach times of 550-650ms. These innovations, supported by extensive experimental verification and design studies (see Nuyujukian et al, BCATS 2010), offer a significant performance advance, thereby increasing clinical viability.

talk no.

9 Generalization and robustness of a high performance cortically-controlled motor prosthetic

Paul Nuyujukian
Stanford University

Vikash Gilja
Cindy A. Chestek
John P. Cunningham
Joline M. Fan
Byron M. Yu
Stephen I. Ryu
Krishna V. Shenoy

Brain-machine interfaces (BMIs) translate neural signals into useful control signals. Using the algorithmic advances discussed in Gilja et al. (SFN 2010), we describe here our attempt to systematically test the performance and robustness of this BMI on two rhesus macaques, as well as to generalize the behavioral tasks beyond the center-out task. Monkeys were implanted with 96 channel electrode arrays, and experiments were carried out 19-27 (4-8) months post implantation for monkey L (monkey J). To control for inter-day variations in array or monkey performance, experiments were carried out intra-day in an A-B-A block design, thereby directly assessing the difference between various algorithms and modes of control. All decodes were performed with one 96 channel electrode array per animal (PMd in monkey L; M1 in monkey J), using neural threshold crossings set at -4.5×RMS without spike sorting at the beginning of each experimental day.

Using a standard Kalman filter without any of the new algorithmic advances, monkey L required two weeks of training under BMI control before 4 cm targets could be acquired and held for 500 ms on an 8 cm center-out task. However, using our algorithmic advances, training appeared considerably streamlined and high performance was achieved on the first day with monkey J.

Using this BMI on the same center-out task, monkey L had a mean target acquire time (and success rate) of 671 ms (96%) vs 543 ms (100%) for hand. Enforcing a maximum 1000 ms acquire time during BMI control improved mean target acquire time (618 ms) but decreased success rate (87%). This was 88% the performance of hand control and was repeatable with similar results over a span of six months.

On a more generalized pinball task, 4 cm targets appeared randomly in a 16x16 cm workspace with a 500 ms hold time. Using the BMI, monkey L achieved 69% the performance of hand control, the BMI at 722 ms (99%) vs hand at 496 ms (98%). This performance could be maintained for nearly two hours under BMI control, matching the duration of the longest hand-controlled sessions.

On a more complex, obstacle avoidance task, monkey J was trained to acquire 6 cm targets while avoiding an intervening visual barrier. Monkey J successfully navigated around the obstacle by instructing a curved trajectory and acquired targets 60% of the time with the BMI vs 62% of the time with hand control. BMI performance was 73% that of hand, the acquire time of BMI at 1219 ms vs hand at 889 ms.

We believe this new feedback-control design substantially increases the performance, robustness, and generalization of BMIs and may help bring BMIs closer to clinical viability.

talk no.

10

Atomistic models of protein-ligand interactions reveal a role for both conformational selection and the induced fit mechanism

Gregory R. Bowman
Stanford University

Daniel-Adriano Silva
Alejandro Sosa-Peinado
Xuhui Huang

Purpose: Many drugs work by altering the structure and dynamics of proteins. Unfortunately, it is difficult to dissect exactly how and this lack of understanding often thwarts drug design efforts. For example, it is common for drugs designed based on a single structure to interact in unexpected ways with other conformations that are sampled in solution and, as a result, such drugs typically fail to have the desired effect. Computer simulations have the potential to provide atomistic models of protein dynamics and protein-ligand interactions; however, traditional simulation methods fall short of biologically relevant timescales and suffer from limited statistics. We have developed methods that combine network theory and Bayesian statistics to enable statistically significant simulation of biologically relevant timescales. Using these methods, we are now able to better understand the roles of a protein's intrinsic dynamics (conformational selection) and the effects of ligands on a protein's structure and dynamics (induced fit) in protein-ligand interactions. As a proof of concept, we have investigated the mechanism by which the LAO protein binds its ligand, arginine.

Methods: We constructed network models, called Markov State Models (MSMs), for LAO binding. These networks are maps of a protein's free energy landscape, with nodes corresponding to metastable states (or free energy basins) and edges representing the probabilities of transitioning between pairs of these states. By combining many simulations into a single statistical model, these networks are capable of capturing millisecond timescale events (over 1,000 times slower than the previous state of the art!).

Results: Our model quantitatively predicts both the structure of the bound state and the binding kinetics of the LAO protein. While ligand binding is often modeled as a two-state process, LAO binding is actually a multi-step process with numerous parallel pathways. All of these pathways must pass through a single gatekeeper state, which we refer to as the encounter complex state because the protein is partially closed and only weakly interacting with its ligand. Since LAO completely encompasses its ligand in the bound state, it has been assumed that it must operate via an induced fit mechanism (i.e. the protein cannot first close and then encompass its ligand, so it must first bind to its ligand and then close). However, we have found that conformational selection is primarily responsible for transitions from the unbound state to the encounter complex state and then the ligand induces a transition to the bound state.

Conclusion: Network models built from atomistic simulations are capable of elucidating the mechanisms of protein-ligand interactions. For example, we have shown that both conformational selection and the induced fit mechanism may play important roles in LAO binding. We look forward to applying this methodology to other systems to understand general principles of protein-ligand interactions and improve drug design.

References:

1. Bowman et al. PNAS 2010;107:10890-10895.
2. Silva et al. PNAS 2010;submitted.

talk no.

11

Computer simulation of the structural changes of the antimicrobial peptide cecropin near a bacterial cell inner membrane

Amirali Kia

Stanford University

Eric Darve

Purpose

The main purpose of this research is to study the antimicrobial peptide (AMP) cecropin and the mechanisms by which it destroys bacterial cell membranes. The size, sequence and structure of the AMP play a significant role in its activity. In this research, we studied cecropin and its structural changes in bulk water and in the vicinity of bacterial-like membranes using Molecular Dynamic (MD) simulations. The changes in the secondary structure of cecropin while in contact with the bacterial cell membrane show which parts of the structure are the most important for its activity. Folding / unfolding free energy of the structure in both environments was computed using the Adaptive Biasing Force (ABF) method, developed in our group. These calculations show how the environment of the protein affects the kinetics of structural changes. Furthermore, a comprehensive study was performed to estimate the modeling and statistical errors in computing the free energy.

Methods

MD simulations were performed on the polypeptide cecropin with and without a bacterial-like membrane. To model the effect of the bacterial cell membrane, we used a sodium dodecyl sulfate (SDS) micelle. The fast relaxation times of small micelles allow a more accurate modeling of the structural changes within the limited available computation time. We computed the free energy of folding / unfolding using ABF and with over 1000 nanosec (1 microsec) of MD simulation time. ABF is an efficient and simple method to calculate free energy, which requires little prior knowledge of the energy landscape. However, it remains sensitive (as most methods) to the choice of reaction coordinate. Many different reaction coordinates were tried along with comprehensive statistical error analysis to make sure the best choice was adopted before performing the production runs.

Results and Discussion

MD simulations of cecropin revealed two different regions in the structure. The first region, consisting of the first 15 residues, tends to remain folded in a helical shape. On the other hand, the second region switches frequently between a helix and a random coil. This suggests that the first region plays a more important role in the cecropin's activity. Furthermore, the first region shows a clear change of structure from normal helix to pi-helix (not as tightly wound) when at the water / SDS interface. Free energy calculations reveal that, in the presence of SDS, there is a stable partially unfolded structure. This corresponds to a re-arrangement of SDS around the peptide, shielding hydrophobic residues from water. Such stabilizing mechanism does not exist in bulk water.

talk no.

12

Medical Risk Interpretation of a Whole Genome

Alexander A. Morgan

Stanford University

Rong Chen
Joel Dudley
Atul J. Butte

Purpose:
Patients are beginning to present to healthcare providers with the results of high throughput individualized genotyping, and interpreting these results in the context of the explosive growth of literature linking individual variants with disease may seem daunting. However, we suggest that results of a personal genomic analysis may be viewed as a panel of many tests for multiple diseases. By using well-established methods of evidence based medicine, these very many parallel tests may be combined using likelihood ratios to report a post-test probability of disease for use in patient assessment.

Materials and Methods:
Using a large database of reported gene-disease associations, we have created risk profiles across disease, modeling genetic tests of variants as independent tests, using the product of likelihood ratios to shift pre-test probabilities to post-test probabilities for 188 genomes. We have also developed visualization and interpretation tools to guide clinical care.

Results and Conclusions:
Our results show that we can generate clinically meaningful risk profiles from genotyping data that have informed clinical decision making and shifted patient care for one individual. We have also found that the risk profiles reported by extensive sequencing for 187 individuals differ dramatically from those that would be obtained from the variants measured or imputable by genotyping array.

Publications:
Ashley, E. A., A. J. Butte, et al. (2010). Clinical assessment incorporating a personal genome. Lancet 375(9725): 1525-1535.
Morgan, A. A., R. Chen, et al. (2010). Likelihood ratios for genome medicine. Genome Med 2(5): 30.

talk no.

13 Integrating Medical Images and Transcriptomic Data in Non-Small Cell Lung Cancer

Olivier Gevaert, Jiajing Xu
Stanford University

Ann Leung
Andy Quon
Daniel Rubin
Chuong Hoang
Sandy Napel
Sylvia Plevritis

Objective
To build an association map between medical images (CT/PET) and gene expression microarrays for Non-Small Cell Lung Cancer (NSCLC) from which to derive relationships between imaging features and gene expression.
Methods
We studied 25 cases of NSCLC using CT and PET images and microarray data from excised tumors. An experienced thoracic radiologist annotated the CT image using "semantic features" from a controlled vocabulary, a nuclear medicine physician extracted the Standard Uptake Value (SUV) from the PET scan, and we developed and applied algorithms to extract "computational features" that characterized the lesion's image texture using Gabor and other texture features, the sharpness of lesion boundaries and the lesion boundary shape, including notions of compactness, roughness, and other shape signatures. We preprocessed the microarray data using log transformation and quantile normalization, obtained 100 co-expressed gene clusters using k-means clustering, and computed a metagene for each cluster using its first principal component. We performed (a) univariate and (b) multivariate analyses to integrate imaging features and metagenes using (a) the rank-sum test, Kruskalwallis test or Spearman's rank correlation coefficient where appropriate, and (b) Sparse Canonical Correlation Analysis (SCCA), respectively.
Results
Image features included 44 semantic terms, 107 computational features, and SUV for each tumor. In a univariate analysis, 76 CT-features and SUV were significantly associated with at least one metagene, and on average 5.6 metagenes (uncorrected $p < 0.05$). Several of these associations were thought provoking. For example Metagene 79 was enriched for target genes of TNF and upregulated in the left apical upper lobe (LAUL) tumors but downregulated in the right apical upper lobe (RAUL) tumors. Using public domain data linking gene expression to survival, we found that Metagene 79 is positively correlated to survival (log-rank $p=0.017$), suggesting LAUL tumors have better prognostic significance than RAUL tumors, a finding corroborated by local surgical experience. In multivariate analysis, a set of 10 metagenes was significantly correlated with a group of 5 Gabor texture features and explains over 82% of the correlation between both data sources using SCCA analysis ($p=0.02$). These metagenes are enriched with genes associated with good survival outcome in NSCLC ($p<1.11$ e-16).
Conclusion
The integration of medical image features and gene expression promises to reveal molecular characteristics underlying medical image features. For translational purposes, this work highlights the potential use of medical image features as predictive markers for molecularly-targeted therapeutics.

talk no.

