bcats2014

biomedical computation at stanford
14th annual symposium

Abstract Book

January 30, 2014

Welcome

to the 14th annual symposium on Biomedical Computation at Stanford (BCATS). The purpose of this student-organized conference is to provide an interdisciplinary forum for students and post-docs to discuss their recent work in computational biology and medicine. BCATS brings together attendees from around the Bay Area and beyond. This year's conference presents cutting-edge research on a diverse set of topics ranging from population genetics to disease biology.

We would like to thank our keynote speakers, student presenters, judges, sponsors, and all 2014 attendees.

BCATS 2014 Organizing Committee

Zoe June Assaf (PhD candidate in Genetics)

Amy Goldberg (PhD candidate in Biology)

Kimberly McManus (PhD candidate in Biology, MS candidate in BMI)

Shaila Musharoff (PhD candidate in Genetics)

Chris VanLang (PhD candidate in Chemical Engineering)

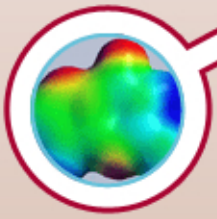
What is Bio-X?

Bio-X is Stanford's pioneering interdisciplinary biosciences program. Bio-X brings together biomedical and life science researchers, clinicians, engineers, physicists, and computational scientists to unlock the secrets of the human body. Rather than study cells and tissues in isolation, Bio-X investigators work to understand the entire organ systems in all their complexity.

BIO-X
STANFORD
UNIVERSITY

Who is affiliated with Bio-X?

Over 600 Stanford faculty from the Schools of Humanities & Sciences, Earth Sciences, Engineering, Law, and Medicine, representing 60+ departments.



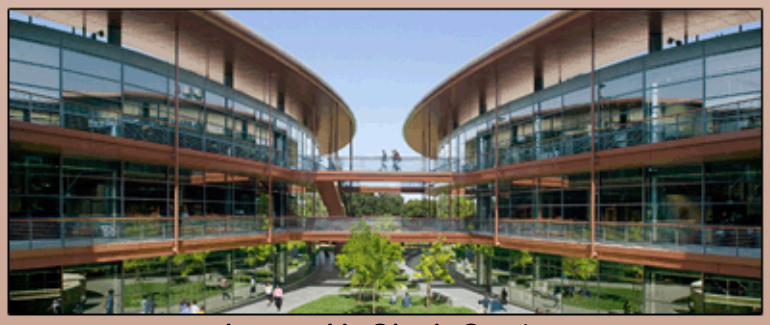
What are Bio-X's programs?

- Seed Grants for Success (Interdisciplinary Initiatives Program)
- Bio-X Stanford Interdisciplinary Graduate Fellowships in Human Health
- Bio-X Graduate Student Fellowships
- Bio-X NeuroVentures Program
- Bio-X Undergraduate Research Awards
- Bio-X Travel Awards

What is the Bio-X Corporate Forum program?

A networking portal for industry to collaborate and create stronger and lasting relationships with Stanford faculty. The union of academic and corporate research further enables innovative discoveries and technological advances. Opportunities for companies include symposia/seminars/mixers, customized technical summits, faculty liaison, and more!

STANFORD
BIO-X



James H. Clark Center

LELAND STANFORD JUNIOR UNIVERSITY
DIE LUFT DER FREIHEIT WEHT
1891

http://biox.stanford.edu

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

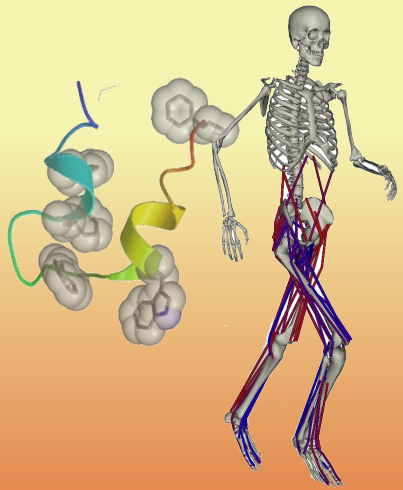
Want to develop, share or find biosimulation software or data?

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

Need open-source, high performance tools to simulate biological structure movement?

Go to www.simtk.org and download:

- ***Simbody*** **for rigid body dynamics**
- ***OpenSim*** **for state-of-the-art musculoskeletal simulations**
- ***OpenMM*** **for accelerated molecular dynamics on GPUs**



Interested in collaborating on important computational biological problems?

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

Running image courtesy of Sam Hamner
NIH Grant R01 GM107340; NIH Roadmap Grant U54 GM072970

Stanford Center for Computational, Evolutionary and Human Genomics

Upcoming Talks

Accurate estimation of heritability in case-control GWAS
David Golan, Tel-Aviv University
Wednesday, February 5, 2014
1:15pm Clark S360
Hosted by Jonathan Pritchard

Decoding autisms using machine intelligence and systems medicine
Dennis Wall, Stanford University
Wednesday, February 12, 2014
1:15pm Clark S360
Hosted by the Wall Lab

Functional genomics of vascular smooth muscle cell differentiation
Olga Sazonova, Stanford University
Wednesday, February 19, 2014
1:15pm Clark S360
Hosted by Stephen Montgomery

Unraveling gene promoter and 3' end effects on expression strength and noise using many designed sequences
Eilon Sharon, Stanford University
Wednesday, February 26, 2014
1:15pm Clark S360
Hosted by Jonathan Pritchard

Circular RNA expression is conserved from humans to the simplest eukaryotes
Julia Salzman, Stanford University
Wednesday, March 5, 2014
1:15pm Clark S360
Hosted by the Salzman Lab

The Stanford Center for Computational, Evolutionary and Human Genomics (CEHG) was founded in 2012 to foster interdisciplinary research at the University. A collaboration between the School of Humanities and Sciences and the School of Medicine, the Center is the intellectual home of 40 professors and over 200 postdoctoral scholars and graduate students. CEHG funding opportunities include the Fellowship Program, Trainee Research Grants, and Faculty Seed Grants.

CEHG researchers at Why We Can't Wait 2013 in San Francisco, CA

Contact Us

cehg.stanford.edu
facebook.com/StanfordCEHG
twitter.com/StanfordCEHG
StanfordCEHG@stanford.edu

Stanford Biomedical Informatics Training Program

The Biomedical Informatics Training Program (BMI) is an interdisciplinary graduate and postdoctoral training program, part of the Biosciences Program at Stanford University's School of Medicine. We offer MS and PhD degrees, and other coursework and research options.

Full-Time Graduate Programs:

- PhD in Biomedical Informatics
- Research MS degree (primarily, but not exclusively, for those with PhD and/or MD)

Part-Time Distance Education Programs:

- Professional/Honors Cooperative Program MS degree
- Certificate in Clinical Research Informatics or Bioinformatics
- Non-degree option for individual classes

For Stanford Students:

- Coterminal MS degree in Biomedical Informatics in addition to BA/BS
- PhD Minor in Biomedical Informatics for Stanford graduate students

For Stanford Medical Students:

- Coursework and research opportunities through Scholarly Concentration and Medical Scholars programs

All degree programs require rigorous coursework in Biomedical Informatics core courses, a coherent set of electives from Computer Science, Statistics, Math and/or and Engineering, and training in Social, Legal and Ethical issues.

For more information, contact:

Student Services Officer
Stanford Biomedical Informatics Training Program
Medical School Office Building, Room X-215
1265 Welch Road, Mail Code: 5479
Stanford, CA 94305-5479

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

Supported by National Library of Medicine Training Grant LM007033

STANFORD CENTER FOR GENOMICS & PERSONALIZED MEDICINE

Thank you to the Stanford Genome Training Program (SGTP) for their financial support

Schedule

8:00 am Registration, badge pickup and breakfast
8:45 am Opening Remarks

Session I (9:00 am - 11:00 am)
9:00 am Keynote Address: John Novembre
Addressing challenges for population genetic inference from next-generation sequencing (pg 10)
10:00 am Scientific Talks
Nandita Garud: *Soft selective sweeps are the primary mode of recent adaptation in Drosophila melanogaster* (pg 17)
Brian Maples: *RFMix: A discriminative modeling approach for rapid and robust local ancestry inference* (pg 19)
Kun-Hsing Yu: *Identifying thyroid carcinoma subtypes and outcomes through gene expression data* (pg 23)
Roshni Cooper: *Exploring the nervous system more effectively with electrical engineering: Employing computer vision to accelerate synaptic research in C. elegans* (pg 16)

11:00 am **Poster Session** (pg 24)

12:00 am Lunch

Session II (1:00 pm - 2:45 pm)
1:00 pm Keynote Address: Katherine S. Pollard
Quantifying taxonomic and functional diversity of microbial communities from shotgun metagenomics data (pg 11)
2:00 pm Scientific Talks
Daniel Himmelstein: *Heterogeneous network link prediction prioritizes GWAS* (pg 18)
Sandeep Venkataram: *Epistasis and the repeatability of evolution in Fisher's geometric model* (pg 22)
Zachary Szpiech: *Long runs of homozygosity are enriched for deleterious variation* (pg 21)
2:45 pm Coffee Break

