Computational Aspects of Biological Information 2010

Microsoft Research New England
Thursday December 9, 2010
First Floor Conference Center
One Memorial Drive, Cambridge, MA

Workshop Program & Speakers

08:00-08:55	Breakfast
08:55-09:00	Opening Remarks
09:00-09:40	**The Future of Connectomics** Jeff Lichtman, Harvard University
09:45-10:25	**It's about Time: Rewiring in Regulatory Circuits** Aviv Regev, The Broad Institute
10:25-10:45	Coffee Break
10:45-11:25	**All about Interactions** Jun Liu, Harvard University
11:30-12:00	**Protein Synthesis Errors in Budding Yeast** Allan D Drummond, Harvard University
12:05-12:35	**A Mathematical Framework to Determine the Temporal Sequence of Somatic Genetic Events in Cancer** Franziska Michor, Dana Farber Cancer Institute
12:35-13:30	Lunch Break
13:30-14:00	**Metabolic Interactions in Microbial Ecosystems** Daniel Segre', Boston University
14:05-14:35	**Genomic Signatures for Interpreting Disease Variants** Manolis Kellis, MIT
14:40-15:10	**Computation in Biophysically-Plausible Neural Circuits** Gabriel Kreiman, Harvard University
15:10-15:30	Coffee Break
15:30-16:10	**Linear Algebra and Darwin's Finches** Micheal Brenner, Harvard University
16:15-16:55	**Estimating Ultra-Large Phylogenies and Alignments** Tandy Warnow, University of Texas and Microsoft Research New England
17:00-17:30	**Systems-level Analysis of Regulation and Signaling Dynamics** Edo Airoldi, Harvard University
17:30	Closing Remarks

Speaker Abstracts

The Future of Connectomics
Jeff Lichtman, Harvard University

NA

It's about Time: Rewiring in Regulatory Circuits
Aviv Regev, The Broad Institute

From cells to organisms to species, regulatory systems are constantly changing and evolving. Temporal changes in transcriptional regulation can be realized epigenetically within minutes, mediating responses to environmental stimuli, or genetically over millions of years through slow accumulation of evolutionary events. Recent advances in genomic profiling and manipulation technologies have opened up the way to chart and dissect expression dynamics at unprecedented scope and detail. Here, I present our work on using integrative experimental and computational approaches to reconstruct dynamic changes in regulatory circuits. We follow fast responses in the response of mammalian cells to pathogens, to the slower changes that govern cell differentiation in the blood, and up to the evolution of transcriptional circuits over hundreds of millions of years.

All about Interactions
Jun Liu, Harvard University

In this talk we will report some of our recent efforts in developing methods for detecting various interactions, such SNP-SNP interactions and amino-acid mutation interactions. We consider problems such as Genome-wide association studies (GWAS), eQTL studies, and the HIV mutation study. Typical in these problems, a large number of covariates (such as SNP genotypes of the individuals) are available. It is extremely challenging to discover how a small number of them interact with each other to affect the response (e.g., the disease risk). We propose a Bayesian partition model to tackle the problem. Our extensive simulation studies mimicked the data structures in real applications and demonstrated that our Bayesian approach is much more powerful than the standard two-stage step-wise approach as practiced by statisticians and geneticists. We also used the method to analyze some GWAS and eQTL datasets.

Protein Synthesis Errors in Budding Yeast
Allan D Drummond, Harvard University

Protein synthesis, the conversion of genetic information into phenotypic information by transcription and translation, is strikingly error-prone: errors are believed to occur at a rate of 1 in 1000 to 10000 codons translated. At an error rate of 5 in 10000, one in five typical-length (400-amino-acid) proteins will contain at least one error. Despite their high frequency, protein-level errors have eluded broad-scale measurement for almost 50 years due to purely technical barriers. Using a novel approach, we have quantified protein-synthesis errors proteome-wide in budding yeast. Errors depend on the nucleotide sequence, and occur at rates consistent with previous estimates derived by monitoring single codons in diverse organisms. We discuss the significance of widespread

A Mathematical Framework to Determine the Temporal Sequence of Somatic Genetic Events in Cancer

Franziska Michor, Dana Farber Cancer Institute

Human cancer is caused by the accumulation of genetic alterations in cells. Of special importance are changes that occur early during malignant transformation because they may result in oncogene addiction and represent promising targets for therapeutic intervention. Here we describe a computational approach, called Retracing the Evolutionary Steps in Cancer (RESIC), to deduce the temporal sequence of genetic events during tumorigenesis from cross- sectional genomic data of tumors at their fully transformed stage. When applied to a dataset of 70 advanced colorectal cancers, our algorithm accurately predicts the sequence of APC, KRAS, and TP53 mutations previously defined by analyzing tumors at different stages of colon cancer formation. We further validate the method with glioblastoma and leukemia sample data and then apply it to complex integrated genomics databases, finding that high-level EGFR amplification appears to be a late event in primary glioblastomas. RESIC represents the first evolutionary mathematical approach to identify the temporal sequence of mutations driving tumorigenesis and may be useful to guide the validation of candidate genes emerging from cancer genome surveys.

Metabolic Interactions in Microbial Ecosystems

Daniel Segre', Boston University

Cellular metabolism consists of a complex network of chemical reactions, which provide the cell with a reliable supply of energy and building blocks. Optimization approaches based on steady state approximations have been increasingly successful at predicting how individual microbial species allocate available resources under different environments and perturbations. Possible extensions of current approaches that incorporate interactions between different species can help understand microbial ecosystems, with applications ranging from synthetic ecology to the organization of the human microbiome.