14 Discovering Informative Representations of Clinical Temporal Data

Suchi Saria
Stanford University

Daphne Koller
Anna Penn

Purpose: Physiological data are routinely recorded in intensive care, but their use for rapid assessment of illness severity has been limited. The data is high-dimensional, noisy, and changes rapidly; moreover, small changes that occur in a patient's physiology over long periods of time are difficult to detect, yet can lead to catastrophic outcomes. A physician's ability to recognize complex patterns across these measurements is limited. We propose a nonparametric Bayesian method for discovering informative representations in such continuous time series that aid both exploratory data analysis and feature construction.
Methods: Time Series Topic Model (TSTM) is a flexible generative modeling framework that can incorporate varying degrees of clinical supervision (completely unsupervised to partially supervised). It models each series as switching between latent "topics" (diseases), where each topic is characterized as a distribution over "words" (physiologic characteristics) that specify the series dynamics. TSTM discovers both the underlying topics and words given the data. Word and topic frequencies provide a new feature representation. Clustering in this new representation and analyzing learned model parameters can help discover clinical hypothesis supported by the data.
Results: TSTM, when applied to physiology data from premature infants in the neonatal ICU, obtains novel clinical insights. For instance, analysis of the learned model highlighted separability between healthy and unhealthy infants based on the word mixing proportions, suggesting different dynamics profiles for these two populations. Furthermore, words highly associated with healthy infants had higher entropy. Thus, we developed Physiscore*, a risk stratification score, that combines such signatures from physiological signals from the first 3 hours of life with birth weight and gestational age. Physiscore was validated on 138 infants with the leave-one-out method to prospectively identify infants at risk of short- and long-term morbidity. Based on noninvasive measurements alone, PhysiScore provided higher accuracy prediction of overall morbidity (86% sensitivity at 96% specificity) than other neonatal scoring systems, including the standard Apgar score (63% sensitivity at 70% specificity).
Conclusion: We demonstrate a flexible methodology of fitting generative models to discover highly informative representations in clinical temporal data. This approach can be extended to other types of patient data contained in Electronic Health Record databases for data-driven discovery of clinical hypothesis.

*Saria et. al., Integration of Early Physiological Responses Predicts Later Illness Severity in Preterm Infants, Science Trans. Med., Sept 2010

Poster Spotlight Abstracts

poster

S1 Functional Interpretation of GWAS and Epigenetic Markers using GREAT

Aaron M. Wenger
Stanford University

Cory Y. McLean
Michael Hiller
Shoa L. Clarke
Bruce T. Schaar
Gill Bejerano

Purpose
Recent technological advances in DNA sequencing provide an unprecedented view of the regulatory genome in action. It is now possible to sequence all binding events of transcription factors and complexes, examine the dynamics of chromatin marks, assay for open chromatin modifications between healthy and disease samples, perform genome wide association studies for different human traits and diseases, and more. Interpretation of this new data with tools designed for the microarray platform typically falls short. The microarray model of one probe per gene is a poor proxy for the rich regulatory landscape of the human genome. While the microarray model considers only proximal binding events, most binding events revealed by sequencing are distal to genes.

Materials and Methods
We have recently published the Genomic Regions Enrichment of Annotations Tool (GREAT)1, the first computational method that properly models whole genome cis-regulatory data. GREAT accurately incorporates multiple proximal and distal binding events by applying a statistical test that adjusts for biases in the distribution of genes throughout the genome.

Results
Our published work focused on successfully applying GREAT to reinterpret ChIP-Seq datasets of multiple transcription-associated factors in different developmental contexts. Here, we significantly extend the same methodology, demonstrating how to analyze high throughput datasets from Genome Wide Association studies (GWAS) and measurements of epigenetic markers. Specifically, we analyze histone-modified enhancers and GWAS markers for human traits and diseases including height and schizophrenia to reveal relevant known and novel pathways and biological processes.

Conclusion
We show that GREAT improves analysis of many types of genome-wide cis-regulatory datasets, including GWAS markers and regions with certain epigenetic modifications. For each of these datasets, GREAT identifies genomic regions putatively regulating components of particular pathways and cellular processes – some known, others novel. GREAT incorporates rich biological annotations from 20 ontologies and is available to the community as an intuitive web tool at http://great.stanford.edu/. Direct submission is also available from the UCSC Genome Browser.

References
1. McLean et al, GREAT improves functional interpretation of cis-regulatory regions. Nature Biotechnology, May 2010.

Webpage
http://great.stanford.edu/

poster

S2 Bigger, Longer, and Uncut: Chemical Informatics at Scale

Imran S. Haque
Stanford University

Vijay S. Pande

Purpose

Fifteen years ago, the advent of modern high-throughput sequencing revolutionized computational genetics with a flood of data. Today, high-throughput biochemical assays promise to make biochemistry the next data-rich domain for machine learning. However, existing computational methods, built for small analyses of about 1,000 molecules, do not scale to emerging multi-million molecule datasets. I will describe how new algorithms [1] and new hardware (GPUs) [2-3] allow us to cross a 10,000-fold scalability barrier to do large-scale biochemical machine learning.

Materials and Methods

Chemical machine learning (CML) usually works on the principle that "similar" molecules will have similar biochemical activities. Definitions of chemical similarity vary from 2D methods comparing chemical graphs (atoms and bonds) to 3D methods comparing the shapes and functionalities (anion/cation, etc.) of molecules in 3D space.

Similarity calculation is a critical bottleneck in CML. Our example is the calculation of all pairwise similarities on the PubChem3D database (N=17e6). On this dataset, existing similarity methods would take months on 200,000 CPUs and require 2 PB of storage.

Our first technique is to use GPUs (graphics processing units) to accelerate similarity computations [2,3]. Our second technique, named SCISSORS [1], uses linear algebra to efficiently approximate similarity metrics, and is applicable to domains beyond chemical informatics. SCISSORS allows asymptotic speedups, making O(N^2) problems scale effectively linearly in both space and time.

Results

Our GPU codes achieve 80-100x speedup relative to the existing best-in-class implementations. SCISSORS achieves >2,800x reductions in both space and time on PubChem. Combining GPUs and SCISSORS reduces the PubChem analysis from months on 200,000 CPUs to about a week on a single desktop computer and reduces storage requirements from 2PB to around 20 GB.

Conclusions

Computing chemical similarities is the critical bottleneck preventing the scaling of machine learning methods to emerging datasets. Our combination of GPU acceleration and new algorithms speeds up this critical path by > 100,000x, enabling a new class of large-scale machine learning to tackle emerging data-rich problems in chemical biology.

References

[1] Haque IS and Pande VS. "SCISSORS — A Linear-Algebraical Technique to Rapidly Approximate Chemical Similarities". J. Chem. Inf. Model. 50(6), 1075-1088 (2010).
[2] Haque IS, Pande VS, and Walters WP. "SIML: A Fast SIMD Algorithm for Calculating LINGO Chemical Similarities on CPUs and GPUs". J. Chem. Inf. Model. 50(4), 560-564 (2010).
[3] Haque IS and Pande VS. "PAPER — Accelerating Parallel Evaluations of ROCS". J. Comput. Chem. 31, 117-132 (2010)

poster

S3

Teaching computers to read the pharmacogenomics literature ... so you don't have to.

Yael Garten
Stanford University

Adrien Coulet
Russ B Altman

Purpose:
Pharmacogenomics is the study of how variation in the human genome impacts drug response in patients. It is a major driving force of personalized medicine in which drug choice and dosing decisions are informed by individual information such as DNA genotype. The field of pharmacogenomics is in an era of explosive growth; massive amounts of data are being collected and knowledge discovered, which promises to push forward the reality of individualized clinical care. However the knowledge being discovered is dispersed in many journals in the scientific literature, and findings are discussed in a variety of non-standardized ways. It is therefore challenging to identify important associations between drugs and genes with the result that these critical connections are not easily available to investigators or clinicians who wish to survey the state of knowledge for any particular gene, drug, disease or variant. Natural Language Processing and text mining techniques allow us to convert free-style text to a computable, searchable format in which pharmacogenomic concepts are identified and important links between these concepts are recorded and thus easily accessible.

Materials and Methods:
In this work, we use natural language processing techniques and a semi-automatically generated ontology to create a semantically rich model of pharmacogenomics, and then extract such knowledge from all of MEDLINE. Our ontology of PGx relationships was built starting from a lexicon of key pharmacogenomic entities and a syntactic parse of more than 87 million sentences from 17 million Medline abstracts.

Results:
The result is a network of over 41,000 relationships. Our extracted raw relationships have a 70 to 87.7% precision and involve not only key PGx entities such as genes, drugs, and phenotypes (e.g., VKORC1, warfarin, clotting disorder), but also critical entities that are frequently modified by these key entities (e.g., VKORC1 polymorphism, warfarin response, clotting disorder treatment).
Normalization of these relationships by mapping them to a common schema using the ontology allows creation of a network between over 200 types of entities with clearly defined semantics. This network is used to guide the curation of pharmacogenomic knowledge and provide a computable resource for knowledge discovery. For example, we use this network to infer latent knowledge implicit in the literature, to predict drug-drug interactions, and to identify conflicting facts described in the literature.

Conclusion:
These achievements provide us with new ways of interacting with the literature and the knowledge embedded within it, and help ensure that we do not bury the knowledge embodied in the publications, but rather connect the often fragmented and disconnected pieces of knowledge spread across millions of articles in hundreds of journals.
References:
Garten Y, Coulet A, Altman RB. Recent progress in automatically extracting information from the pharmacogenomic literature, Pharmacogenomics, October 2010.
Coulet A, Shah NH, Garten Y, Musen M, Altman RB. Using text to build semantic networks for pharmacogenomics, J Biomed Inform. 2010 Aug 17.

poster

S4 An Automated Workflow for Characterising Potential Chemotherapeutics

Tiffany J. Chen
Stanford University

Matthew R. Clutter
Nikesh Kotecha
Garry P. Nolan

Purpose:
Cancer therapeutics are grouped into general classes. Within classes it is difficult to determine differences without time-intensive studies and trials. While cancer drug classification is dependent on our knowledge of direct targets, it does not typically consider how a number of global cellular processes are ultimately affected. Quantifying the relationships between drugs is challenging. There are standard ways to quantify individual drug efficacy. These measurements are taken at a heterogeneous population level and thus ignore effects of drug action or mechanism in single cells or cell populations. Because our knowledge is limited in this way, we are often surprised to find that similarly classified cancer drugs can have disparate effects in patients. Single-cell techniques such as flow cytometry allow us to uncover relationships between drugs. We have created an automated workflow to analyze the effect of cancer drugs at different stages of the cell cycle. We have discovered subgroups of therapeutics that act very differently, as well as the cells responsible for these differences.
Materials and Methods: We used intracellular flow cytometry to obtain single cell measurements of DNA content, cell cycle proteins, and state-specific signaling proteins in tandem after application of cancer drugs. Next, we used a method we call percentile mesh, which gives an overall description of the data. Then, we clustered these mesh-based feature vectors to determine groups of drugs which act similar to each other. Next, we ranked and determined regions of interest which best differentiated different drug samples from different groups of drugs. Finally, we compared features of these regions across control and drug-treated samples, and mapped these features back to locations in the cell cycle.
Results: We applied 4-6 concentrations of >10 approved cancer drugs to an acute lymphoblastic leukemia (ALL) cell line, and performed analysis with our workflow. Although there exist clearly defined drug classes, we found profiles of underlying biochemical events that differed between drugs of the same class. In addition, we determined in which stages of the cell cycle drugs differentially affected the cells.
Conclusions: In recent years, advancements have allowed researchers to analyze the effects of drugs on individual cells in great detail. With our high-parameter flow cytometric approach we extended these efforts, monitoring the effects of drugs on proteins during the cell cycle. As a result, we achieved an even finer-grained view of how a drug affects a cancer cell. This new approach will be useful in finding differences between new potential therapeutics and cells in an era of personalized medicine.

poster

S5

Methods for Adjusting Dietary Recall Data

Joanna Lankester
Stanford University

Margaret Brandeau
Julie Parsonnet

Purpose
A quantitative understanding of dietary energy intake is essential to understanding its effects on body weight in the population. Unfortunately self-reported dietary intake recalls are notoriously skewed as individuals tend to under-report their intake [1]. A method of correcting for systematic under-reporting is needed.

Materials and Methods
We develop two methods of adjusting a dietary recall distribution. Using experimental data [2] with both energy expenditure and dietary recall, we predict under-reporting using biomarkers; as a second method, we adjust a dietary recall distribution (normalized to theoretical resting metabolic rate [3]) to its expected distribution.

Results
We validate our methods via a population simulation using NHANES (National Health and Nutrition Examination Survey) datasets from both 1970 and 2007. Values of normalized energy intake outside of a range from experimental data are designated "failures". The simulation shows a failure rate of 1-2% for the first method using biomarkers, and 6-11% for the second distribution adjustment method, compared to over 33% for the unadjusted data.