Session III (3:15 pm - 5:00 pm)
3:15 pm Scientific Talks
Franco Pestilli: *Model-based neuroanatomy: Validation and statistical inference in white-matter connectomes* (pg 20)
Christopher Baldassano: *Supervoxel parcellation of visual cortex connectivity* (pg 14)
Berenice Benayoun: *Buffer domains promote transcriptional stability at key cell identity and function genes* (pg 15)
4:00 pm Keynote Address: David C. Van Essen
The Human Connectome Project: Progress and prospects (pg 12)
5:00 pm Closing Remarks
5:30 pm Dinner with Keynote Speakers (Selected Participants Only)

bcats2014

biomedical computation at stanford
14th annual symposium

Keynote Speakers

Keynote Speaker

Addressing challenges for population genetic inference from next-generation sequencing

---

John Novembre

Associate Professor, University of Chicago, Department of Human Genetics

Next-generation sequencing data open up immense opportunities for discovery in biology. However, the structure of next-generation sequencing data also poses many inherent challenges, especially for population genetic inference. Reads have high error rates, coverage is variable and in many cases low, and when pooling strategies are used the reads can be an unlabeled mixture across multiple individuals. These challenges make accurate individual-level genotype calls difficult. Thankfully, many of these problems can be addressed in the context of model-based inference. In this talk, I will present recent work addressing these challenges. I will demonstrate biases that arise in inferring frequency spectra from low coverage sequencing data and an efficient algorithm that can ameliorate them. I will also show how haplotype frequencies can be inferred from pooled sequencing data derived from experimental evolution experiments with known founders. I will discuss how the same inference algorithm is applicable to the estimation of microbiome community composition. Throughout there will be an emphasis on accuracy and computational efficiency.

About the speaker: The lab of Dr. Novembre focuses on analytical methods to understand the evolutionary forces that shape the genome of non-model organisms, such as Canids and humans. He completed his Ph.D. under Dr. Montgomery Slatkin at UC Berkeley as an HHMI fellow, and moved to the University of Chicago as an NSF Postdoctoral Fellow. Dr. Novembre has also been named as a Searle Scholar. After joining the faculty at the University of California, Los Angeles, he became Associate Professor of Human Genetics at the University of Chicago, where he continues his focus on methods in spatial and population genetics.

Keynote Speaker

Quantifying taxonomic and functional diversity of microbial communities from shotgun metagenomics data

---

Katherine S. Pollard

Associate Professor, University of California San Francisco, Division of Biostatistics
Associate Investigator, University of San Francisco California,Gladstone Institutes

Microbes are a key component of essentially every ecosystem on earth from the human gut to deep sea vents. However, most microbes have never been studied or even sequenced, and many cannot be isolated and investigated in the lab by traditional microbiological methods. Analysis of shotgun sequenced environmental DNA, known as metagenomics, promises insight into the taxonomic and functional composition of microbial communities. The major challenges are that we do not know which sequencing read came from which genome, genomes have very different amounts of coverage in the data due to differences in abundance and genome size, and existing tools for identifying taxa or protein families from sequence data are not optimized for short reads. I will describe several new phylogenetic methods developed in my lab that address these challenges, as well as simulation approaches to statistically evaluate the performance of analysis tools for shotgun metagenomes. The pitfalls of metagenome data analysis and emerging solutions to these problems will be illustrated with examples from our studies of microbial community ecology and the human body.

About the speaker: Dr. Pollard received her M.A. and Ph.D. from UC Berkeley with Dr. Mark van der Laan. Her work focused on statistical methods development for large and complex data sets related to cancer biology. After a post-doc with Dr. Sandrine Dudoit, she became interested in comparative genomics and genetics as an NIH Postdoctoral Scholar at UCSC. Previously faculty at the University of California, Davis, Dr. Pollard is currently an Associate Professor at the University of California, San Francisco, where she develops novel computational and statistical approaches to identify human accelerated regions, and the genetics of regions of ecological or biomedical relevance.

Keynote Speaker

The Human Connectome Project: progress and prospects

David C. Van Essen

Professor, Washington University in St. Louis, Department of Anatomy and Neurobiology

Recent advances in noninvasive neuroimaging have set the stage for the systematic exploration of human brain circuits in health and disease. The Human Connectome Project (HCP) is systematically characterizing brain circuitry, its variability, and its relation to behavior in a population of 1,200 healthy adults (twins and their non-twin siblings). This talk will review progress by the HCP consortium in acquiring, analyzing, and freely sharing these massive and highly informative datasets. The HCP obtains information about structural and functional connectivity using diffusion MRI and resting-state fMRI, respectively. Additional modalities include task-evoked fMRI and MEG, plus extensive behavioral testing and genotyping. Each of these methods is powerful, yet faces significant technical limitations that are important to characterize and be mindful of when interpreting neuroimaging data. Advanced visualization and analysis methods developed by the HCP enable characterization of brain circuits in individuals and group averages at high spatial resolution and at the level of functionally distinct brain parcels and brain networks. Comparisons across subjects are beginning to reveal aspects of brain circuitry that are related to particular behavioral capacities and which are heritable or related to specific genetic variants. Data from the HCP is being made freely available to the neuroscience community via a user-friendly informatics platform. Altogether, the HCP is providing invaluable information about the healthy human brain and its variability.

References:

Van Essen DC et al.,(2013) The WU-Minn Human Connectome Project: an Overview. Neuroimage 80:62-79.

Smith SM et al., (2013) Functional connectomics from resting-state fMRI. Trends Cogn Sci. 17:666-82

Van Essen DC and Ugurbil K (2012) The future of the human connectome. NeuroImage 62: 1299-1310.

About the speaker: Department at Washington University in St. Louis. Along with Kamil Ugurbil, he is Principal Investigator of the Human Connectome Project, a 30 million dollar NIH grant to map brain circuitry in a large population of healthy adults using cutting-edge neuroimaging methods. Van Essens physiological and anatomical studies of macaque visual cortex provide many insights into functional specialization within this distributed hierarchical system. He has pioneered the use of surface-based atlases for visualizing and analyzing cortical structure, function, development, and connectivity and for making comparisons across studies and across species. His tension-based theory of morphogenesis accounts for how and why the cortex gets its folds. His studies of human cerebral cortex provide insights regarding normal variability, abnormalities in specific diseases, and patterns of cortical development. He has served as Editor-in-Chief of the Journal of Neuroscience, founding chair of the OHBM, and President of the Society for Neuroscience. He is a fellow of the AAAS and has received the Raven Lifetime Achievement Award from the St. Louis Academy of Sciences and the Krieg Cortical Discoverer Award from the Cajal Club.

bcats2014

biomedical computation at stanford
14th annual symposium

Scientific Talks

Scientific Talk

Supervoxel parcellation of visual cortex connectivity

Christopher Baldassano
Additional authors: Diane M. Beck, Li Fei-Fei

New large-scale studies using fMRI and dMRI have begun to reveal the fine-scale functional and anatomical connectome of the human brain. We have developed a new approach for understanding these massive datasets, allowing us to discover and visualize how connectivity changes over the entire cortical surface. Given a matrix describing the functional or anatomical connectivity strength between each pair of voxels, we apply a nonparametric clustering algorithm based on the distance-dependent Chinese Restaurant Process (ddCRP) in order to group voxels with similar connectivity properties into spatially contiguous supervoxels. Our method is hypothesis-free, requires no specification of seed voxels, and produces a true parcellation of the brain into spatially-connected subregions rather than ignoring location information. We first validate the clustering method by dividing the Parahippocampal Place Area (PPA) into two subregions based on functional connectivity properties, matching previous work. We then cluster the 59,412-voxel whole-brain group functional and anatomical dataset from the Human Connectome Project, producing a $\sim$200 supervoxel parcellation which provides a compact summary of the full connectivity matrix. For example, resting-state connectivity clustering in early visual cortex divides peripheral V1 from the foveal confluence of V1-V4, revealing that these regions have different functional connectivity patterns with other occipital regions (which match anatomical connectivity differences from probabilistic tractography). In addition to aiding in the discovery of more fine-grained connectivity patterns (allowing us to move beyond a localizer approach to region discovery), the learned parcellation is a general-purpose atlas that can be used to aid in other experiments such as whole-brain decoding. We plan to publicly release the connectivity atlas using an interactive 3D browser-based visualization tool, which will allow anyone to explore the rich connectivity structure of the brain.