Genomic Signatures for Interpreting Disease Variants

Manolis Kellis, MIT

Our group at MIT aims to further our understanding of the human genome by computational integration of large-scale functional and comparative genomics datasets. (1) Using alignments of multiple closely related species, we have defined evolutionary signatures for the systematic discovery and characterization of diverse classes of functional elements, including protein-coding genes, RNA structures, microRNAs, developmental enhancers, regulatory motifs, and biological networks. (2) Using epigenomics datasets of multiple chromatin marks across the complete genome, we have defined chromatin signatures that reveal numerous classes of promoter, enhancer, transcribed, and repressed regions, each with distinct functional properties. (3) Using diverse functional datasets across many cell types, we have defined multi-cell activity signatures for chromatin states, regulator expression, motif enrichment, and target gene expression, and have used their correlations to link candidate enhancers to their putative target genes, infer cell type-specific activators and repressors, and to predict and validate functional regulator binding in specific chromatin states.

We have used these evolutionary, chromatin, and activity signatures to elucidate the function and regulatory circuitry of the human and fly genomes, to reveal many new insights on animal gene regulation and development, including abundant translational read-through in neuronal proteins, functionality of anti-sense microRNA transcripts, and thousands of novel large intergenic non-coding RNAs. We have also used these signatures to revisit previously uncharacterized disease-associated single-nucleotide polymorphism (SNP) variants linked to several diseases and phenotypes from genome-wide association studies, which has enabled us to provide mechanistic insights into their likely molecular roles. Overall, our genomic signatures dramatically expand the

annotation of the non-coding genome, providing a systematic annotation of chromatin functions, new insights on diverse regulatory mechanisms, and shining new light on previously uncharacterized disease-associated variants.

Computation in Biophysically-Plausible Neural Circuits
Gabriel Kreiman, Harvard University

Our brains routinely solve multiple challenging problems in a fault- tolerant, robust, rapid and almost effortless manner. Understanding the computational principles that govern how densely interconnected, parallel and hierarchical neuronal circuits solve cognitive problems can provide inspiring ideas for the development of engineering algorithms and hardware. Using computational tools and biophysically- plausible modeling can in turn help us untangle the functions of complex biological circuits. I will provide examples of the two-way communication between computation and neuroscience centered around problems in visual recognition and high-level cognition. I will use these examples to emphasize open questions and the multiple opportunities at the interface between computation/mathematics and biology/neuroscience.

Linear Algebra and Darwin's Finches
Micheal Brenner, Harvard University

Abstract: Evolution by natural selection has resulted in a remarkable diversity of organism morphologies that has long fascinated scientists and served to establish the first relations among species. Despite the essential role of morphology as a phenotype of species, there is not yet a formal, mathematical scheme to quantify morphological phenotype and relate it to both the genotype and the underlying developmental genetics. In this talk, I will discuss our recent work (joint with O. Campas, R. Mallarino, A. Herrell and A. Abzhanov) that demonstrates that the morphological diversity in the beaks of Darwin's Finches is quantitatively accounted for by the mathematical group of affine transformations. All beak shapes in Darwin's most famous genus (Geospiza) are related by scaling transformations. Previous work has shown that there are two genes (BMP4 and Calmodulin) whose expression controls beak shape in Geospiza. After summarizing the evidence for these conclusions, I will delve into preliminary investigations as to what it might mean: I will discuss simple models of beak development, and discuss necessary constraints for generating shapes related by affine transformations. I will then discuss our efforts to extend these ideas to a larger phylogeny of birds, in particular trying to identify the rules for changing one beak shape into another.

Estimating Ultra-Large Phylogenies and Alignments
Tandy Warnow, University of Texas and Microsoft Research New England

Phylogenetic (evolutionary) tree estimation is fundamental to biology, and has applications to much biomedical research. Phylogenetic tree estimation is enormously difficult: the best approaches are NP-hard, and many real datasets take weeks or months of analysis. In particular, the estimation of multiple sequence alignment turns out to be one of the most challenging and important steps in a phylogenetic analysis. In this talk I will present several methods that produce greatly improved trees as compared to other phylogeny estimation methods. These methods utilize divide-and-conquer strategies in order to improve the accuracy of traditional phylogeny estimation methods. Among the methods I will present are SATe (Liu et al. 2009, Science Vol 324, no. 5934) and DACTAL (in preparation). SATe simultaneously estimates a tree and a multiple sequence alignment, while DACTAL estimates a tree without ever computing an alignment on the entire set of sequences. Both methods provide great improvements in tree accuracy over other methods, and do so fairly efficiently.

Systems-level Analysis of Regulation and Signaling Dynamics
Edo Airoldi, Harvard University

Abstract: A fundamental systems-level challenge for living organisms is the regulation of cellular activity in a fluctuating environment. While the complete set of regulatory and functional pathways supporting growth and cellular proliferation are not yet known, portions of them are well understood. In particular, cellular proliferation is governed by mechanisms that are highly conserved from unicellular to multicellular organisms, and the disruption of these processes in metazoans is a major factor in the development of cancer. In this talk, I will present new statistical and computational methods to identify quantitative aspects of regulatory mechanisms, and describe recent work on a statistical characterization of cellular proliferation and growth in yeast. We found that the expression levels of a small set of genes accurately predict the instantaneous growth rate of any cellular culture, robust to changing biological conditions, experimental methods, and technological platforms. Our model also predicts growth rates for related yeast species, suggesting that the underlying regulatory signature is conserved across a wide range of unicellular evolution. Most importantly, statistical and computational methods enable substantive biological insights about growth at instantaneous time scales inaccessible by direct experimental methods.