Conclusion
Both methods give a significant improvement in dietary recall data for a cross-section of the population. The biomarker data method gives particularly reliable output that could be used in further population modeling applications.

References
1. Briefel et. al. Am J Clin Nutr. 1997 Apr;65(4 Suppl):1203S-1209S.
2. Subar AF et. al. Am J Epidemiol. 2003 Jul 1;158(1):1-13.
3. Harris JA, Benedict FG. Proc Natl Acad Sci USA. 1918 Dec;4(12):370-3

poster

S6

Computational Prediction of Human Cell Cycle Genes

Debashis Sahoo

Stanford University

Purpose

Human cell cycle genes have been traditionally characterized in a genome-wide
scale by identifying periodically expressed genes along the different cell cycle phases. However, the identified genes are very noisy because of loss of
synchronization, few mitosis, noisy microarray data, abnormal culture conditions, and artificially transformed human cells for rapid proliferations.

Materials and Methods

We present a new approach to identify human cell cycle genes by analyzing 25,000+ human microarrays from GEO database. We used Boolean implication [2], which is a simple if-then relationship between the high and low expression values of two genes. For example if gene X is high, then gene Y is low, which is stated as X high => Y low, is one of the six different types of Boolean implications. We also used MiDReG algorithm [1], which can predict intermediate markers in a given developmental pathway given two or more end point markers. Our new approach of identifying human cell cycle genes is based on the concept used by MiDReG using two known end point markers: CCNB2 for M phase of cell cycle and CCND1 for G1 phase of the cell cycle.

Results

This approach identified a list of 64 genes, 40 genes (62.5%) are common with 480 Bar-Joseph et al. 2007 cell cycle genes, and 41 genes (64%) are common with 874 genes from Whitfield et al. 2002 cell cycle dataset. Therefore, our approach independently identifies periodically expressed genes without using any timecourse microarray experiment. Further, our predicted genes are mostly associated with known S-phase and G2/M-phase genes.

Conclusions

Our analysis is able to predict genes associated with an intermediate stage, starting with genes from two known stages of the cell cycle. Additional benefit of this analysis is that it will be able to improve the current noisy list of human cell cycle genes.

References

[1] Debashis Sahoo, Jun Seita, Deepta Bhattacharya, Matthew A. Inlay, Sylvia K. Plevritis, Irving L. Weissman, David L. Dill. MiDReG: A Method of Mining Developmentally Regulated Genes using Boolean Implications. Proc Natl Acad Sci U S A. 2010 Mar 30;107(13):5732-7. Epub 2010 Mar 15.
[2] Debashis Sahoo, David L. Dill, Andrew J. Gentles, Rob Tibshirani, Sylvia K. Plevritis. Boolean implication networks derived from large scale, whole genome microarray datasets. Genome Biology.

poster

S7

A Model of Regulatory Program Differentiation in Immune Cell Development

Vladimir Jojic
Stanford University

Tal Shay
Aviv Regev
Daphne Koller

Purpose

We introduce a novel model of gene expression aimed at uncovering the differential regulation that underlies immune cell development. This model builds on a body of work in module network reconstruction and extends it by permitting within module regulatory program variation.

Materials and Methods

This new framework trades off two competing desiderata. First is the preference for preserving regulatory programs among related stages in cell development. Second is the flexibility to capture sudden and substantial changes in regulator roles, for example, a switch from activator to repressor. The problem of fitting this model is cast as a penalized linear regression problem based on tree-guided fused lasso and elastic net penalties. These penalties jointly promote discovery of correlated lineage specific regulators. In addition, we employ a coarse-to-fine strategy, whereby fine modules nested in coarse modules are encouraged to share regulatory programs with their parent coarse modules but can contain deviations specific to their gene set. Importantly, the quality of the model fit is assessed by the BIC score and accounting for the degrees of freedom particular to the model. These considerations prevent the introduction of spurious regulators or regulator role change events.

Results

We demonstrate the utility of this model by analyzing ImmGen compendium data. This data consists of gene expression measurements from 211 distinct cell types of the mouse immune system. In this analysis, the model leverages a developmental tree to encourage conservation of the regulatory programs between parent and daughter cells. Thus, regulatory programs preserved throughout a particular lineage are recovered, in addition to deviations specific to sublineages. We present examples of differential regulation specific to T-cells, B-cells and NK lineages, highlighting lineage-specific roles of regulators.

Conclusions

This novel method enables discovery of condition specific regulatory networks while encouraging conservation of regulatory programs across related conditions. Hence it can uncover regulatory relationships that span different stages in differentiation and disease progression.

poster

S8 Highly Occupied Target (HOT) regions in the C. elegans genome

Eric Van Nostrand
Stanford University

modEncode consortium
Michael Snyder
Stuart K. Kim

Purpose: As part of the modEncode project, regulatory targets for all C. elegans transcription factors (TFs) are being identified by ChIP-seq. In initial analyses of the first 23 factors to be completed (1), we unexpectedly discovered a new type of DNA domain that is bound by most or nearly all TFs, which we term a HOT (high occupancy target) region. Computational analysis has revealed that HOT regions are a new mechanism for transcriptional regulation of ubiquitously expressed and essential genes.

Materials and Methods: This analysis made use of numerous datasets generated as part of the C. elegans modEncode project. ChIP-seq experiments were performed using a GFP antibody in transgenic worms expressing a GFP-tagged transcription factor (2). Other datasets included RNA expression measurements from both whole worms as well as individual tissues, GFP expression in specific cells of the worm, and gene knockdown phenotypes (1, 3, 4).

Results: Upon profiling the binding sites for only 23 different TFs, we identified 304 regions of clustered TF binding of 15 or more TFs, which we have termed HOT (Highly Occupied Target) regions. Using multiple controls, we showed that HOT regions are not an artifact of the experimental procedure but are true incidences of high TF occupancy. HOT regions are not enriched for known TF binding motifs, suggesting TF association to HOT regions results from protein-protein instead of protein-DNA interactions. Genes located near HOT regions are characterized by ubiquitous, high expression, and tend to be essential for viability.

Conclusions: A large-scale analysis of worm ChIP SEQ data has led to the discovery of an unexpected gene regulatory mechanism in which most or all TFs are bound to the same small DNA domain (termed a HOT region). HOT regions result in high-level expression of housekeeping genes in most or all tissues in the worm, and are typically associated with genes with an essential function..

References:
(1) Gerstein*, Lu*, Van Nostrand*, Cheng*, Arshinoff*, Liu*, Yip*, Robilotto*, Rechtsteiner*, Ikegami*, Alves*, Chateigner*, Perry*, Morris*, Auerbach*, Vielle*, Niu*, Rhrissorrakrai,* modEncode Consortium, et al. Integrative Analysis of Functional Elements in the Caenorhabditis elegans Genome by the modENCODE Project. (submitted)
(2) M. Zhong et al., Genome-Wide Identification of Binding Sites Defines Distinct Functions for Caenorhabditis elegans PHA-4/FOXA in Development and Environmental Response. PLoS Genet. 6, (2010)
(3) X. Liu et al., Analysis of Cell Fate from Single-Cell Gene Expression Profiles in C. elegans. Cell 139, 623-633 (2009)

poster

S9

Transmission distortion in Crohn's disease risk gene ATG16L1 leads to sex difference in disease association

Linda Liu
Stanford University

Marc Schaub
Marina Sirota
Atul Butte

Purpose
Crohn's disease (CD) is an autoimmune inflammatory disease of the bowel that affects millions of people around the world. Genome-wide association studies in multiple populations have demonstrated that variants in the autophagy gene ATG16L1 confer increased risk of developing CD. There is also evidence suggesting that disease onset and pathogenesis, as well as general immune processes involving autophagy, are different between males and females, however there have been no previous reports of a sex difference in this gene.

Materials and Methods
We used genotyping data from a Caucasian WTCCC cohort of 1748 CD cases and 2938 controls to investigate variants in ATG16L1 for sex differences in association with disease risk.

Results
The single nucleotide polymorphism rs3792106 showed a significant sex effect with heterozygous odds ratios (Woolf p=0.037). Surprisingly, the difference was found to arise from discrepancy in allele frequencies between male and female healthy control subjects (Chi-square p=0.0045) rather than CD cases. Using HapMap 3 populations (ASW, CEU, MEX, MKK and YRI), we found evidence of transmission distortion at this locus based on parental origin of alleles. We detected significant maternal under-transmission of the risk allele in these healthy individuals (Chi-square p=0.02).

Conclusion
We hypothesize that different transmission patterns between males and females may be responsible for sustaining the disparate allele frequencies in healthy populations, and furthermore that a disease mechanism implicated in CD alters the allele ratio seen in diseased patients. To our knowledge, this is the first report of a sex difference in the ATG16L1 gene. The possible implications in Crohn's disease and basic human biology present interesting areas for future investigation.

poster

S10

Host Genetic Interactions During Viral Infection: A Systems Virology Approach to λ Phage Infection of E. Coli

Elsa Birch
Stanford University

Nathaniel Maynard
Markus Covert

The complex infection process initiated by the viral genome arises from the interaction of the infectious agent with its host. Understanding this relatively small set of instructions as it interprets and co-opts its host's resources is crucial to addressing the problem of viral infection. Using the fundamental system of lambda phage and E. coli, we investigate the host-viral genetic interactions the disruption of which lead to a reduced infectivity. We gain additional insight into gene roles and interactions by observing population dynamics during infection.

Poster Abstracts

poster

1 Matching Cancer Genomes to Established Cell Lines for Personalized Oncology

Joel Dudley
Stanford University

Atul J. Butte

The diagnosis and treatment of cancers, which rank among the leading causes of mortality in developed nations, presents substantial clinical challenges. The genetic and epigenetic heterogeneity of tumors can lead to differential response to therapy and gross disparities in patient outcomes, even for tumors originating from similar tissues. High-throughput DNA sequencing technologies hold promise to improve the diagnosis and treatment of cancers through efficient and economical profiling of complete tumor genomes, paving the way for approaches to personalized oncology that consider the unique genetic composition of the patient's tumor. Here we present a novel method to leverage the information provided by cancer genome sequencing to match an individual tumor genome with commercial cell lines, which might be leveraged as clinical surrogates to inform prognosis or therapeutic strategy. We evaluate the method using a published lung cancer genome and genetic profiles of commercial cancer cell lines. The results support the general plausibility of this matching approach, thereby offering a first step in translational bioinformatics approaches to personalized oncology using established cancer cell lines.

poster

2

The Imaging Biomarker Ontology: ontology-based support for imaging biomarker research

Tiffany Ting Liu
Stanford University

Erica Savig
Daniel Rubin
David Paik

Purpose:
With the advancement of pre-clinical molecular imaging techniques, a wide array of novel imaging biomarkers have been developed and demonstrated effectiveness in quantifying biological processes and in early prediction of therapeutic outcomes. However, integration and re-use of imaging biomarker data and knowledge is limited because most quantitative measurements are only available in the text-based literature that requires expertise and time to synthesize. The purpose of this project is to develop an ontology to represent and integrate the heterogeneous knowledge in the domain of imaging biomarkers, for the goal of developing applications to enable the storage and retrieval of desired imaging biomarker measurements, to mine the expanding imaging literature, and to discover novel imaging biomarkers.

Materials and Methods:
We conducted a literature review of the journal Molecular Imaging and Biology. We focused on abstracting terms and relationships that characterize and annotate an imaging biomarker. This approach allowed us to examine specific instances of biomarkers and to be able to build the ontology from the bottom up.

Also, we reused publicly available ontologies, including MeSH (Medical Subject Headings), NCI thesaurus, GO (Gene Ontology), Disease Ontology, FMA (Foundational Model of Anatomy), and BIRNLex (a controlled terminology for Biomedical Informatics Research Network).

We have built the frame-based ontology using Protégé, a widely used ontology editing tool.

Results:
The ontology we have developed integrates imaging biomarker knowledge in different fields represented by 9 top-level classes, including Experimental Subject, Biological Intervention, Imaging Agent, Imaging Instrument, Image Processing and Analysis Algorithm, Biological Target, Indicated Biology, and Biomarker Application. The ontology currently contains about 490 classes, of which Imaging Instrument (115 classes) and Indicated Biology (106 classes) are the two largest classes. We also created synonyms and definitions for classes in the ontology. To ensure consistency, is-a relationship is strictly conserved in every branch of the ontology.