Scientific Talk

Buffer Domains promote transcriptional stability at key cell identity and function genes

---

Berenice Benayoun

Additional authors: Elizabeth A. Pollina, Duygu Ucar, Salah Mahmoudi, Edith D. Wong, Kalpana Karra, Elena Mancini, Benjamin C. Hitz, Keerthana Devarajan, Rakhi Gupta, Thomas A. Rando, Julie C. Baker, Michael P. Snyder, J. Michael Cherry, Anne Brunet

Trimethylation of Histone H3 at Lysine 4 (H3K4me3) is a chromatin modification traditionally known to mark transcriptional start sites. Here we demonstrate that H3K4me3 domains that spread more broadly over gene bodies preferentially mark the genes that are essential for cell identity and function across species. The broadest H3K4me3 domains are able to predict novel regulators of neural progenitor cells. Remarkably, the broadest H3K4me3 domains are not associated with increased transcriptional levels; instead, these domains are associated with transcriptional stability in response to a fluctuating environment. Integrative models reveal that the broadest H3K4me3 domains represent a distinct biochemical entity, defined by specific sets of transcription and chromatin regulators. The knock-down of Wdr5, one component of H3K4me3 methyltransferase complexes, leads to a preferential shortening of the broadest H3K4me3 domains. Shortening H3K4me3 breadth in neural progenitor cells increases the transcriptional variability of the novel regulators we identified in these cells, suggesting a causative relationship between H3K4me3 breadth and transcriptional stability. We propose that broad H3K4me3 domains ensure the stable expression of key cell identity genes, and we have termed them Buffer Domains.

Scientific Talk

Exploring the nervous system more effectively with electrical engineering: Employing computer vision to accelerate synaptic research in C. elegans

---

Roshni Cooper
Additional authors: Kang Shen

Synapses are the basic functional unit of the nervous system. They enable neurons to communicate with other cells in the body by trafficking organelles called synaptic vesicles. As fluorescently-tagged proteins are used to visualize increasingly complex aspects of synapses, computer vision can improve both the efficiency and efficacy of synaptic research by automating the imaging of synapses and revealing information otherwise inaccessible to the human eye. This talk highlights two applications of computer vision to the study of synapses in the nematode C. elegans.

The first application analyzes the intermittent motion of synaptic vesicles. This motion can be visualized in C. elegans by tagging the organelles with fluorescent proteins. Kymographs, or images of intensity in position versus time, are created by imaging the tagged neurons. This talk presents an algorithm for quantifying vesicle flux, speed, and duration of vesicle motion. This automated system is more efficient and accurate than the manual processing currently carried out by biologists. In contrast with previous work, this algorithm has been designed to handle biological variations and poor SNR.

Another application of computer vision to neuroscience is for accelerating forward genetic screens. In wild-type C. elegans, two motorneurons (DA8 and DA9) are located adjacent to each other without overlapping, but in synaptic tiling mutants, they do intersect. Hence, muscles may receive conflicting neurotransmitters, ultimately hampering movement. Understanding the genes underlying neuronal development can help explain neuromuscular function and diseases. The software system presented in this talk is designed for the genetic screening of C. elegans based on colocalization of DA8 and DA9. The software isolates the synapses using methods such as clustering and filtering, and then classifies worms as wild-type or mutant by constructing a support vector machine (SVM). Mutant worms can then be studied further by biologists.

Scientific Talk

Soft selective sweeps are the primary mode of recent adaptation in Drosophila melanogaster

---

Nandita Garud
Additional authors: Philipp Messer, Erkan Buzbas, Dmitri Petrov

Adaptation is typically thought to proceed by the rapid increase in frequency and ultimate fixation of a single adaptive allele. This process results in the signature of a hard sweep, specified by one haplotype bearing the adaptive allele present at high frequency in the population.

However, not all modes of adaptation necessarily lead to the presence of a single common haplotype. For instance, in some cases, adaptation might involve subtle changes in frequency at a large number of sites, leaving no signatures of selective sweeps. In other cases, adaptation might drive multiple haplotypes to high frequency, generating signatures of soft sweeps. Soft sweeps can occur when adaptation involves standing genetic variation, where the adaptive allele was already present in the population prior to the onset of positive selection, or when multiple de novo adaptive mutations arise in the population independently on different haplotypes and sweep through the population simultaneously.

We developed a haplotype statistic (H12) that identifies hard and soft selective sweeps with similar power in population genomic data and a second statistic (H2/H1) that can determine whether a given sweep identified with H12 was hard or soft. We used these statistics to carry out a genome scan for adaptation in D. melanogaster sequenced by DGRP. We found evidence of pervasive haplotype structure suggestive of abundant, recent, and strong adaptation in this population. Interestingly, when we applied our H2/H1 statistic to the 50 most prominent H12 peaks, we rejected the hard sweep hypothesis in every case. On the other hand, the vast majority of the peaks are compatible with a simple model of soft sweeps from multiple de novo mutations. We conclude that recent adaptation in D. melanogaster has led primarily to soft sweeps either because it utilized standing variation or because short-term effective population sizes are extremely large.

Scientific Talk

Heterogeneous network link prediction prioritizes GWAS

Daniel Himmelstein
Additional authors: Sergio Baranzini

Genome Wide Association Studies (GWAS) remain the preeminent strategy for identifying disease-associated variants. However, small effect sizes and multiple comparisons limit the pace of discovery. Intelligent prioritization approaches afford an increase in study power while avoiding the constraints related to expanded sampling. We develop a prioritization method that integrates information across multiple relevant domains by predicting relationships from a network with multiple entity and relationship types.

We created a network of 19,088 genes, 16 tissue types, and 128 complex diseases. Using publicly-available resources, we connected these entities with six different relationship types. Disease-gene relationships indicate reported associations from published GWAS studies. By adapting an existing algorithm from social network analysis, PathPredict, we provide a framework for computing the probability that a gene-disease pair represents a true association. The method calculates the prevalence of different network topologies connecting a gene and disease. A logistic ridge regression model identifies influential topological features and enables prediction of gene-disease pairs with an unknown association status.

Our method successfully recalled known GWAS associations with a majority of disease-specific AUCs exceeding 0.76. For multiple sclerosis (MS), we discover 8 novel susceptibility genes, 5 of which (JAK2, REL, IRF1, CD48, and IKZF3) achieve Bonferroni validation on a 9,772-case GWAS masked from our analysis. Regions containing three of these genes were uncovered in a recent MS ImmunoChip-based study highlighting our ability to identify the causal gene within a locis region of indetermination. In addition to prioritizing susceptibility genes, our method enables comparison of the informativeness of each included network component. For identifying disease-associated genes, we find the diseasome the most informative data source followed by functional gene relationships, semantic disease similarity, protein protein interactions, and finally tissue localization.

Scientific Talk

RFMix: a discriminative modeling approach for rapid and robust local ancestry inference

---

Brian Maples
Additional authors: Carlos Bustamante

Local ancestry inference is an important step in the genetic analysis of fully sequenced human genomes. Current methods can only detect continental-level ancestry (i.e., European vs. African vs. Asian) accurately even when using millions of markers. Here, we present RFMix, a powerful discriminative modeling approach that is faster (∼30X) and more accurate than existing methods. We accomplish this by using a conditional random field (CRF) parameterized by random forests trained on reference panels. RFMix is capable of learning from the admixed samples themselves to boost performance and autocorrect phasing errors. RFMix shows high sensitivity and specificity in simulated Hispanic/Latinos, African Americans, and admixed Europeans, Africans, and Asians. Finally, we demonstrate that African Americans in HapMap contain modest (but non-zero) levels of Native American ancestry (∼0.4 percent).

Scientific Talk

Model-based neuroanatomy: Validation and statistical inference in white-matter connectomes

---

Franco Pestilli

Additional authors: J. Yeatman, A. Rokem, K. Kay, H. Takemura, B. Wandell

Magnetic resonance diffusion imaging and computational tractography are the only technologies that enable neuroscientists to measure white matter in the living human brain. In the decade since their development, these technologies revolutionized our understanding of the importance of the human white-matter for health and disease.

Prior to these technologies, the white matter was thought of as a passive cabling system. But modern measurements show that white matter axons and glia respond to experience and that the tissue properties of the white matter are transformed during development and following training. The white matter pathways comprise a set of active wires and the responses and properties of these wires predict human cognitive and emotional abilities in health and disease. We can now predict confidently that to crack the neural code in mapping the human brain, neuroscientists will have to develop an account of the connections and tissue properties of these active wires. Whereas there are many impressive findings, it is widely agreed that there is an urgent need to keep developing and improving tractography methods. The need for a systematic approach to tractography validation and for a framework to perform statistical model testing can be seen in recent reports in Science that set out to characterize human white matter structure.

I will present new methods to perform both tractography validation and statistical hypotheses testing on the network of brain connections. These new methods improve current techniques in fundamental ways and can be applied to any type of diffusion data. I will show that by using the methods we were able to identify a major white-matter pathway communicating information between the dorsal and ventral visual streams, the Vertical Occipital Fasciculus (VOF). This pathway is large and its organization suggests that the human ventral and dorsal visual streams communicate substantial information through areas V3A/B and hV4/VO-1. We suggest that the VOF is crucial for transmitting signals between regions that encode object properties including form, identity and color information and regions that map spatial location to action plans.

Scientific Talk

Long runs of homozygosity are enriched for deleterious variation

---

Zachary Szpiech

Additional authors: Jishu Xu, Trevor J. Pemberton, Weiping Peng, Sebastian Zoellner, Noah A. Rosenberg, Jun Z. Li

Exome sequencing offers the potential to study the population-genomic variables that underlie patterns of deleterious variation. Runs of homozygosity (ROH) are long stretches of consecutive homozygous genotypes probably reflecting segments shared identically by descent as the result of processes such as consanguinity, population size reduction, and natural selection. The relationship between ROH and patterns of predicted deleterious variation can provide insight into the way in which these processes contribute to the maintenance of deleterious variants. Here, we use exome sequencing to examine ROH in relation to the distribution of deleterious variation in 27 individuals of varying levels of apparent inbreeding from 6 human populations. A significantly greater fraction of all genome-wide predicted damaging homozygotes fall in ROH than would be expected from the corresponding fraction of nondamaging homozygotes in ROH ($p < 0.001$). This pattern is strongest for long ROH ($p < 0.05$). ROH, and especially long ROH, harbor disproportionately more deleterious homozygotes than would be expected on the basis of the total ROH coverage of the genome and the genomic distribution of nondamaging homozygotes. The results accord with a hypothesis that recent inbreeding, which generates long ROH, enables rare deleterious variants to exist in homozygous form. Thus, just as inbreeding can elevate the occurrence of rare recessive diseases that represent homozygotes for strongly deleterious mutations, inbreeding magnifies the occurrence of mildly deleterious variants as well.

Scientific Talk

Epistasis and the repeatability of evolution in Fisher's geometric model

---

Sandeep Venkataram
Additional authors: Diamantis Sellis, Dmitri A. Petrov

Steven J Gould's classic thought experiment contemplated how life would evolve differently if we "replayed the tape of life" from a common starting point. There have been multiple methodologies developed to study the repeatability of evolution. The first methodology, which we call forecasting, studies repeatability using convergent evolution in natural systems and parallel laboratory evolutions. These studies test forward predictability- the similarity of evolution at the genetic and phenotypic levels with a fixed initial genotype but no constraints on the final adapted state. The second methodology, which we call retrocasting, tests backward predictability- how likely different adaptive walks are when both the starting point and ending genotype are fixed. It is unclear how forward and backward predictability are related to each other, as they have never been studied in the same system. In addition, theory states that sign epistasis plays a major role in determining the repeatability of adaptive walks by modifying the underlying fitness landscape. We utilize Fisher's classic Geometric Model to study epistasis and repeatability in a single-peaked, smooth fitness landscape.

Recent work has shown that diploid populations in the geometric model frequently accumulate overdominant mutations resulting in balanced polymorphisms, which are never observed in haploids. We therefore test the effect of ploidy on epistasis and forward and backward predictability in the adaptive walks generated by our simulations. We find that overdominant mutations strongly increase the levels of sign epistasis observed adaptive walks. However, there is little difference between haploids and diploids in the frequency of sign epistasis. We observe that diploid simulations are less forward predictable but more backward predictable than haploid simulations, suggesting that these metrics are not equivalent. Finally, we observe that the presence of extinct lineages in diploid adaptive walks can cause incorrect inference of backward predictability in diploids.

Scientific Talk

Identifying thyroid carcinoma subtypes and outcomes through gene expression data

---

Kun-Hsing Yu

Additional authors: Wei Wang, Chung-Yu Wang, Michael Snyder

Unlike most cancers, thyroid cancer has an ever-increasing incidence rate over recent years. In order to better understand its molecular mechanisms, we acquired gene expression data from The Cancer Genome Atlas (TCGA), leveraged supervised machine learning methods to predict stages and outcomes, and utilized unsupervised machine learning methods to gain biomedical insights on the gene activity patterns. Results showed that support vector machine with Gaussian kernel could distinguish patients with different survival outcomes with 81% accuracy, and factor analysis identified important biological processes in the tumor development. With continuing effort to improve classification accuracy, we envision personalizing patient treatments based on their predicted disease outcomes with larger sample size, thereby increasing the quality of care and reducing the cost of cancer management.

bcats2014

biomedical computation at stanford
14th annual symposium

Posters

Poster	Presenter	Title	Page
22	Subarna Sinha	Wilms' Tumor 1 mutation is a driver of DNA methylation in Acute Myeloid Leukemia	48
23	Erika Strandberg	A comparison of data completion methods for predicting hospital acquired deep vein thrombosis from medical record data.	49
24	Timothy Sweeney	Disease specific expression analysis reduces GBM molecular subgroups to two robust clusters	50
25	Federica Torri	Graphically driven analysis and interpretation of RNAseq in the cloud	51
26	Julia Udell	Association of mismatched polymorphic amino acid residues in HLA with chronic rejection of kidney transplants.	52
27	Maya Varma	Parallel algorithm for the DNA sequence alignment problem in a Raspberry Pi Cluster	53
28	Charles Zheng	Resolving multiple fibers in a single voxel: Application of continuous basis pursuit to diffusion MRI data	54

*Please note that these posters will not longer be presenting: Frederica Torri (poster #25)

Poster No.

1 **Performance of methods used to identify coevolving residues between proteins**

Aram Avila-Herrera
Additional authors: Katie Pollard

Co-evolving residues between proteins potentially reveal the importance of proteins in a complex. Few methods for analyzing inter-protein residue co-evolution have been implemented, however many methods used to identify coevolving residues within proteins can be adapted for an inter-protein analysis. We developed tools to compare the performance of 9 such methods on randomly sampled sub-alignments of the bacterial two-component system that vary in the number of sequences, phylogenetic diversity, and conservation.

Poster No.

2 Predicting evolutionary histories of cancer

Tracy Ballinger

Cancer arises when the genome of a normal, healthy cell evolves into a genome with key tumor suppressor genes incapacitated and oncogenes, or tumor drivers, made over-active. This evolutionary process involves single nucleotide mutations as well as structural rearrangements of the genome such as duplications, deletions, and inversions, and many current efforts in sequencing analysis focus on detecting these mutations. However, because of the high mutation rate of cancer cells, many mutations in a late stage tumor are merely passenger mutations which are inconsequential to the disease. A complete mutational analysis would segregate these passenger mutations from the driver mutations, and we do this by reconstructing the evolutionary history of the tumor. Although the evolutionary history of the tumor cannot be precisely known, we have developed a method for generating parsimonious orderings of mutation events. Using copy number changes and structural rearrangements, the Copy Number Ancestral Variation Graph (CN-AVG), samples possible sets of rearrangement events which explain how the mutated tumor genome arose from the normal. From these potential evolutionary histories, we can predict which mutation events occurred early in the formation of the tumor and are therefore likely to be driving events, versus those that occurred later and are probably passenger mutations.

Poster No.

3 **Fragment Mixing: Improved protein design through coupled backbone and side chain flexibility**

Kyle Barlow
Additional authors: Tanja Kortemme

Fragments, data structures that typically store phi and psi torsion angles for short stretches of protein backbones, have been used successfully within the Rosetta macromolecular modeling suite for protocols such as ab-initio prediction and loop modeling. In these protocols, a library of fragments is assembled from known protein structures. Fragments are inserted into a target structure by setting the backbone torsion degrees of freedom to those stored in the fragment. I am expanding this method by now creating fragments that store all protein degrees of freedom. This implicitly creates plausible coupled side chain/backbone moves, which can be difficult to generate using other methods. I am currently benchmarking my method by predicting the peptide sequence specificity of mutated PDZ domains, and by predicting contact networks using fragments derived from room temperature crystal structures.

Poster No.

4 Improving differential diagnostic accuracy of pathologically similar dermatological conditions

---

Jason Bouhenguel

The erythemato-squamous dermatological diseases share many clinical and histopathological attributes making their differential diagnosis a challenging task for physicians. In this paper, we develop a classification algorithm based on previous clinical data to assist dermatologists in differentiating among these conditions. The final embodiment of the model was an optimized SVM with a Guassian kernel combined with principal component analysis (PCA), achieving a maximum classification accuracy of 97.8 percent, sensitivity of 97.77 percent, and specificity of 99.55 percent. The model was optimized using LOOCV and feature selection was performed on subsets of the data (clinical and histological) to understand the relative contribution of these feature subsets. Removing certain features improved the classifier accuracy while others only had a minor effect, suggesting a potential for increased diagnosis efficiency, minimizing invasive procedures, and ultimately patient cost reduction.

Poster No.