Conclusion:
We have developed the Imaging Biomarker Ontology to support imaging biomarker research. It integrates heterogeneous knowledge of imaging biomarkers, as well as bridges pre-clinical and clinical imaging biomarker research. To our knowledge, this is the first ontological representation of knowledge in imaging biomarker research. The complete Imaging Biomarker Ontology will soon be made available to the public domain. Following the ontology development, we will validate the ontology using published imaging data and demonstrate its utility in various applications.

poster

3 Robustness to mutations as a key property of functional RNAs

Pablo Cordero
Stanford University

Rhiju Das

Structural characterization of currently known functional RNAs that distinguish them from random ribonucleotide sequences is a common problem found in areas ranging from de novo non-coding RNA discovery to molecular design. Many of the current approaches carefully dissect the structural ensemble of functional RNAs to obtain measures of stability and modularity. These methods, however, usually do not provide a statistically significant signal to noise ratio when comparing biological versus random sequences. Here, we investigate the structural regularity of the set of an RNA's single point mutants. By simulating chemical footprinting experiments through RNA secondary structure predictions, we calculate statistical measures of functional RNAs' mutational robustness versus various sets of random sequences. These measures are shown to distinguish the different groups better than single sequence structural signals. Ultimately, they could be used as a tool to scan for novel functional RNAs or calculate the exact lengths of currently known non-coding transcripts.

poster

4

Comparison of Cartesian and torsional normal modes for describing conformational changes in proteins

Jenelle Bray

Stanford University

Purpose:
Normal mode analysis can be used to model large scale conformational changes in proteins. However, in Cartesian space, this can lead to unrealistic deformation of the proteins. This deformation can be avoided by performing the normal mode analysis in torsional space. This has the added benefit of a large decrease in the degrees of freedom. This study will compare the performance of Cartesian and torsional normal modes in describing protein conformational changes.

Materials and Methods:
Thirteen pairs of protein structures were chosen at a test set. Each pair represents two states of the same protein before and after a conformational change. Cartesian and torsional normal mode analyses were performed on all the test cases. The performance of each method was analyzed by RMSD and dot products.

Results:
Torsional modes do as well or better than Cartesian normal modes at describing the conformational changes between pairs of proteins in all thirteen test cases.

Conclusions:
Torsional normal mode analysis is a good alternative to Cartesian normal mode analysis. Torsional normal mode analysis prevents unrealistic deformation of the protein, the calculation is less computationally expensive, and it describes protein conformational changes as well or better than Cartesian normal mode analysis.

poster

5 Using Computational Algebraic Topology to Characterize Chromosomal Instability in Cancer

Alex Pankov
San Francisco State University

D. DeWoskin
M. Nguyen
R. Scharein
M. Vazquez
C. Park and J. Arsuaga

Purpose
DNA copy number aberrations (CNAs) are often associated with cancer initiation and progression and can be detected using microarray technologies. We propose a method Multidimensional Analysis of CGH (MDaCGH), which draws from the theory of computational algebraic homology to find recurrent CNAs.

Materials and Methods
MDaCGH method:
CGH data is mapped using a sliding window method to n–dimensional space. Next we construct the 1-skeleton of the Vietoris-Rips simplicial complex which is given by the graph where each pair of points whose distance is less than an ε, in a fixed dimension n, are connected. For each sample and in each dimension, β0 is calculated at each value of ε. For each dimension, samples are grouped by phenotype, and β0 at each epsilon is calculated. Finally, a p-value for each dimension is calculated using the permutation test, and combined over all dimensions using the Hartung method, yielding the significance for the region.

Results
Applying this method, we show that MDACGH analysis of 147 primary glioblastoma tumor samples published by the TCGA network [1] detects all reported deletions and 60% of the amplifications while using a subset of the samples. It also finds two new regions of CNAs on arms 5q and 18p not previously reported by TCGA.

We also applied MDACGH to the Climent et al. [2] breast cancer dataset, with the goal of finding CNAs associated with disease recurrence. MDACGH finds two novel regions of CNAs (1q and 6q) that were not found by the original study.

Conclusions
The current implementation of MDaCGH combines the results for all dimensions into a single p-value. We hypothesize that there is a relationship between the size of the aberration and the dimension at which the difference between β0 curves are significant. We performed simulations and tested different combinations of amplifications from dimension 2 to 10 and found that dimensions smaller than the size of the aberration had significant p-values while larger ones lost significance. Some CNAs may occur in combination rather than independently. Higher order homology, such as the first homology group that computes two-dimensional holes in the cloud of points, has the potential to detect more complex interactions, including these combinations.

References
[1] Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455
[2] Beroukhim R, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc Natl Acad Sc U S A. 2007
[3] Climent JA. et al. Deletion of chromosome 11q predicts response to anthracycline-based chemotherapy in early breast cancer. Cancer Res, 2007

poster

6 Using confirmed dicistronic gene structures from several Drosophilid species to understand eukaryotic

Henry C. Hunter
San Francisco State University

Christopher D. Smith

RNA secondary structures have been found in humans to act as 5′-UTR regulatory elements to allow proper translation. Mutations or any other alterations to those secondary structures can drastically affect the translation of the mRNAs, causing diseases such as Multiple myeloma, fragile X syndrome, and numerous cancers [1]. Similar RNA secondary structures have been found in multicistronic genes. Encoding multiple genes in a single mRNA, multicistronic gene structures are common among prokaryotes and viruses, but their existence in eukaryotic cells has only recently been observed. High-quality genomic annotation of *Drosophila melanogaster* (fruit fly) has revealed over 100 dicistronic genes, where two non-overlapping genes are expressed from a single mRNA transcript [2]. The mechanisms that mediate downstream dicistronic expression in eukaryotes are poorly described, but evidence has shown that RNA structures in the mRNA may promote translation.

Interestingly, these same genes can appear as standard monocistronic transcripts in some distantly related Drosophila species, suggesting that there exist mechanisms that allow genes to merge and become expressed as dicistronic products. Recently, 11 other Drosophilid genomes have been completed, providing an excellent system to study the structure of dicistronic genes over evolutionary time. While full-length cDNA evidence supports the existence of these genes in *D. melanogaster*, to date there has been no direct evidence that these genes are also dicistronic in other species.

We identified and annotated orthologs of several *D. melanogaster* dicistronic genes and their putative gene structures in multiple Drosophilid species. We used these annotations to develop PCR primers to confirm the existence of each dicistronic transcript in cDNA samples. RT-PCR cDNAs were sequenced and compared to verified dicistronic genes to identify gene structural changes over evolution. Currently, we are focusing on prediction of conserved RNA secondary structures and location of the secondary structures on the mRNA that give further insight into how dicistronic genes function in eukaryotes. These results are useful to refine existing dicistronic gene annotations, discover their function in eukaryotes, and provide a detailed comparative system to study the forces that shape the evolution of gene structures over evolutionary time.

References:
1. Chatterjee, Sangeeta, and Pal, Jayanta K., Role of 5′- and 3′-untranslated regions of mRNAs in human diseases, *Biol. Cell* **101**: 251-262. (2009)
2. Kozak, Marlyin, Comparison of Initiation of Protein Synthesis in Procaryotes, Eucaryotes, and Organelles, *Microbiological Reviews* **47**: 1-45. (1983)

poster

7

Work in Progress: Identification of Protein Functional Sites using Machine Learning with Support Vector Machines

Gurgen Tumanyan

San Francisco State University

Ljubomir Buturovic
Dragutin Petkovic
Mike Wong
Russ Altman
Grace Tang

Detection of functional sites in proteins is an important problem in computational biology and has wide implications in computational drug discovery. FEATURE by Stanford Helix Group (http://feature.stanford.edu/webfeature/ is a system for identifying active sites in proteins based on a set of 3D structural and biophysical properties (features) of functional microenvironments. FEATURE examines protein micro environments of 7A radius, generates feature vectors for structural elements in 1A shells and identifies structural and bio-physical motifs that characterize the site. These feature vectors are then processed by machine learning algorithms to identify functional sites based on models that leverage prior training data. Current FEATURE uses Naïve Bayesian classifier as a machine learning algorithm.

In this collaborative work between Stanford Helix Group and SFSU Computer Science Department and SFSU Center for Computing for Life Sciences, we explore the application of Support Vector Machine Learning algorithm (SVM) and compare it with the currently used Naïve Bayesian classification. We analyze in depth the performance of the two classification algorithms and examine the misclassified Thioredoxin binding sites. Additionally we compare performance of the classifiers on a set of 56 functional site datasets from SeqFEATURE library using cross-validation and Receiver Operating Characteristic

Our initial results over the 56 datasets, with approximately 300 positive and 3000 negative samples indicate that for some functional families, the functional and non-functional sites are linearly separable in the input space. For linearly non-separable cases experimental results indicate that Support Vector Machine may be advantageous for some functional sites. Future work includes improved characterization of functional sites, comparisons on larger data sets and on additional functional models.

poster

8 UniPeak: Quantification of parallel sequencing experiments

Joseph W. Foley
Stanford University

Cheryl L. Smith
Arend Sidow

Purpose: High-throughput sequencing has enabled highly quantitative investigations of DNA-protein interaction and gene expression. Peak callers are used to discover enriched genomic regions independent of reference annotations, but the available methods generally consider each sample separately, so multiple peak lists must be reconciled to compare different experimental conditions. We present a next-generation peak caller, UniPeak, the first tool that efficiently combines data from an unlimited number of parallel experiments and provides output that can be analyzed with existing methods.

Materials and Methods: UniPeak accepts as input short-read alignments from common mappers. Alignments are represented as frequency of 5 read start positions and pooled across all samples. As in QuEST [1], a combined density profile is generated by kernel density estimation, and enriched regions are called where the density exceeds a threshold of fold-enrichment relative to a uniform background. 5 read start positions from each sample are counted within these regions, yielding a regions × experiments matrix analogous to those used for microarray analysis, but representing tag enrichment as raw counts.

Results: We demonstrate applications for parallel sequencing analysis using a variety of public ChIP-seq data from the ENCODE Consortium as well as novel 3SEQ [2] data from mammalian embryogenesis. With the ChIP-seq data, we find that clustering analysis recovers the biological similarity between cell types, and comparison with histone methylation marks reveals a relationship with the activity of transcription factors and transcriptional machinery. Examination of the 3SEQ data produces an inventory of both annotated and unannotated transcripts whose expression is related to sex differentiation and development in mammals.

Conclusions: UniPeak offers an efficient platform for analysis and comparison of multiple high-throughput sequencing experiments. It can be used conveniently in a bioinformatic pipeline, preceded by short-read alignment, and followed by normalization and standard methods for analysis of parallel microarray experiments. This enables large-scale studies of DNA-protein interaction and gene expression across varying experimental conditions.

References:

1. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. Nat Methods 2008.

2. Beck AH, Weng Z, Witten DM, Zhu S, Foley JW, Lacroute P, Smith CL, Tibshirani R, van de Rijn M, Sidow A, West RB. PLoS ONE 2010.

poster

9

Defining the relationship between sequence and three-dimensional structure conservation in RNA

Emidio Capriotti

Stanford University

Marc A. Marti-Renom

Emidio Capriotti1,2 and Marc A. Marti-Renom3

1 Department of Bioengineering, Stanford University,
Stanford (CA), United States of America.
2 Department of Mathematics and Computer Sciences, University of Balearic Islands,
Palma de Mallorca, Spain.
3 Structural Genomics Unit. Bioinformatics and Genomics Department.
Centro de Investigación Príncipe Felipe. Valencia. Spain

In the last decade, the number of available RNA structures has rapidly grown reflecting the increased interest on RNA biology [1] Similarly to the studies carried out in the late eighties for proteins [2], which gave the fundamental grounds for developing comparative protein structure prediction methods, we are now able to quantify the relationship between sequence and structure conservation in RNA. Here we introduce an all-against-all sequence- and three-dimensional (3D) structure-based comparison of a representative set of 451 RNA structures using the SARA algorithm [3]. Selecting a set of 589 high similarity alignments from 114 RNA chains, we have quantitatively confirm that: (i) there is a measurable relationship between sequence and structure conservation, (ii) evolution tends to conserve more RNA structure than sequence, and (iii) there is a twilight zone for RNA homology detection. We found that similar to proteins, the structure identity decreased with the decrease of the sequence identity and in agreement with previous work [4] and we observed a higher mean value of percentage of structure identity with respect to the average sequence identity. Using the Infernal program [5], an e-value threshold of ~5x10-4 has been found to be the lower limit of the "twilight zone" for sequence alignment longer than 100 nucleotides. The computational analysis here presented quantitatively describes the relationship between sequence and structure for RNA molecules and defines a twilight zone region for detecting RNA homology. Our work could represent the theoretical basis and limitations for future developments in comparative RNA 3D structure prediction.