5 SLO Triage: a software tool for rapid assessment of Scanning Laser Ophthalmoscope data sets

---

Arlie Capps

Additional authors: Robert J. Zawadzki,John S. Werner, Bernd Hamann

We present SLO Triage, a software tool to evaluate a Scanning Laser Ophthalmoscope (SLO) data set for its suitability for further analysis, at interactive speed. Using SLO, time series of 2D images of the human retina can be acquired conveniently and non-invasively. High-resolution SLO shows retinal microstructure such as the photoreceptor mosaic and captures the eye's transverse motion over the duration of the scan. Because we acquire the images as non-invasively as possible, SLO data reflects blinks and saccades, or sudden jerks in gaze direction, which can interfere with motion stabilization. For safety's sake, SLO data are captured using very low laser light levels, so the resulting images contain laser speckle and high levels of noise. Blinks, saccades, high noise, incorrect focus, and lack of image detail (for example, a missing photoreceptor mosaic in a patient with geographic atrophy) are all factors which can cause errors in an in-depth motion study, which can take five to seven minutes. Implemented as a plugin to ImageJ, SLO Triage quickly provides a repeatable, quantitative preliminary screening of the suitability of an SLO data set for further study, allowing researchers to concentrate on the most interesting and useful data sets.

SLO Triage is concerned not with tracking the motion of the eye, but with detecting whether saccades, blinks and other unfavorable conditions have occurred. This task is much simpler than the full motion study and allows SLO Triage to run much faster. In order to come up with a score for an SLO data set, the SLO Triage software measures the cross correlation of a small central patch from each frame to its successor. A single distinct peak with low overall variance indicates a good match, implying that the in-depth study will find good matches across the entire image pair. SLO Triage also marks frames which are significantly darker than the average brightness as blinks. The combination of few (or no) blinks with good central patch matching across most or all of the data set implies that the in-depth motion study will likely succeed. We are currently working on extending SLO Triage to detect other failure modes such as incorrect resonant-scanner artifact preprocessing.

Poster No.

6 **Using statistical methods to develop analysis tools for trend identification in microbial and metabolite communities**

Ellen Casavant

Additional authors: Danielle Kain, Josh Elias, Susan Holmes

* Co-presented by Danielle Kain

With the recent advancements in 16S sequencing, microbial taxonomic data is readily attainable. However, as data becomes easier to collect, the importance in interpreting and analyzing this data accurately becomes paramount to understanding biological significance, like underlying causation for a shift in microbial composition. Using microbial data from 454 sequencing and metabolic data from HPLC targeting, statistics was applied to novel mediums to create a method of analysis. The mediums this statistics was applied to were different diets– vegetarian and omnivore– and different marine locations. Careful decisions in normalizing data, developing principal component analysis graphs, applying sparse canonical correlation analysis, and implementing sparse linear discriminants were necessary to draw confident conclusions upon involvement of different microbes and metabolites to establish each condition. We believe these tools can be applied to data collected from either technique and could help biologists find general correlations and trends in data, which could eventually simplify the identification of biomarkers.

Poster No.

7 **A cross-validation evaluation of sparse fascicle models of diffusion MRI data from the Human Connectome Project.**

Kimberly Chan

Additional authors: Brian A. Wandell, Ariel Rokem

Diffusion-weighted MRI (DWI) is a non-invasive brain imaging technique that allows for characterization of white matter fibers in vivo. In DWI, the signal is made sensitive to the diffusion of water and multiple measurements are taken with the signal intensity sensitized to different directions of diffusion to reveal information about different fiber populations. The sensitivity to diffusion is determined by a set of gradient amplitudes and durations summarized as so-called b-values. Measurements with high diffusion weighting have a high angular resolution and can reveal more information about different fiber populations, but come at the cost of a reduced signal-to-noise ratio. One proposed hypothesis is that including multiple diffusion weightings in models of the data can improve the accuracy by including the advantages of both low b values and high b values. We have previously demonstrated that DWI models that summarize the signal in each voxel as a linear combination of fibers oriented in different directions (sparse fascicle model, or SFM) fit data collected at single b values well. Here, we evaluate the precision and accuracy of SFM in modeling data collected at multiple b values with one fiber orientation distribution function (single fODF). We compare it to the SFM fitted separately to each b-value (multi fODF). This was performed on DWI data collected by the Human Connectome Project, with high spatial (1.5 mm isotropic) and angular (270 directions) resolution, with multiple b-values. We analyzed the SNR of these data and found it to be comparable to data collected at much lower spatial resolution (2 mm isotropic) on a standard scanner. Models of the mean signal were evaluated using cross-validation: a model was fit to a portion of the data, and then used to predict the remaining data. The accuracy of the model was quantified as the error in this prediction. The use of cross-validation precludes the possibility of bias in model selection, due to overfitting. Cross-validation was first used to find the appropriate model for the mean signal. Then, the SFM was fit to the deviations from the mean signal. The reliability of the model was found by evaluating the correlation between the fODFs from each fold of cross-validation. We demonstrate that while the single fODF model and multi fODF model are equally accurate, the multi fODF model is more reliable.

Poster No.

8 Texture-based computational models of biomedical tissue in radiological images: Unveiling the visible

---

Adrien Depeursinge

Additional authors: Camille Kurtz, Christopher F. Beaulieu, Sandy Napel, Daniel L. Rubin

Modern multi-dimensional imaging protocols in radiology yield much more information than the naked eye can appreciate. As a result, errors and variations in interpretations are currently representing the weakest aspect of clinical imaging. Computerized image analysis and management is expected to provide solutions for ensuring the quality of medical image interpretation by yielding exhaustive, comprehensive and reproducible data analysis. It can also reveal more information from medical images than it is currently possible to see with the naked eye, even in a 2D slice.

We build computational models of multi-dimensional morphological properties of biomedical tissue. The Riesz transform and support vector machines are used to learn the organization of image scales and directions that is specific to a given biomedical tissue type. The obtained models can be steered analytically to enable rotation-covariant image analysis. While most rotation-invariant approaches discard precious information about image directions, rotation-covariant analysis enables to model the local organization of image directions independently from their global orientation. Although already being important in 2D, this becomes crucial to adequately leverage directional image information in 3D data.

The proposed approach is first evaluated and compared to other state-of-the-art methods using collections of natural textures. Experimental evaluation reveals average classification accuracies in the range of 97 to 98 percent for the Outex-TC-00010, the Outex-TC-00012, and the Contrib-TC-00000 suite for even orders of the Riesz transform, and suggests high robustness to changes in global image orientation and illumination. In a second step, we demonstrate the ability of the framework to model biomedical tissue. Computational models of descriptive terms about liver lesions seen in CT slices are built. The models are used to predict the presence of 18 terms with an average area under the ROC curve of 0.853. The distances between all models are calculated to establish a nonhierarchical computationallyderived ontology containing inter-term synonymy and complementarity. It was found to be complementary to the RadLex ontology, and constitutes a potential method to link the image content to visual semantics.

Overall, the proposed computational models were able to fit a wide range of textures and tissue structures. They can be trained to segment, classify and retrieve desired tissue categories. We are currently extending the framework to 3D, which is expected to fully leverage the wealth of modern radiology images.

Poster No.

9 **Deciphering the lexicon of cis-regulatory elements in the spatiotemporal control of gene expression in plants**

Shahram Emami

Additional authors: Rui Wu, Muh-ching Yee, Jos R. Dinneny

We seek to gain a systematic understanding of the role that cis-regulatory elements (CREs) play in orchestrating the regulation of spatiotemporal gene expression in the Arabidopsis root tissue. We have assembled a pipeline for discovery and validation of CREs and their corresponding network of transcription factors (TFs) using bioinformatics, synthetic biology and genomics tools and the existing high-resolution root expression data. Based on the data gathered, a computational predictive model of the effects of the validated CREs on the gene expression will be generated, which will: 1) advance our understanding of the mechanisms controlling gene expression in multicellular organisms 2) be used to generate synthetic promoters inducing gene expression to suit a particular application.

Poster No.

10 Kinematic sampling characterizes RNA dynamics in solution

Rasmus Fonseca

Additional authors: Dimitar Videv Pachov, Julie Bernauer, Henry van den Bedem

The cellular functions of biomolecules such as RNAs and proteins are mediated by the way they twist, open, or close to interact with each other. Structurally characterizing biomolecular conformational dynamics can provide mechanistic insights to exploit the structure-dynamics-function relationship in RNA engineering and design. We have created a kinematic representation of RNA based on theory from robotics to model the dynamic properties of biomolecules. By using inverse kinematics our system, KGSrna, can efficiently perform large conformational motions without breaking distance constraints such as hydrogen- or disulphide bonds while maintaining ideal geometry. We evaluated our method on a benchmark set by generating thousands of samples near a native state and show that distributions of structural properties agree with those from experimental sources. In addition, KGSrna conformational sampling can accurately predict residual dipolar coupling (RDC) experimental data, which is a sensitive reporter of the amplitude of motions on the sub-millisecond time-scale. We structurally characterized a ground state and previously hidden excited state from HIV-1 Trans-Activation Response (HIV1-TAR) RDC data, and offer insight for the transition paths between these two states.

Poster No.

11 **Identifying changes in combinatorial transcription factor interactions across stages of cardiomyocyte differentiation using co-occurrence network dynamics.**

Tara Friedrich

Precise temporal expression of genes is essential for cellular progression during different stages of development. The activation of a particular class of cis-regulatory elements called enhancers helps fine tune gene regulation. Enhancers are composed of multiple transcription factor binding sites that can be found distal to promoters. Characterizing the functional combinations of transcription factor binding site motifs present within enhancers will help elucidate how these regions regulate their targets. We hypothesize that interactions between different sets of specific transcription factors characterize and potentially drive cellular development from pluripotency to more committed states. We use gene expression and epigenetic information from four stages characterized during mouse cardiomyocyte differentiation. With this, we predict expressed transcription factor binding sites within enhancers for each stage. We then build stage-specific transcription factor co-occurrence networks to identify transcription factors that combinatorially regulate genes. Confidence in network edges, or interactions between transcription factors, are updated as ChIP-seq and mass spectrometry data becomes available. We can then characterize how these networks change over the course of differentiation. By observing transcription factor co-occurrence network dynamics, we infer combinations of transcription factors that potentially bind different sets of enhancers consecutively to drive cells to diverse fates. We then ask how well these combinations of transcription factors explain expression variation of nearby genes. To further investigate how these combinations regulate expression, we will attempt to test the functionality of these combinatorial regulatory interactions using a quantitative high-throughput reporter assay. This assay will allow us to quantify the importance of predicted combinations of transcription factor binding sites to the enhancer activity.

Poster No.

12 **Machine-learning estimation of prenatal exposure to polybrominated diphenyl ethers (PBDEs) and dichlorodiphenyltrichloroethane (DDT) from levels in maternal and child blood 9 years after delivery**

Fraser Gaspar

The prenatal and early life periods are of critical importance for human development and prenatal exposure to endocrine disrupting compounds like polybrominated diphenyl ethers (PBDEs), a class of flame retardant, and 1,1,1-trichloro-2,2-bis(4-chlorophenyl)ethane (DDT), an insecticide, have been shown to have long-term health effects. However, the costs and logistics of measuring prenatal exposures in a large number of children and following them throughout development are often prohibitive. To develop predictive models of prenatal exposure, PBDE and DDT were analyzed in blood collected prenatally and 9 years after birth in mothers and children (n=89). Using blood levels and demographic characteristics, SuperLearner and Deletion/Substitution/Addition (DSA) algorithms, both loss-based cross validation methods, were used to systematically select the best predictive model. SuperLearner showed the strongest prediction ability, with correlation coefficients ranging from 0.88-0.99 when using both maternal and child blood levels. Child blood levels at 9 years were shown to have considerably less ability to predict prenatal levels than maternal blood levels at 9 years. Overall findings show that machine-learning back-extrapolation techniques can strongly predict prenatal exposure of persistent compounds like PBDEs and DDT using blood and demographic data collected after birth.

Poster No.

13 Natural stimuli acquire basic-level advantage in object-selective cortex

---

Marius Catalin Iordan

Additional authors: Marius Ctlin Iordan, Michelle R. Greene, Diane M. Beck, Li Fei-Fei

Most object categories elicit a distributed representation throughout visual cortex, however, little is known about the principles behind this representation. Behaviorally, of all the taxonomic levels at which we can categorize an object, the basic-level is considered privileged; a mid-level of generality (i.e. cat) is named, learned, and recognized faster than subordinate (i.e. very distinct: Mr. Whiskers, my orange tabby) or superordinate (i.e. very broad: natural object) levels. Eleanor Rosch first put forward this idea in her seminal 1976 work and argued that the basic level is the taxonomic step that maximizes within-category similarity and minimizes between-category similarity in a cognitively useful way. We used this basic- or entry-level perceptual advantage to guide our search for a fine-grained organizational principle for object categories in the brain: we asked whether patterns of activity across visual cortex adhere to this very powerful cognitive principle of maximizing within-group similarity and minimizing between-group similarity, and if so at what level of the object taxonomy is this organization predominant. We conducted an fMRI experiment in which participants were shown 1,024 images from 32 subordinate categories (i.e. chihuahua) grouped into 4 basic categories (dog, plane, flower, shoe) and 2 superordinate categories (natural object and man-made object). Using several multi-voxel pattern analyses, we provide the first evidence that activity patterns in visual cortex simultaneously maximize within-basic-level similarity (cohesiveness) and between-basic-level dissimilarity (distinctiveness), despite the fact that real-world objects do not group by basic-level category in terms of their appearance. This effect is strongest in object-selective cortex, but is also shared by scene- and face-selective areas. Furthermore, the basic-level advantage emerges gradually as we move up the visual cortical hierarchy, suggesting that successive levels in the visual system may be optimizing basic-level categorizations.

Poster No.

14 Multi-scale data integration framework to predict disease outcome

---

Haruka Itakura

Additional authors: Olivier Gevaert

The emerging paradigm of leveraging multidimensional data to foster the discovery of disease subtypes, prognostic indicators, or treatment predictors has generated both heightened optimism and a new set of challenges. Multi-scale datasets expand the opportunity for novel biomarker and knowledge discovery in a systems biology framework by enabling simultaneous evaluation of evidence at multiple levels molecular (genomic, proteomic), cellular, tissue, radiographic, clinical. The oft-embraced assumption is that having diverse data from heterogeneous sources necessarily and automatically leads to more information and development of better prediction models. However, a number of studies have demonstrated the inadequacy of methods that rely on simple concatenation of different datasets as a means for integrating data. There is a dire need for effective integration strategies that intelligently integrate each layer of data from heterogeneous sources. A number of studies have proposed methods for data integration that improves prediction model accuracy, but these have largely focused on integration of two datasets (e.g., clinical + molecular). Much work remains to be done in developing effective strategies for multi-scale integration using more than two datasets, an area that will be the topic of my research. The overarching goal of my study is to develop a systematic data integration and analysis framework that can effectively generate a high-performing classifier when there are more than two heterogeneous datasets available. The framework will then be applied on two independent disease processes for prognostication of outcome.

Poster No.

15 Understanding the relationship between genetic variation and differential DNA methylation

---

Irene Kaplow

Additional authors: Sarah Mah, Yiqi Zhou, Michael S. Kobor, Hunter B. Fraser

Previous studies associating genetic variation with DNA methylation have been limited to a small subset of the CpGs in the genome, thus providing an incomplete picture of the effects of genotype on methylation. In this study, we implemented a novel pooling technique to collect genome-wide bisulfite sequencing data from a pooled sample of sixty Yoruban individuals. Using this new data-set, we identified genetic differences that are associated with near-by differences in DNA methylation, including methylation in locations that had not previously been assayed in more than a few individuals. Future work involves integrating this data-set with existing RNA-seq and DNase hypersensitivity data from these individuals to obtain a more complete picture of the effects of genetic variation on gene regulation.

Poster No.

16 **TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements**

Rajiv McCoy

Despite tremendous advances in high-throughput sequencing technologies and computational methods, de novo assembly of complex eukaryotic genomes remains a challenging task mostly due to the presence of repetitive elements. Transposable elements, a class of dynamic mobile repeats, are particularly problematic due to high sequence identity and high copy number. Current assembly approaches cannot accurately reconstruct and place transposable elements, so they often remain absent from genome assemblies despite their importance for genome structure, function, and evolution. Here, we showcase the ability of high-quality (less than 0.05 percent per base error rate) Illumina TruSeq synthetic long-reads (2-15 Kbp) to facilitate de novo assembly and resolve complex transposable element sequences. We sequenced and assembled the genome of the model organism Drosophila melanogaster (strain yw;cn,bw,sp) achieving an NG50 contig size of 77.9 Kbp and covering 97.2 percent of the reference genome (including heterochromatic chromosome arms). We entirely recover and accurately place more than 80 percent of annotated transposable element sequences identical to the current reference genome. Because transposable elements are a common feature of genomes across the tree of life, TruSeq synthetic long-reads offer a powerful new approach to genome assembly.

Poster No.