REFERENCES

1. Capriotti E, Marti-Renom MA (2008) Computational RNA structure prediction. Current Bioinformatics. 3:32-45.
2. Chothia C, Lesk AM. (1986). The relation between the divergence of sequence and structure in proteins. Embo J. 5:823-826.
3. Capriotti E, Marti-Renom MA. (2009). SARA: a server for function annotation of RNA structures. Nucleic Acids Res. 37:W260-W265.
4. Abraham M, Dror O, Nussinov R, Wolfson HJ. (2008). Analysis and classification of RNA tertiary structures. RNA. 14:2274-2289.
5. Nawrocki EP, Kolbe DL, Eddy SR. (2009). Infernal 1.0: inference of RNA alignments. Bioinformatics. 25:1335-1337.

poster

10

An open-source library of data clustering algorithms for GPUs

Kai J. Kohlhoff
Stanford University

Marc H. Sosnick
William Hsu
Vijay S. Pande
Russ B. Altman

Purpose

Data clustering techniques are an essential component of many data analysis toolkits and are frequently used in biological applications[1-4]. With the availability of simplified APIs for the use of graphics processors (GPUs) as general purpose computing platforms, scientific applications can benefit from speed-ups of one to two orders of magnitude[5, 6]. For large data sets, such as those retrieved from genome sequencing efforts, gene expression profiles determined by gene arrays, or protein structural information, it is thus becoming possible to reduce time-intensive tasks from hours of computing time down to just minutes. To make the GPUs' unparalleled floating-point operations-to-cost ratio available to researchers with little or no background in parallel programming, we have developed CAMPAIGN, an open-source library of data clustering algorithms. This library can be included in existing projects and is designed to completely hide the underlying GPU-specific program code behind an easy to use interface.

Material and Methods

Clustering algorithms were implemented using Nvidia's freely available programming API 'C for CUDA'[7]. Performance tests were carried out between conventional sequential data clustering code on an Intel Xeon E5420 processor at 2.5 GHz, and parallel code executed on an NVIDIA Tesla C1060 graphics multiprocessor at 1.3 GHz. Multiple types of data were used for the tests: Two synthetic data sets of uniformly and Gaussian-distributed data, a collection of data vectors from the FEATURE analysis software for key biophysical and biochemical features of biological structures[8], and gene expression data from 9,395 Affymetrix arrays testing a set of 20,099 human genes[9].

Results

We have developed CAMPAIGN, an open-source library of 'Clustering Algorithms for Massively Parallel Applications Including GPU Nodes'. The library contains GPU versions of four popular clustering algorithms, K-means, K-centers, Hierarchical clustering, and Self-organizing map, with speed-ups ranging from one order of magnitude for hierarchical clustering to more than 200-fold speed-up for K-means. The source code is provided for free download and the library is kept modular, flexible, and extensible. By following best programming principles, the library assures ease of use and simplifies future upgrading and maintenance. Furthermore, the clustering algorithms are structured to achieve speed-ups for a wide range of numbers of data points, cluster centers (where applicable), and dimensionality of data points, in practice only limited by the amount of available physical memory.

Conclusion

The CAMPAIGN library makes it feasible to perform highly accelerated data clustering tasks in reasonable time. This allows processing larger data sets than previously possible, and to develop new types of data analysis protocols or refining existing ones, including, e.g. running clustering repeatedly based on different starting configurations, or increasing the number of iterations until convergence.

References

1. Andreopoulos B, et al, Brief Bioinform, 2009. 10(3): p. 297-314

poster

11

Toward the Use of cloud computing for bioinformatics: case study of Amazon Elastic Compute Cloud EC2

Jinesh H. Lalan
San Francisco State University

Mike Wong
Dragutin Petkovic

Cloud computing provides computation services on demand (at the time of convenience, with scalable resources, e.g. CPUs, memory, storage, etc.), thus reducing maintenance and facilities costs. These features make Cloud Computing an attractive alternative to traditional high-performance computing (HPC) approaches for bioinformatics researchers. Several manufacturers offer this service, among others Amazon, with its Elastic Compute Cloud http://aws.amazon.com/ec2.

This work is a collaboration between San Francisco State University CS Department, SFSU Center for Computing for Life Sciences and Stanford Helix group. FEATURE by Stanford Helix Group (http://feature.stanford.edu/webfeature) is a system for identifying active sites in proteins based on a set of 3D structural and biophysical properties (features) of functional microenvironments. Its usage involves high volumes of data (including those generated by molecular dynamics) that can be performed in parallel thus making it much more effective research tool.

Goals of this work are several: a) understand this new technology in terms of computational infrastructure; b) investigate usage of Amazon EC2 services for frequently-used bioinformatics software (e.g. BLAST) and FEATURE; c) producing a tutorial, documentation and software tools to help other researchers in using this service.

Progress today includes: a software system that securely controls parallel computing on the cloud from a laptop or desktop computer; simplified documentation that enables bioinformatics researchers to harness the cloud; and a performance comparison of BLAST on the Amazon EC2 Cloud and on a traditional Beowulf cluster.

Future work includes: Exploring and enabling usage of Amazon EC2 for FEATURE application; investigating high memory architectures on the cloud; and securing cloud machine images for bioinformatics research with sensitive data.

poster

12

The Beginnings of the Phenologue: An Ontology of the Genetic Causes of Autism Spectrum Disorders and Their Resulting Phenotypes and Endophenotypes

Daniel Y. Li

Stanford University

Amar K. Das

Purpose

Autism is a highly heritable disease with about 90% heritability, yet no specific genes have been successfully identified that are causal. Legions of studies that attempt to identify these genetic origins have been carried out, but as far as we know, they are all published as free-text papers dispersed across numerous journals. We felt it would be helpful for researchers to have an ontology that presents what is currently known and suspected about autism's causes at one central resource.

Methods

To design this ontology, I decided to focus on formalizing the claims made in a few specific papers and building the ontology around what I learned from them (illustrated in Steps 1–4 below). Constructing the claims and supporting claims for a paper involved identifying the statements and hypotheses the authors made, reasoning what other statements were needed to support those hypotheses, and searching for additional papers that provided those statements. In this way, the ontology is a hierarchy of evidence and statements that support hypotheses made about autism, and those hypotheses in turn can be support for even higher-level hypotheses, forming a blueprint of what we currently know about autism spectrum disorders.

Results

Results are shown in the figures in the poster.

poster

13

Development of FEATURE and WebFEATURE 2.0 Software using Modern Software Engineering

Trevor Blackstone
San Francisco State University

Gemma Lee Fu-Sun
Mandar Modgi
Teague Sterling
Mike Wong
Grace Tang
Russ Altman
Dragutin Petkovic

Detection of functional sites of proteins is an important problem in computational biology and has wide implications in computational drug discovery. WebFEATURE by Stanford Helix Group (http://feature.stanford.edu/webfeature) is an Internet based software system for predicting protein functions based on a set of 3D structural and physicochemical properties (features) of functional microenvironments. It offers users the ability to submit their own protein structure for analysis and to get the results via e-mail. In its backend, it contains FEATURE software which uses machine learning teachniques to predict the functional sites based on learned models which leverage prior training data. Current FEATURE uses Naïve Bayesian classifier as a machine learning algorithm.

This work is a collaboration between San Francisco State University CS Department, SFSU Center for Computing for Life Sciences and Stanford Helix group. The goal is to enhance current WebFEATURE and FEATURE software systems and to make them easily available and attractive tools for external research community. We plan to achieve this in several ways: a) Make FEATURE more scaleable, maintainable, documented and efficient; b) provide FEATURE as a standalone desktop software; c) Enhance WebFEATURE by making it faster and easier to use; d) Create APIs (application programming interfaces) to allow easy integration of different machine learning algorithms and addition of new functions; and e) to develop modern software development and maintenance environment for the whole system. The challenge is to do all this while not impeding researchers' and students' freedom to develop new research software.

Our approach is based on most modern software engineering methods using agile and iterative processes, User Centered Design, continuous build and test process, use of modern code profiling tools, peer code reviews, and in creation of an integrated build and test environment to allow reliable software upgrades and maintenance. User Interface was redesigned after numerous interactions with typical users.

Our accomplishments so far include: completion of FEATURE 2.0 desktop version which is ten times faster; more resource efficient and easy to use; improvement of UI for WebFEATURE; creation of automated software development, build and test system for FEATURE maintenance and upgrades; and software development and maintenance polices that allow flexible experimentation while at the same time ensuring stable main code base. We plan to launch all these enhancements as WebFEATURE 2.0 and FEATURE 2.0 in early 2011.

poster

14

Adaptive genetic variation in Quercus lobata

Eric Levy
UCLA

Victoria Sork
Stephanie Steele
Andrew Eckert
Karen Lundy
Kevin Squires

Climate change has altered ecosystems and compromised habitats for many species throughout the world. Of particular concern is the survival of long-lived trees, which shape local biodiversity. Understanding the genetic structure of local populations could provide key information regarding the potential for adaptation as a response to such climate change. We are studying adaptive genetic variation in Quercus lobata Née, a California valley oak.

Current genomic technology has allowed for such information to be discovered in a non-model species such as Q. lobata. We used a candidate gene approach to target genes with a possible functional association with geographical gradients and climate variables, such as cold and heat tolerance, bud burst, and osmotic stress. We obtained high-quality sequences for 49 genes sampled from a total of 45 individuals spread across 13 sites in California.

Analysis of these sequences has indicated that SNPs from several genes indeed show an association with the climate variables (Sork et al. in prep). We have also used existing databases to align our genes against available amino acid sequences from closely related plant species. Using this information, we have determined the reading frames of our genes and separated them into coding and non-coding regions. We have been able to annotate SNPs from these coding regions as causing synonymous or non-synonymous changes. We are now using this information to find more evidence of positive selection by looking for elevated ratios of non-synonymous to synonymous mutations of the genes that appear highly correlated to the climate variables. Finally, we are looking at in what part of the genes the significant SNPs occur and the frequency of those that cause non-synonymous changes in the coding regions.

poster

15

The Stanford EEG Viewer: A High-Throughput Platform for the Visualization and Analysis of Sleep Data

Hyatt Moore IV
Stanford University

Simon Warby
Emmanuel Mignot

Purpose
Although the function of sleep is not entirely clear, the human brain requires it to function properly. During sleep, the brain generates a complex and dynamic pattern of electrical activity that can be measured by electroencephalography (EEG). EEG data can be analyzed by traditional spectral analysis, but the EEG signal also contains many fast and discrete features, such as sleep spindles and k-complexes that can be thought of as an 'EEG fingerprint' - unique to each individual and strongly influenced by the genes we inherit. Changes to the EEG fingerprint have been associated with several neurological diseases.

The purpose of this project is to develop a software platform that can provide immediate visual and analytical tools to examine EEG data. The software will be able to analyze data using traditional spectral analysis, but also incorporates novel analytical methods and machine-learning to characterize specific EEG features. In addition to visualizing data from a single individual, the platform is optimized to process data in a high-throughput and automated fashion. In the US alone, over 1 million EEG sleep studies are conducted annually. By automating these methods of analysis, we can use this information on the large datasets that are required to perform genome-wide association studies to establish genetic links to these EEG features.

Materials and Methods
The Stanford EEG Viewer (SEV) is a Matlab based platform that is capable of reading and analyzing EEG data stored in standardized format. Modular plugins, such as Artifact detection, Power-Spectral analysis and Sleep Spindle characterizer can be easily added and customized as needed.