17 Transcriptome analysis reveals differential splicing events in IPF lung tissue

---

Tracy Nance

Additional authors: Kevin S. Smith, Vanessa Anaya, Rhea Richardson, Lawrence Ho, Mauro Pala, Sara Mostafavi, Alexis Battle, Carol Feghali-Bostwick, Glenn Rosen, Stephen B. Montgomery

Idiopathic pulmonary fibrosis (IPF) is a complex disease in which a multitude of proteins and networks are disrupted. Interrogation of the transcriptome through RNA sequencing (RNA-Seq) enables the determination of genes whose differential expression is most significant in IPF, as well as the detection of alternative splicing events which are not easily observed with traditional microarray experiments. We sequenced messenger RNA from 8 IPF lung samples and 7 healthy controls on an Illumina HiSeq 2000, and found evidence for substantial differential gene expression and differential splicing. 873 genes were differentially expressed in IPF (FDR less than 5 percent), and 440 unique genes had significant differential splicing events in at least one exonic region (FDR less than 5 percent). In particular, we found a strong signal of differential cassette exon usage in periostin (adjusted pval = 2.06e-09), an extracellular matrix protein whose increased gene-level expression has been associated with IPF and its clinical progression, but for which differential splicing has not been studied in the context of this disease. We confirmed the differential exon usage in periostin by qPCR (Wilcoxon pval = 3.11e-4). Our results suggest that alternative splicing of periostin and other genes may be involved in the pathogenesis of IPF. We have developed an interactive web application which allows users to explore the results of our RNA-Seq experiment, as well as those of two previously published microarray experiments, and we hope that this will serve as a resource for future investigations of gene regulation in IPF.

Poster No.

18 **Estimation and correction for average genome size in the human microbiome**

Stephen Nayfach
Additional authors: Katie Pollard

Next-generation sequencing has the potential to shed light on the functional role of the human microbiome in disease processes through identification of microbial genes differentially abundant between host phenotypes. However, identification of these candidate genes can be confounded by differences in the average size of genomes between microbial communities, leading to increases in both false positives and false negatives. Specifically, a larger average genome size in one phenotype compared to another reduces the expected number of reads per gene family and therefore the estimated relative abundance of that family if genome size is not accounted for in the analysis. However, due to a lack of tools designed for short-read metagenomic data, it has not been possible to characterize and correct for genome size variation in the human microbiome.

We developed a tool to rapidly and accurately estimate the average size of microbial genomes in a microbiome from shotgun sequence data by aligning short reads to a set of 30 universally distributed single copy genes. Using this tool, we show that average genome size varies significantly both within and between body sites in the human microbiome and that average genome size variation can be largely explained by species level taxonomic differences. Lastly, we explore whether normalizing for average genome size is able to reveal a more biologically relevant set of candidate genes associated with inflammatory bowel disease and type-II diabetes in the gut microbiome.

Poster No.

19 Automated catalytic site detection

Jerome Nilmeier

Additional authors: Nilmeier, J.P., Kirshner, D.A., Wong, S.E., Lightsone, F.C.

As a component of a function prediction platform, we present a catalytic site identification procedure. It uses a template-matching algorithm and a scoring procedure that allows for rapid, scalable protein-to-template matching from a catalog of binding sites. We develop the procedure using the Catalytic Site Atlas (CSA) of Thornton. The Catalytic Site Identification web server provides the innovative capability to find structural matches to a user-specified catalytic site among all Protein Data Bank (PDB) proteins very rapidly (in less than a minute). The server also can examine a user-specified protein structure or model to identify structural matches to a library of catalytic sites. Finally, the server provides a database of pre-calculated matches between all PDB proteins and the library of catalytic sites. The database has been used to derive a set of hypothesized novel enzymatic function annotations. In all cases matches and putative binding sites (protein structure and surfaces) can be visualized interactively online. The website can be accessed at http://catsid.llnl.gov/, and online examples will be discussed.

Poster No.

20 Accessing conformational ensembles of protein complexes

Dimitar Pachov

Additional authors: Rasmus Fonseca, Damien Arnol, Julie Bernauer, Henry van den Bedem

Macromolecules often exhibit significant conformational flexibility, especially when they interact with ligands or protein partners to form assemblies. A divide and conquer approach is typically used for structure determination, where structures of individual components (domains or proteins) are determined at high resolution while the entire complex is studied at lower resolution. Typically, diverse experimental techniques are combined in this process. However, there is a lack of tools to fit such large assemblies to experimental data or to explore their conformational landscape while allowing for flexibility of their components and to connect that to structural mechanisms for biomolecular dynamics.

We extended a kinematics-based computational algorithm to include protein complexes and experimental data (kgsX: kino-geometric sampling for X-ray data). kgsX represents a protein complex as a multi-chain kinematic linkage. In our representation, hydrogen bonds form inter- or intra-chain kinematic loops. By using inverse kinematics, kgsX deforms this complex network of interdependent cycles while maintaining ideal geometry for the entire assembly. Our method can sample and characterize conformational sub-states of large complexes and provide insight into transition pathways. We show its applicability to test cases of 2AR:Gs and PCSK9:LDL assemblies. Sampling such large systems in atomic detail by molecular dynamics simulations are prohibitively expensive, but kgsX has a significant performance benefit.

Poster No.

21 Integrated data mining analysis of RB1

Manisha Sapre

Additional authors: Xuan Guo, Yi Pan , William Walthall, Irene Weber

Retinoblastoma (RB1) is an aggressive and sporadic form of childhood intra-ocular (eye) cancer, originating from the retina, and metastasizing to other parts of the brain in later stages. Children between the age group of 0-9 years are most susceptible. Mutations of RB1 gene can be detected early, observing a 70kb locus on the chromosome band 13q14, which gets repeatedly deleted in Retinoblastoma using computational methods. A package of software, algorithms is applied for analyzing mutation data. The total number of mutations per gene is computed, this tally is converted to a score, and then to a significance level. A threshold is chosen to control False Discovery Rate (FDR), genes exceeding this threshold are reported as significantly mutated. Simulation results filter the data. Major advantage of Bioinformatics based computer aided analysis is, well in advance cost effective treatment.

Poster No.

22 Wilms' Tumor 1 mutation is a driver of DNA methylation in Acute Myeloid Leukemia

---

Subarna Sinha

Additional authors: Max Jan, Thomas M. Snyder, M. Ryan Corces-Zimmerman, Paresh Vyas,Irving L. Weissman, Stephen R. Quake, Ravindra Majeti

Acute myeloid leukemia (AML) is associated with widespread deregulation of methylation at gene promoters but the key signalling events responsible for perturbing epigenetic landscapes are poorly understood. Sequencing of purified cell populations in our laboratory and others have revealed mutations in enzymes that modify cytosine (such as TET2 and IDH) not only in bulk leukemia but also in residual non-leukemic stem cells at the time of diagnosis, suggesting that deregulation of DNA methylation is an early step in the evolution of AML (Jan et al, 2012). In order to find additional mutations that may drive methylation, we developed a novel application of Boolean implications (IF-THEN rules) to identify subset and mutual exclusion relationships between the presence of a mutation and methylation. We found that mutation in the Wilms Tumor 1 (WT1) gene strongly linked to CpG hyper-methylation, similar to mutation in IDH2 but acting upon a different gene set. Expression of mutant WT1 protein in AML cells also induced hypermethylation as measured by 450 bead-chip arrays, confirming WT1 mutation to be an active driver of methylation. The Boolean-derived pattern of methylation, gene expression and response to EZH2 inhibition in WT1mut AML was consistent with a differentiation block caused by WT1mut through deregulated silencing of polycomb targets. Boolean implications may be a useful tool to understand mutation-specific epigenetic states in cancer and find leads for therapy.

Max Jan, Thomas M. Snyder, M. Ryan Corces-Zimmerman, Paresh Vyas,Irving L. Weissman, Stephen R. Quake, Ravindra Majeti (2012) Clonal Evolution of Preleukemic Hematopoietic Stem Cells Precedes Human Acute Myeloid Leukemia. Sci Transl Med (4) 149 :149

Poster No.

23 **A comparison of data completion methods for predicting hospital acquired deep vein thrombosis from medical record data.**

Erika Strandberg
Additional authors: Mohsen Bayati

Objective: Electronic medical record data holds a wealth of information on patient health status, but the high percentage of missing data can make predictions difficult. While previous studies have successfully evaluated the recovery of missing values on biological data with up to 30 percent incompleteness, few studies have compared methodology for missing data completion with the objective of optimal predictive performance using highly incomplete medical record data1,2. We look at the performance of different completion methods on improving prediction of hospital-acquired deep vein thrombosis (DVT) from laboratory results and vital measurements of ICU patients. This particular data set has over 80 percent missing values.

Results: We compared mean, k-nearest neighbors (kNN), singular value decomposition (SVD) imputation and matrix completion methods to singular vector regression (SVR) using zero imputation, all optimized for maximal cross-validation area under the receiver operator characteristic curve (AUC). Logistic regression with Lasso penalty was used as a model for prediction. Matrix completion methods performed the best in terms of cross validation AUC and test set receiver operator characteristic curves (ROC); however, matrix completion was by far the most computationally complex method evaluated. For DVT prediction, SVR performs comparably to matrix completion with AUC of 0.912, 1.5 times better than mean imputation and a computation speed 45 times faster than matrix completion.

Conclusions: The choice of method for missing data completion can greatly improve predictions made on medical record data with high levels of missingness.