Results
The SEV is currently on its 25th local release. The visualization tools were initially used to fine tune detection algorithms that identified poor quality data, as well as EEG features that could be used for genetic analysis. Using these parameters, the SEV was then used to analyze >2,000 initial studies (>160 terabytes in 12 hours) and produce statistical summary reports. These reports are currently being used to help locate stable biomarker traits.

Conclusion and Future Aims
The SEV is a software platform for the analysis of EEG data that is capable of processing large datasets for power spectral analysis and discrete EEG features such as sleep spindles. We are looking to deploy new machine-learning algorithms to find novel EEG features.

Website
http://www.stanford.edu/~hyatt4/BCATS/abstract.html

poster

16

Deconvolving Individual Gene Expression Data into Cell-Type Specific Profiles

Chuan-Sheng Foo
Stanford University

Shai S. Shen-Orr
Daphne Koller

Purpose
Gene expression assays are commonly carried out on samples that consist of a mixed population of cells. For instance, expression data obtained from blood is a mix of different cell types with varying proportions. As a result, the measured expression of a gene will be an aggregate measurement of its expression levels in the individual cell types. Variation in cell proportions across samples confounds downstream analysis of the data: averaging attenuates the signal within a given cell type, and differences in gene expression between samples may simply be due to differences in cell proportions. It would thus be ideal to run such assays on purified samples. However, experimental purification methods such as cell sorting are time consuming, costly and affect the gene expression in the samples.

Materials and Methods
We build upon the recent work of Shen-Orr et al. [1] that estimates the average cell-specific gene expression profile across a population of samples. However, our work presents a computational solution that takes an aggregate gene expression measurement and the cell proportions from an individual sample, and estimates cell-specific gene expression profiles for that specific sample. Our method works by performing inference in a probabilistic model that exploits the correlation patterns in the expression data for the different cell-types; these priors are learned from expression profiles of purified samples.

Results
We demonstrate our method on both pseudo-synthetic and real data. Our synthetic data is derived from real cell-type specific expression profiles (not used in training the model) that are artificially convolved with realistic mixing proportions to produce aggregate samples. Preliminary results show that our method is able to recover the original cell-specific gene expression profiles from these aggregate samples to the extent that these estimated expression profiles may be used as surrogates for measured expression on purified samples for each of the individual cell-types in traditional analyses of such data. Our method provides significantly better performance in recovering differentially expressed genes and gene-gene correlation networks than baseline methods applied to the aggregate samples. We also show some suggestive results on real aggregate data from the auto-immune disease SLE.

Conclusion
We have demonstrated the feasibility of obtaining cell-specific gene expression profiles from an individual sample given cell proportions within the sample. Our method enables higher-resolution analysis of gene expression in aggregate samples that elucidates the role of the individual cell-types present in the sample.

References
[1] Shen-Orr et al. Nature Methods 7, 287 - 289 (2010)

poster

17 Developing a Similarity Reference Standard for CBIR Using Matrix Completion

Jessica Faruque
Stanford University

Daniel Rubin
Christopher Beaulieu
Ronald Summers
Aya Kamaya
Grace Tye
Sandy Napel

Purpose: This project develops a way to create a reference standard for visual similarity between liver lesions seen at CT for validating a Content-Based Image Retrieval (CBIR) system by predicting all similarity scores in a large similarity matrix from a small subset of the scores.

Materials and Methods: We displayed 19 portal venous CT images containing liver lesions in all 171 pair-wise combinations to 3 radiologists in random order. For phase 2, each pair of images was rated for overall similarity between the images, and determined if the overall similarity matrix, averaged over the readers, could be predicted from a subset of the entries. We removed a random subset of the entries, and predicted the remaining entries by computing the set of entries completing the matrix that minimized the maximum sum of the singular values of the matrix. We ran 180 iterations of this at a variety of random subsets containing between 8% and 99% of the matrix, computed the RMS error between the calculated and actual entries for each iteration, binned the results into bins of width 10% of the actual entries included in the matrix, and computed the means and standard deviations of RMS error in each bin. We compared its performance to that of an "ideal similarity matrix" created by randomly assigning each image a score and calculating the pair-wise image similarity as the maximum similarity minus the difference between scores between the images.

Results: In the 20-30% bin and all bins containing greater percentages of points of the matrix, the mean RMS error was below 3 points for a 9-point scale. Compared to the "ideal similarity matrix," the actual data followed a similar trend, though as expected, its performance was poorer.

Conclusion: We can predict the remaining entries to within an average of 3 points out of 9 when 30% or more of the entries are included, which may be a feasible method for creating a similarity reference standard for a CBIR system.

poster

18

Simulation modeling of the impact of prophylactic surgery and screening on life expectancy in BRCA1/2

Diego F. Munoz
Stanford University

Bronislava M. Sigal
Sylvia K. Plevritis
Allison W. Kurian

Purpose
Due to high risks of developing breast and ovarian cancer, women with inherited mutations in the BRCA1/2 cancer susceptibility genes are recommended a number of intensive risk-reducing strategies, including prophylactic mastectomy (PM), prophylactic oophorectomy (PO) and screening. The outcomes of these alternatives, however, have not been compared directly and there remains significant controversy about their overall health benefits. We simulated the life histories of women that are found with BRCA1/2 mutations at different ages, comparing the resulting lifetime expectancies when undergoing different risk-reducing strategies.

Methods
We built a Monte Carlo model that simulates the life histories of women with BRCA1/2 mutations, using data from published meta-analyses to estimate their elevated cancer incidence risks and hazard ratios associated to undergoing PM and PO. Given a cancer event, we model the natural history of the disease following the histopathological characteristics of BRCA1/2-associated breast cancers. Screening, which includes annual mammography plus magnetic resonance imaging (MRI), was superimposed on the pre-clinical course of the breast cancer to determine if detection is due to screening or by symptoms. Based on these outcomes, we obtain the patient's overall survival by taking the minimum between the age of other-cause death and the age of cancer death.
Different estimates for life expectancies were obtained by modeling risk-reducing strategies alone and in combination, performed immediately after mutation status determination or delayed by several years.

Results
The greatest gains in life expectancy are achieved by performing PM and PO immediately after carrier status is determined; these gains vary according to age at carrier status determination and range between 7.3-10.7 years for BRCA1, and 3.6-4.6 years for BRCA2, carriers. Dependant on age of carrier status determination and intervention, delaying prophylactic surgery can still provide valuable gains in life expectancy; these gains range between 1-9.9 years for BRCA1, and 0.9-4.3 years for BRCA2, carriers. Adding yearly breast screening provides gains that range between 2.0-10.2 years for BRCA1, and 1.6-4.4 years for BRCA2, carriers. Sensitivity analysis was performed on the estimates for BRCA1/2 cancer incidence and the duration of PO's protective effect.

Conclusions
Our analysis shows that, despite the uncertainties associated to BRCA1/2 cancer incidence and risk-reducing strategy effectiveness, these procedures can provide relevant gains in life expectancy. In particular, even when procedures are delayed, the benefits can still be substantial if performed prior to ages when risk is likely to increase.

poster

19

Assessment of Diffusion Parameters at Pre-, Mid-, and Post-Radiation Therapy in Glioblastoma Patients Receiving Bevacizumab Therapy

Laleh Jalilian
UCSF

Emma Essock-Burns
Soonmee Cha
Susan Chang
Nicolas Butowski
Sarah J. Nelson

Purpose: The purpose of this study was 1) to evaluate diffusion parameters at pre-, mid- and post-radiation therapy (RT) in the contrast-enhancing lesion (CEL) and non-enhancing lesions of postsurgical glioblastoma multiforme (GBM) patients treated with RT, temozolomide and the anti-angiogenic bevacizumab, 2) to assess if diffusion parameters may serve as early markers for disease progression by relating them to clinical outcome of 1-year progression-free survival (PFS).

Methods: 27 newly diagnosed GBM patients treated with surgical resection, radio-, chemo- and bevacizumab therapy underwent standard anatomic magnetic resonance (MR) imaging and diffusion-weighted imaging prior to the beginning of therapy (post resection, pre-RT) and serially following therapy initiation at 1-month (mid-RT), 2-months (post-RT), and every 2 months thereafter. ADC values were obtained at pre-, mid-, and post-RT for normal appearing white matter (NAWM), CEL, T2 hyperintense lesion (T2ALL), and areas of T2ALL that did not include contrast enhancement (NEL). 15 patients with progression data were grouped into early progressors (EP) and late progressors (LP) based on radiographic progression at one year following therapy initiation. Differences in imaging parameters for volumes, median values, and percent changes were assessed amongst scans using signrank tests and between EP and LP using a ranksum test.

Results: CEL volume decreases significantly between pre- and mid-RT ($p < 0.0074$) but not between mid- and post-RT. T2ALL and NEL volume decrease significantly between mid- and post-RT ($p < 0.0005$ and $p < 0.0006$) but not between pre- and mid-RT. Median nADC values for CEL, T2ALL, and NEL demonstrate no significant change at pre-, mid-, and post-RT. There are no significant differences in CEL, T2ALL, and NEL volumes and median nADC between EP and LP at pre-, mid-, and post-RT. EP showed greater decrease in T2ALL and NEL percent change in volume between mid- and post-RT scans as compared to LP ($p < 0.0155$).

Conclusion: Studies have suggested that brain tumors that respond favorably to RT or chemotherapy show an increase in the ADC values shortly after treatment (1). In the present study, an increase in ADC values during RT, chemo-, and bevacizumab therapy was not observed. Bevacizumab may exert a steroid-like effect affecting tumor edema, as demonstrated by decreases in T2ALL and NEL volumes, while maintaining stable ADC values. Further investigation will examine long term effects of bevacizumab on diffusion parameters and assess whether they may serve as biomarkers for disease progression by relating them to clinical outcome of 1-year PFS and overall survival.
References: 1- Mardor et al. J Clin Oncol. 2003;21:1094–1100.

poster

20

Bitrate optimization of a continuous control neural cursor applied to discrete selection tasks

Joline Fan, Paul Nuyujukian
Stanford University

Jonathan Kao
Paul Kalanithi
Vikash Gilja
Cindy Chestek
Stephen Ryu
Krishna Shenoy

Cortically-controlled prosthetic systems (e.g. brain-computer interfaces) enable neural activity describing a subject's motor intent to be translated to useful control signals. Recently, new algorithmic and training advances have led to improved cursor control in an online, closed-loop framework [Gilja et al. and Nuyujukian et al., BCATS 2010]. To further explore the performance characteristics and clinical viability of these systems, we designed and evaluated two discrete selection tasks that point towards interface designs suitable for patient use.

In this study, neural activity was recorded from a 96-channel array implanted in the motor cortex of two rhesus macaques. To decode cursor kinematics from spiking activity, we employed a modified Kalman filter [ibid], which is trained on estimated volitional cursor movement and incorporates a feedback control perspective.

Two selection tasks were investigated. First, the grid task is intended to closely model everyday keyboard design, in which the user is presented with a grid of potential targets and activates instructed targets by dwelling on it for a specified amount of time. Any dwells on non-instructed targets for the same amount of time are considered failures. Second, the radial-arc task presents an alternative keyboard design, in which the user moves the cursor past a radial threshold and towards the instructed target on the perimeter of a circle. Performance was evaluated based on achieved bit rate. Other calculated metrics include selection rate, channel capacity, and success rate.

By sweeping the parameters of both tasks, i.e. number of targets (both), dwell/hold times (grid), and radial thresholds (radial), we show that an optimal task parameter can be achieved for two animal models (one arm free to move or both arms restrained). The optimal task parameter varies with the subject's cursor control and yielded bit rates of around 3 bps. To probe the utility of this framework over long time scales, we ran both monkeys with bootstrapped or previous-day models on the grid task for 7 consecutive days. We show sustained performance throughout the week of bit rates of over 3 bps for over an hour. These findings suggest that the performance of intracortical prostheses may be robust and sustainable for clinically-relevant discrete selection tasks.

poster

21 IRAP: A Knowledge Based Discriminatory Function for Protein Structure

Michael Zhou
University of Washington

Brady Bernard
Jeremy Horst
Ram Samudrala

Proteins are essential in carrying out almost all cellular processes. Despite this, the structure and function of many proteins is unknown. Although protein structures can be determined to great detail by experimental methods, such efforts are time consuming and costly. However, thanks to genome sequencing projects over the last few decades, a large body of genetic and amino acid sequence data is available. By using these sequences, computers can use energy functions to search the various possible conformations of the protein in an effort to model its structure. This project addressed an important part of the computational protein structure puzzle, determining which protein conformations are closest to reality. It took a knowledge-based method developed by Dr. Bernard and Dr. Samudrala for protein-small molecule interactions, and modified it to apply to overall protein structure. This method takes nonbonded interatomic distances in protein conformations and compares them with distances in known protein structures. From this, the favorability of a certain protein conformation is evaluated. We found that the method compared favorably with others in the field, showing high correlations between a conformation's ranked score and its average distance from the actual structure. From these results, we expect that this function has potential to aid the modeling protein structure, especially in concert with other methods in the field.

poster

22

Reconstruction of Signaling Transduction Networks from Mass Cytometery Data

Robert Bruggner
Stanford University

Michael Linderman
Karen Sachs
Nikesh Kotecha
Garry Nolan

Aberrant intracellular signaling plays a key role in numerous lethal diseases related to cellular malfunctions (e.g. cancer). Accordingly, an understanding of cell signaling cascades provides crucial insight into disease mechanism and can play a critical role in patient diagnosis and treatment. To facilitate high-throughput analysis of signaling components, instrumentation technologies such as flow cytometry have emerged that enable high-throughput, simultaneous measurement of intra and extra-cellular molecules of a single cell. Prior work in Bayesian network inference demonstrates the ability to automatically reconstruct signaling cascades from flow cytometry data.

We expand on this work and present here a hardware--accelerated Markov Chain Monte Carlo (MCMC) implementation to learn signaling cascades from single cell data. Additionally, we present the results of utilizing this pipeline to reconstruct T-Cell receptor signaling cascades in Jurkat cells and discuss implementation issues such as discretization of cytometry data, order and graph sampling methods, and execution time in CPU and GPU based architectures. As the number of simultaneous, single--cell measurements continues to increase, automated approaches such as the one described here will play a crucial role in describing aberrant signaling and provide key insights into disease mechanism and potential disease causing populations.

poster

23 Quantitative Unmixing of Surface-Enhanced Raman Scattering Spectra

Chinyere Nwabugwu
Stanford University

K. Kode
D. Van de Sompel
C. Zavaleta
S. Gambhir
David Paik

Purpose: Surface-enhanced Raman scattering (SERS) microscopy is a light scattering technique of vibrational microspectroscopy for the selective detection of specific biomolecules. This technique combines the advantages of Raman imaging and biofunctionalized gold nanoparticles for visualizing and quantifying the distribution of target molecules. It is a useful tool to analyze the molecular composition and structure of a sample on the basis of unique Raman signatures characterized by a series of Stokes shifts with narrow peak widths, which form a unique pattern for different organic molecules. This project involves developing and validating an algorithm to quantitatively unmix the spectra of multiplexed Raman nanoparticles in living subjects.

Materials and Methods: Surface-enhanced Raman scattering nanoparticles were used to demonstrate whole-body Raman imaging, nanoparticle pharmacokinetics, multiplexing, and in vivo tumor targeting, using an optical microscope adapted for small-animal Raman imaging. Three female 8-week-old nude mice were used for all Raman spectroscope studies. The unmixing process was carried out on Raman spectra of 4 nanoparticles injected simultaneously using unmixing algorithm that parameterizes individual Raman peaks and uses Nelder-Mead optimization, accounting for changes in peak position, intensity and width.

Results: Distorting effects such as fluorescence background, peak location and width heterogeneity break the assumptions of linear and instantaneous generative model needed by typical source separation methods. Individual Raman peaks have been modeled to accurately depict the Raman spectra of nanoparticles. Simulated data based upon empirical measurements of peak location, width and height statistics was used to test the accuracy of our previously developed algorithm. Other constrained optimization methods have been explored and tested.

Conclusion: The initial results of this project have shown that many of the similar issues in other Raman spectral unmixing apply to this data where the relatively lower signal to noise ratio in vivo presents an even greater challenge. We are currently working toward optimizing the performance of this quantitative technique for Raman spectra unmixing in living subjects thereby resulting in significant potential for application of Raman spec¬troscopy in the analysis of nanoparticles.

poster

24 Automatic video monitoring system to detect behavioral patterns in neurodegenerative disorders

David R

Aging Research Clinical Center,
VA Hospital, Palo Alto

Mulin E
Romdhane R
Lee J.H
Piano J
Thonnat M
Bremond F
Leroy I
Yesavage J

Introduction

Behavioral disorders, in particular, apathy are important characteristics of Mild Cognitive Impairment (MCI) and early Alzheimer's disease. Rating scales are essential tools for their assessment however these instruments do not fully capture the complexity of behavioral changes. Video processing techniques could therefore be of interest, but currently, standardized assessment procedures are lacking.

Objectives:

(1) To test a standardized clinical scenario for use in a video-monitored assessment session. (2) To demonstrate the feasibility of this technique among MCI participants.

Methods:

Healthy elderly controls (HC) and patients with MCI were recruited and completed a cognitive and behavioral examination included. The experimental room consisted of a 2 video-camera "smartroom" furnished as a home lounge with daily living equipment. Participants performed a 3-part scenario, including: 1) directed activities corresponding to the Short Physical Performance Battery with video-assessed gait parameters, rated by walking speed (S), length of step (L) and cadence; 2) partially-directed activities consisting of a set of ordered activities of daily living, such as using the phone and cooking, rated as the number of activities completed and activity index; 3) un-directed activities for 30 minutes, rated as number and duration of goal-directed activities.

Results

12 participants (8 HC, mean age=78.7±5.5; 4 MCI, mean age= 80.4±6.2) were included. Data were not collected on two. Preliminary results indicated that for the directed activities, compared to HC, those with MCI had: lower S ($p<.05$); L correlated negatively with apathy rating scales.

Conclusion:

The procedures were well-tolerated by the participants. Video processing was able to capture aspects of directed activities, e.g. gait differences. Inclusion of additional subjects is required to confirm the results for partially- and un-directed activities. These will be illustrated by clinical vignettes.

poster

25

Differences in Antipsychotic Adverse Events among Adult, Pediatric, and Geriatric Populations - An Analysis of the FDA Adverse Events Reporting System

Narmadan A. Kumarasamy
Stanford University

Hersh Sagrieya
Yi-Ren Chen
Karthikeyan I. Ponnusamy
Amar Das

Purpose:
In recent years, antipsychotic medications have been increasingly used in pediatric and geriatric populations (1,2). Many of these drugs were approved based on clinical trials in adult patients only. Preliminary studies have shown that the "off-label" use of these drugs in pediatric and geriatric populations may result in adverse events not found in adults (2,3). In this study, we utilized the large-scale FDA Adverse Events Reporting System (AERS) database to look at differences in adverse events from antipsychotics among adult, pediatric, and geriatric populations (4).

Materials and Methods:
We performed a systematic analysis of the FDA AERS database using MySQL by standardizing the database using structured terminologies and ontologies. We compared adverse event profiles of atypical versus typical antipsychotic medications among adult (18-65), pediatric (age < 18), and geriatric (>65) populations. Data was analyzed using descriptive and comparative statistics (Wilcoxon rank sum).

Results:
We found statistically significant differences between the number of adverse events in pediatric vs. adults with 7 antipsychotics, and between adults vs. geriatrics with 10 antipsychotics ($P < 0.05$). We also found statistically significant differences between the number of adverse events for atypical and typical antipsychotics within each population comparison ($P < 0.05$). Furthermore, the types of adverse events (e.g. metabolic or neurological) reported also varied statistically significantly between each population ($P < 0.05$).

Conclusion:
Antipsychotic medications are commonly prescribed in the United States, and are increasingly used in pediatric and geriatric populations for which the drugs were never specifically tested. Our analysis of the FDA AERS database shows that there are significant differences in the number and type of adverse events among age groups and between atypical and typical antipsychotics for each age group. It is important for clinicians to be mindful of these differences when prescribing antipsychotics, especially if they are used off-label.

References:
1. Olfson M, Blanco C, Liu L, Moreno C, Laje G. National trends in the outpatient treatment of children and adolescents with antipsychotic drugs. Arch Gen Psychiatry 2006 Jun;63(6):679-85.
2. Alexopoulos GS, Streim J, Carpenter D, Docherty JP. Using antipsychotic agents in older patients. J Clin Psychiatry 2004;65[suppl 2]:1-105.
3. Credibility crisis in pediatric psychiatry. Nat Neurosci 2008 Sep;11(9):983.
4. Rodriguez EM, Staffa JA, Graham DJ. The role of databases in drug postmarketing surveillance. Pharmacoepidemiol Drug Saf 2001 Aug-Sep;10(5):407-10.

poster

26 Identifying Tumor-Specific Mutations in Acute Myelogenous Leukemia

Kris Weber
University of Washington

Junfeng Wang
William Noble
Mirela Andronescu
Jay Shendure
Alexandra MacKenzie
Anthony Blau

Purpose:
Identification and analysis of tumor-specific mutations in Acute Myelogenous Leukemia (AML).

Materials and Methods:
We perform whole exome sequencing on paired diagnosis and remission marrow specimens from two patients with Acute Myelogenous Leukemia. Paired-end reads are sequenced on an Illumina Genome Analyzer IIe to an average haploid coverage of 90x.
The reads are aligned to reference genome hg19 with Burrows-Wheeler Aligner (BWA) and realigned around indels with the Genome Analysis Toolkit (GATK). PCR duplicates and ambiguous mappings are discarded. Mutations are identified with VarScan, and SeattleSeq is used to annotate them and compare them to known variants. The output from SeattleSeq and VarScan is parsed into categories, corresponding to noncoding, synonymous and non-synonymous coding.

Results:
Preliminary results based on samples from a single patient reveal the following mutation counts (p-value < 0.001 after Bonferroni adjustment):
29174 mutations in total, including 28271 SNVs and 903 indels.
5423 tumor-specific somatic mutations, including 457 novel mutations, of which 144 are novel non-synonymous mutations in coding regions.
We are still in the early stages of examining individual somatic mutations. Interesting results so far include the identification of novel nonsense mutations in the following genes:
RGS5 – regulates G protein signaling. Can promote apoptosis in endothelial cells.
ADORA3 – adenosine 3 receptor. Can trigger apoptosis in lymphoid leukemias.
CARD10 – caspase recruitment domain family, member 10. Involved in apoptosis.
None of these genes were identified in previous sequencing studies of AML(1,2).

Conclusions:
These preliminary results suggest that the quantity of tumor-specific mutations in AML is much larger than previously reported (144 novel non-synonymous coding mutations versus eight(1) or ten(2). This discrepancy is likely due to the deeper coverage obtained here, which facilitates the capture of lower-frequency mutations and ameliorates problems arising from the mixture of tumor cells with normal cells. Furthermore, the presence of nonsense mutations in three functionally relevant genes not previously associated with AML suggests the existence of previously unexplored genetic pathways to this disease. Future work will include analysis of copy number variation and identification of probable driver mutations.

References:
1.Ley TJ, Mardis ER, Ding L, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 2008;456:66-72
2.Mardis ER, Ding L, Dooling DJ, et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. N Engl J Med 2009;361:1058-1066

poster

27

Reversed ORFeome: an evolutionary crossroad between recombination and gene regulation

Jia-Ren Lin

Stanford University

PURPOSE:

Alternative open reading frame has been discovered for its important function in gene regulation and genetic diversity. However, the open reading frame in the antisense strand (herein: revered ORF), even though it has been found in viral genome and prokaryotes, the existence and function of reversed ORF has not yet been studies in eukaryotes. Here we perform a complete analysis of the reversed ORF in different organisms and study potential application of reversed ORFeome in evolution and gene regulation.