References:

1. Troyanskaya, O., et.al. (2001) Missing value estimation methods for DNA microarrays. Bioinformatics, 17, 520-525.

2. Waljee, A.K., et al. (2013) Comparison of imputation methods for missing laboratory data in medicine. BMJ Open, 3, e002847.

Poster No.

24 **Disease specific expression analysis reduces GBM molecular subgroups to two robust clusters**

Timothy Sweeney
Additional authors: Olivier Gevaert

Background: Glioblastoma Mulitforme (GBM) is a disease with poor outcome and limited treatment options. Several groups have attempted to classify GBMs based on sample tissue gene expression, but there is no firm consensus on which clusterings are correct, nor the translational/clinical relevance of a tumor belonging to a particular cluster.

Methods: We first analyzed gene expression data from GBM samples in the TCGA database using disease-specific gene analysis (DSGA). We took the DSGA-transformed data and ran multiple clustering algorithms to determine optimal cluster number. Robust clusterings were overlapped, and samples for which there was agreement in cluster assignment were assigned to the resulting novel core clusters. Gene enrichment in the core clusters was analyzed with SAM. Prediction analysis for microarrays was then used to make a prediction score for testing in external datasets. Similarities with previously published clusters and survival outcomes analyses were performed.

Results: Our core clustering method yielded two robust clusters, one of which overlapped with the previously defined proneural clustering of Verhaak et al., and the other of which we termed inflammatory based on gene enrichment analysis. The proneural cluster was shown to have a survival advantage in the TCGA dataset. Proneural cluster was significantly associated with IDH1 mutation and G-CIMP positivity. The PAM prediction model built from the two core clusters was then applied separately to two external datasets, REMBRANDT and Rotterdam; in each case, there was a significant survival advantage to the proneural cluster. Final re-analysis of the TCGA dataset without initial DSGA transform was unable to produce a robust prediction score.

Conclusions: We showed that after DSGA analysis, GBMs robustly cluster into two groups, proneural and inflammatory, and that the proneural group has a survival advantage, with validation in two external datasets. The effect is not present without DSGA transform. Focusing on disease-specific gene expression changes and using a core clustering technique allows for the inherent heterogeneity of GBM samples to yield robust prediction scores.

Poster No.

25 **Graphically driven analysis and interpretation of RNAseq in the cloud**

Federica Torri

Additional authors: Dragan Bajcic, Kate Blair, Nathan Meyvis, Vladimir Mladenovic, Zoran Rilak, Federica Torri, Predrag Zecevic

Massively parallel cDNA sequencing (RNA-Seq) has emerged as the leading technology for transcriptome-wide expression analysis. While protocols for library preparation have advanced and sequencing costs declined; the size and complexity of data processing, downstream analysis, and interpretation present significant challenges to the widespread accurate implementation of RNA-Seq. High-volume studies require the provisioning of compute clusters and efficient pipeline implementations of analytical toolkits in order to cope with the computational scale of next generation sequence data. The Tuxedo suite is a widely adopted analytical toolkit for RNA-Seq analyses; including read alignment, read quantification, differential expression analysis, and visualization components.

We employ a cloud-based platform for the analysis and interactive visualization of RNA-Seq data. The computational components of the Tuxedo Suite are integrated into a graphical user interface and run on the cloud-computing clusters as pipelines. Differential expression analyses with Cuffdiff produce standard QC charts powered by CummeRbund.

Web-era technologies in combination with cloud computing also enable novel paradigms of visualization-based big data exploration. The human eye is an essential tool in the analysis of biological data; our ability to recognize patterns makes visual inspection of graphical representations of large data sets essential to successful data analysis. We present an interactive visualization suite for interpretation and hypothesis-exploration of RNA-Seq based differential expression data. Interactive plots can be explored by search and gene names are revealed on hover with outbound links to Web-based resources included. This suite enables iterative analysis and presentation of comparative transcriptomic data.

Poster No.

26 **Association of mismatched polymorphic amino acid residues in HLA with chronic rejection of kidney transplants.**

Julia Udell

The success of kidney transplantation relies heavily upon the compatibility of the HLA genes between donor and recipient. The high degree of polymorphism among these genes in the population makes finding an ideal match difficult, and also presents challenges to precision in gene sequencing methodology. It is unclear whether specific amino acids or regions are particularly critical in determining histocompatibility. To investigate this possibility, we introduce a novel method of comparing donor and recipient genotypes and explore its efficacy in predicting transplant rejection. Using typing data of the HLA-A, -B, -C, and -DRB1 genes for 580 paired donors and recipients of a kidney transplantation, we developed a method of allele estimation that used population frequency information to estimate amino acid sequence information. Using this sequence data we then apply a biallelic model measuring differences between donor and recipient sequences, and look for associations between mismatched residues and organ rejection.

Poster No.

27 Parallel algorithm for the DNA sequence alignment problem in a Raspberry Pi Cluster

---

Maya Varma

We present a new parallel algorithm for solving the DNA sequence alignment problem in a cluster of cheap Raspberry Pi computers. The Raspberry Pi is a credit card-sized computer developed by the non-profit Raspberry Pi foundation for the purpose of promoting computer science education. This card incorporates a Broadcom 700 MHz processor with a separate graphics processor to drive a display, 512 Mbytes of RAM, two USB ports and an Ethernet port, all for 35 dollars. Because of its low cost, low power consumption, and availability of software development tools, a cluster of Raspberry Pi offers an attractive alternative to conventional computers for the solution of biological computing problems. The limited amount of memory and communication speed, however, are impediments that must be overcome to design efficient algorithms on these systems.

The DNA sequence alignment problem is concerned with the matching of a very large number of short sequences of DNA, called reads, to a long reference genome. The first step in the DNA sequence alignment problem is to determine the locations where each read sequence appears in the human reference genome. We have designed a novel parallel algorithm to determine the positions of read sequences in a reference genome. The algorithm first creates a hash table of the unique base pair sequences in the reference genome. The algorithm distributes the hash table across a cluster of N worker processors; each worker processor also stores approximately 1/N of the reference genome. A master processor hashes incoming read sequences and directs them to one of the worker processors, which then matches the read sequence with the sequence in the reference genome corresponding to the hashed read sequence. The algorithm avoids collisions in hashing the unique sequences in the reference genome by using a minimal perfect hashing scheme.

We have implemented the algorithm in a cluster of 32 Raspberry Pi processors using the Message Passing Interface (MPI) library. We used the reference genome "human-g1k-v37" from the 1000 Genomes Website, which has approximately 3 billion base pairs. We used a 2.6 GHz 64-bit AMD desktop x86 system with 16 Gbytes of RAM for comparing the power, time, and energy requirements with the Raspberry Pi cluster. The results show that the algorithm runs 3 to 5 times faster in the cluster compared to the desktop system. The power consumption of the entire cluster is under 30 Watts, compared to 65 Watts for the AMD processor. The total energy consumed is 7 to 10 times lower for the cluster. Our results show a cluster of cheap processors is a promising alternative to traditional server-based computing for the solution of biological computing problems.

Poster No.

28 **Resolving multiple fibers in a single voxel: Application of continuous basis pursuit to diffusion MRI data**

Charles Zheng

How much do white matter tracts in the human brain ever cross each other? A number of high-profile publications have used diffusion MRI to answer this question, but coming up with wildly differing conclusions. Indeed, it will remain difficult to determine a conclusive answer for the question as long as we lack a proven protocol for accurately estimating the size and direction of fiber populations from raw diffusion-weighted MRI data. We make a critical assessment of regression-based approaches for analyzing single-voxel data, including L1 and L2 penalized methods, and the continuous basis pursuit method originally introduced by Ekanadham and Simoncelli (2013) as a neuronal-spike sorting method. We use simulations to evaluate the performance of each method both in terms of fitting the diffusion MRI signal and in terms of recovering the true white-matter fibers directions. We simulated varying conditions of true fiber populations, number of measured diffusion directions and MRI parameters, data signal-to-noise ratio, and model misspecification.

BCATS 2014

Thank You!

We would like to express our appreciation in particular for Cody Montana Sam at CEHG for being our administrative support.

Previous Organizers
Anne Mai
Trevor Martin
Jillynne Quinn
Sandeep Venkataram

Platinum Sponsors
Bio-X
Simbios
Department of Biomedical Informatics (BMI)
Stanford Genome Training Program (SGTP)
Stanford Center for Computational, Evolutionary and Human Genomics (CEHG)

Silver Sponsors
Counsyl
Stanford Biosciences Student Association (SBSA)

And the Stanford University Department of Developmental Biology for the kind use of their poster boards!

BIO-X
STANFORD UNIVERSITY

simbios

STANFORD CENTER FOR GENOMICS & PERSONALIZED MEDICINE

Stanford Center for Computational, Evolutionary and Human Genomics

bmi

Counsyl

SBSA STANFORD BIOSCIENCES STUDENT ASSOCIATION