MATERIALS AND METHODS:

Six different RefSeq mRNA sets (Human, Mouse, Rat, Cow, Xenopus, Zebrafish) have been downloaded from NCBI database. The complete reversed ORFeomes of these organisms were generated by rORFinder. The amino acid composition of forward and reversed ORFeomes has been compared and clustered across different species. The specific enrichment of gene ontology of these reversed ORF containing genes were analyzed by PANTHER (http://www.pantherdb.org/panther/). The reversed ORFs of unusual length were also analyzed by BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

RESULTS:

The potential reversed ORFeomes have been identified from lower eukaryotes to human. The length ratio between forward and reversed ORFs exhibits a high correlation to genomic complexity. This finding indicates that the reversed ORFs are tightly regulated by evolutionary constrains. In addition, the amino acid composition of forward and reveresd ORFeomes have distinct signatures, which could be used to identify novel reversed ORFs. Meanwhile, by examining the gene ontology of reversed ORF containing genes, we discovered that proteins involved in receptor/channel signaling or in enzymatic regulation are highly enriched. Finally, by searching for homology of these revered ORF, we found a potential explanation of the origin of these antisense-RNA encoded genes.

CONCLUSIONS

Our results indicate that reversed ORFs could be originated from recombination events throughout evolution. Meanwhile, our findings also suggest that reversed ORFeomes could be used to identify novel regulatory elements of gene, ex: non-coding RNA. Finally, the present of large amounts of revered ORFs within higher eukaryotes could reveal a new source of genetic diversity.

poster

28

A Model for Reasoning about Interaction with Users in Hospital Decision Scenarios

Hyunggu Jung
Stanford University

Robin Cohen

Purpose
Our research aims to develop a model that can be used for scenarios where an agent is reasoning about which human users to enlist to perform decision making, in an environment where decisions need to be made under critical time constraints and where the parameters that serve to model the human users are changing dynamically, to a significant extent.

Materials and Methods
We offer a hybrid transfer of control strategy that takes as its starting point the model of Cheng [1], which includes reasoning about interaction (partial transfers of control or PTOCs) as well as about full transfers of control of the decision making (FTOCs) to another entity. We then project our model into the hospital decision scenarios, reasoning about how to find the right person, at the right time, to assist with the care of patients who are arriving at an emergency room, by effectively modeling the doctors in the environment.
We conduct simulations of hospital settings that reason about which doctors to ask to attend to patients, based on our proposed model. In the setting of our validation simulating hospital emergency scenarios, there are four entities on the entity list and five patients on the waiting list. Every patient has a task criticality for the specific medical problem and the task criticality of each patient is changed dynamically as time passes. Our simulation first selects the patient whose task criticality is highest among those of patients.
We then obtain a strategy chain by calculating formulae reflecting our model with information of each patient. After choosing an entity in the chain, we ask him/her to treat the current patient and update the task criticality of patients who have been treated by entities, as well as, there remaining on the waiting list. When there are no more patients on the waiting list, we finally count the number of problem patients.

Results
Our simulations demonstrate valuable improvements introduced due to the modeling of bother. By comparing the number of problem patients simulated by our model with bother cost and without bother cost, there have been more problem patients with the version without bother cost than one with bother cost.

Conclusions
Our research provides insights into how the dynamic nature of the hospital decision scenarios can be considered as part of the determination of the most effective transfer-of-control strategies.

Reference
[1] Cheng, M., Cohen, R.: A hybrid transfer of control model for adjustable autonomy multiagentsystems. In: Proceedings of AAMAS'05. (2005)
[2] Scerri, P., Pynadath, D., Tambe, M.: Why the elf acted autonomously: Towards a theory of adjustable autonomy. In: Proceedings of AAMAS'02. (2002)

poster

29 Novel statistics reveal cancer universal microRNA activity

Roy Navon
Agilent Technologies

Hui Wang
Israel Steinfeld
Anya Tsalenko
Amir Ben-Dor
Zohar Yakhini

Purpose
microRNAs (miRNAs) regulate genes and play important roles in cancer pathogenesis and development. Variation amongst individuals is a significant confounding factor in miRNA (or other) expression studies. The true character of biologically or clinically meaningful differential expression can be obscured by inter-patient variation. In this study we identified miRNAs with consistent differential expression in multiple tumor types using a matched sample experiment design and a novel data analysis approach.

Materials and Methods
We will present data from microarray profiling of more than 700 miRNAs in 28 matched (same patient) tumor/normal samples from 8 different tumor types (breast, colon, liver, lung, lymphoma, ovary, prostate and testis) – a design that minimizes tissue type and patient related variability. We will then describe novel statistical methods used in analyzing this matched data. These methods are also applicable in various other contexts.

Results and Conclusions
The analysis revealed several miRNA that are consistently differentially expressed over multiple tumor types. These differentially expressed miRNAs include known oncomiRs as well as miRNAs that were not previously universally associated with cancer, such as miR-133b and miR-486-5p, both consistently down regulated in cancer, in the context of our cohort.

References
1. Navon R. et al - Novel rank-based statistical methods reveal microRNAs with differential expression in multiple cancer types. - PLoS One. 2009 Nov 25;4(11):e8003.

poster

30

A Systems Biology Method for Predicting Drug-Gene Interactions in Saccharomyces cerevisiae

Guy Haskin Fernald
Stanford University

Drug targets are often incompletely characterized and most drugs are thought to bind multiple unknown targets, leading in some cases to effective treatment and in other cases to harmful or even lethal adverse events. A thorough understanding of drug targets is essential for rational drug development and may elucidate the biological mechanisms behind adverse events. Testing compounds for all of their potential binding partners would elucidate much of the biochemistry behind their effects, but such large-scale binding experiments are intractable. In this paper, I introduce a systems biology method for predicting drug targets in Saccharomyces cerevisiae. By integrating drug-target databases, chemical similarity metrics, and protein-protein interaction databases I was able to develop a classifier for predicting drug targets in Saccharomyces cerevisiae which achieves an overall performance of 0.87 area under the ROC curve with 10-fold cross validation. This method will be a valuable tool in predicting targets for existing and novel compounds and can be extended to better understand drug targets in other model organisms and even in humans.

poster

31 A flexible estimating equations approach for mapping function-valued traits

Hao Xiong
UCSF

Evan H Goulding
Elaine J Carlson
Laurence H Tecott
Charles E McCulloch
and Saunak Sen

In genetic studies, many interesting traits, including growth curves or skeletal shape, have temporal or spatial structure and are better treated as curves or function-valued observations. Current methods for mapping function-valued traits are mostly likelihood-based, requiring specification of the error structure. However, such specification is difficult or impractical in many scenarios. We propose a general functional regression approach based on estimating equations that is robust to misspecification of the covariance structure. Estimation is based on a two-step least squares algorithm, which is fast, and applicable even when the number of time points exceeds the number of samples. It is also flexible due to the general linear functional model; changing the number of covariates does not necessitate a new set of formulas and programs. In addition, many meaningful extensions are straightforward. For example, we can accommodate missing genotype data using multiple imputation, and the algorithm can be trivially parallelized. Simulation studies indicate that the proposed method maintains the target false positive rate under the null hypothesis regardless of covariance structure. However, likelihood-based procedures with a misspecified covariance structure do not necessarily have the desired false positive rate under the null, and may have lower power compared to our method. We illustrated our method and its advantages using circadian mouse activity data.

poster

32

Detecting ligand-binding sites similarity: A geometric-constraint free method using multiple FEATURE micro-environments (mFEATURE)

Tianyun Liu

Stanford University

Russ B. Altman

The co-evolution of substrates/inhibitors and enzymes is expected to result in a partnership between ligands and the binding sites. It is reasonable to assume that sites that bind to similar ligands share similarity in terms of structural and biochemical properties. Therefore, we developed a fast and efficient approach for comparing binding sites based on the properties extracted from the sites.

We make use of a previously developed system, FEATURE, to represent structural and biochemical properties of a site. Given a site, FEATURE calculates multiple microenvironments that capture properties in the local spherical region centered at each residue observed in the site. Between two sites, every two FEATURE microenvironments from different sites are compared in an exhaustive manner. The similarity between two microenvironments is calculated by using an adjusted Tanimoto coefficient that incorporates background frequency. We then align the two sites by searching for the mutual best scored pairs of microenvironments. The method does not constrain geometrics between multiple microenvironments, increasing the degrees of freedom for comparison.

We benchmark the performance of mFEATURE two folds. First, we evaluate its ability to discriminate pairs of sites bound with adenine-containing ligands from those with non-adenine ones. The method of mFEATURE outperforms most existing algorithms. At 95% specificity, mFEATURE identifies 40% positive pairs (Other methods are lower than 30%). Second, we test the ability of mFEATURE in predicting FAD-binding sites on a structural proteome scale. The method achieves an AUC of 0.92. Due to conformational flexibility of FAD, predicting FAD-binding site remains challenging. The novelty of mFEATURE lies in its independence from sequential and geometric constraints. Therefore it predicts correct alignments between the functional groups from two sites that bind to FAD in different conformations.

Given the reliability of mFEATURE, it has been applied to predict kinase targets for multiple targeted drugs in cancer. We focus on the discovery of binding profile similarities across kinases that are not closely related in sequence space. When high binding site similarity is accompanied by low sequence identity, there is the potential for unexpected ligand cross-reactivity. We predict how similar the inhibitor-binding profiles of two protein kinases are likely to be, based on the properties of the residues surrounding known ATP-binding sites. Two of the top predicted pairs of kinases have been validated by the fact that identical inhibitors were observed to bind to the same pairs of kinases, respectively. Other novel yet high-confidence predictions have potential implications for target selection for multiple targeted drugs in cancer.

poster

33

Cooling Design and Simulation of the Frontend for a 1mm^3 Resolution Breast Cancer Dedicated PET Camera

Jinjian Zhai
Stanford University

Derek Innes
Arne Vandenbroucke
Craig Levin

Purpose:
We are prototyping a 1mm^3 resolution Positron emission tomography (PET) system based on lutetium yttrium oxyorthosilicate (LYSO) - position sensitive avalanche photodiode (PSAPD) modules. The frontend of the system is compactly packed, radioactive and thermal regulated. Special frontend needs to be designed to load, shield and cool the elements inside. There are 2304 PSAPDs in a volume of 9cmX16cmX4cm. It is well known that each PSAPD generates 2mW~4mW power under high voltage bias when the panel is in usage. Because of the dimensional constraints from top and front, we can only cool the PET panel from sides. The main objective of this study is to investigate the frontend design to keep it compact in dimension, provide lead shielding and facilitate heat transfer.

Material and Method:
Because the PSAPD is expensive elements and modules are low in production rate. In order to perform the research, we designed and built the cooling system and then used simulation methods to evaluate the design. We first designed and manufactured a tube-embedded liquid heat sink, and combined with a 3.8 mm tungsten shielding layer to form a compact multilayer shielding/cooling structure. An acrylic arm was extended to hold electrical circuits at the back-end. It is attached to aluminum sidewall by stainless steel screws. Simulation was performed for all the layers and structures including holes and screws between sidewall and acrylic arm. Experiments will be done in the future and the results can be compared with simulation of temperature profile over the structure.

Results:
The design of multilayer sidewall is successful to load the shield material and cooling elements. The total cooling power is 0.08W for one layer of half fin. From the simulation of the sidewall/tungsten/acrylic structure, we found that acrylic prohibits heat from back-end of PET panel and the sidewall/tungsten part has less than 0.5degC temperature rise from boundary condition.

Conclusions:
We concluded that the design of sidewall/tungsten/acrylic structure can thermally isolate the front-end and back-end of the PET panel and protect fin from circuit heat exhaust. It can also load shielding material for radioactive protection.

Thank You

Guidance and Help

Blanca Pineda
Russ Altman
Vijay Pande
Larry Fagan
Carolyn Mazenko
Nancy Lennartsson
Scott Delp
Heideh Fattaey
Susan Parker
Carla Shatz
Mary Jeanne Oliva
Betty Cheng

Previous Organizers

David Chen
Sarah Aerni
Robert Bruggner
Samuel Hamner
Jonathan Karr
Linda Liu
Daniel Newburger
Chirag Patel

2010 Organizing Committee

Konrad Karczewski
Rob Tirrell
Keyan Salari
Matt DeMers
Jessica Faruque
Amir Ghazvinian

Platinum Sponsors

Stanford Biomedical Informatics Training Program
Simbios
Bio-X

Gold Sponsors

Genentech
Sandia National Laboratories
Agilent

Silver Sponsors

23andMe
Butte Lab

Startup Sponsors

DNAnexus