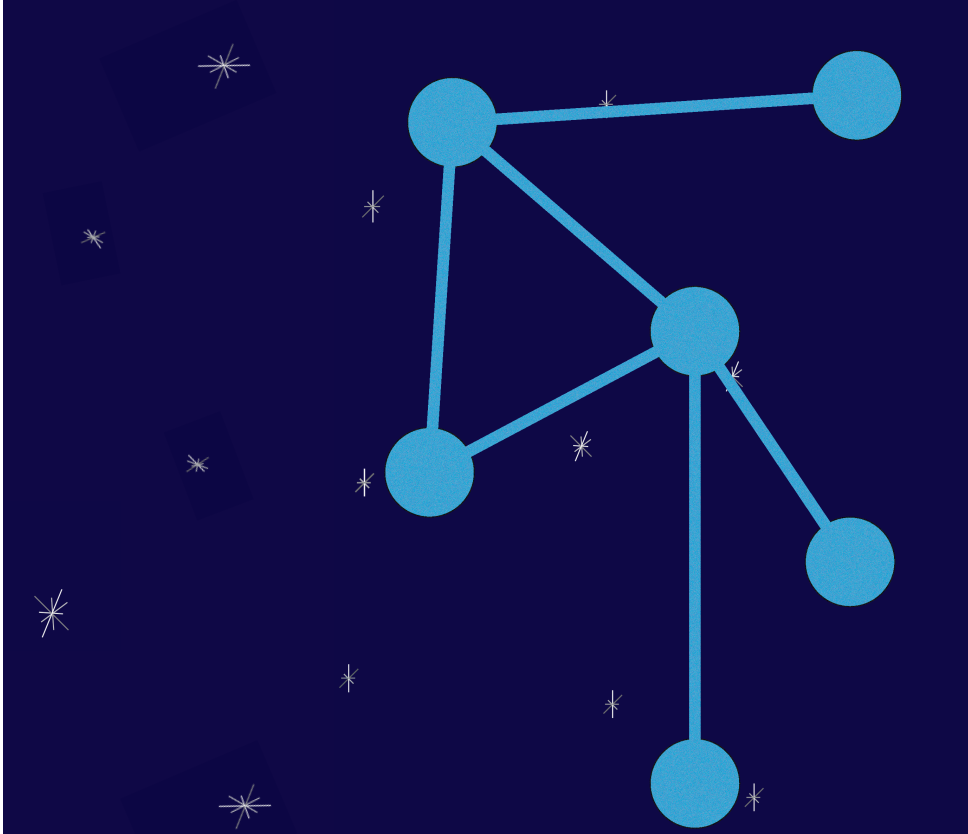
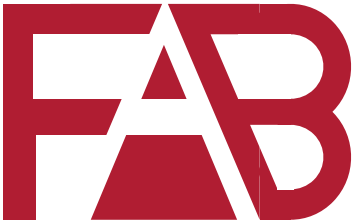


Future of Algorithms in Biology

September 28-29, 2018
Pittsburgh, PA
fab2018.cbd.cmu.edu

Carnegie Mellon University

Workshop on the Future of Algorithms in Biology



September 28-29, 2018

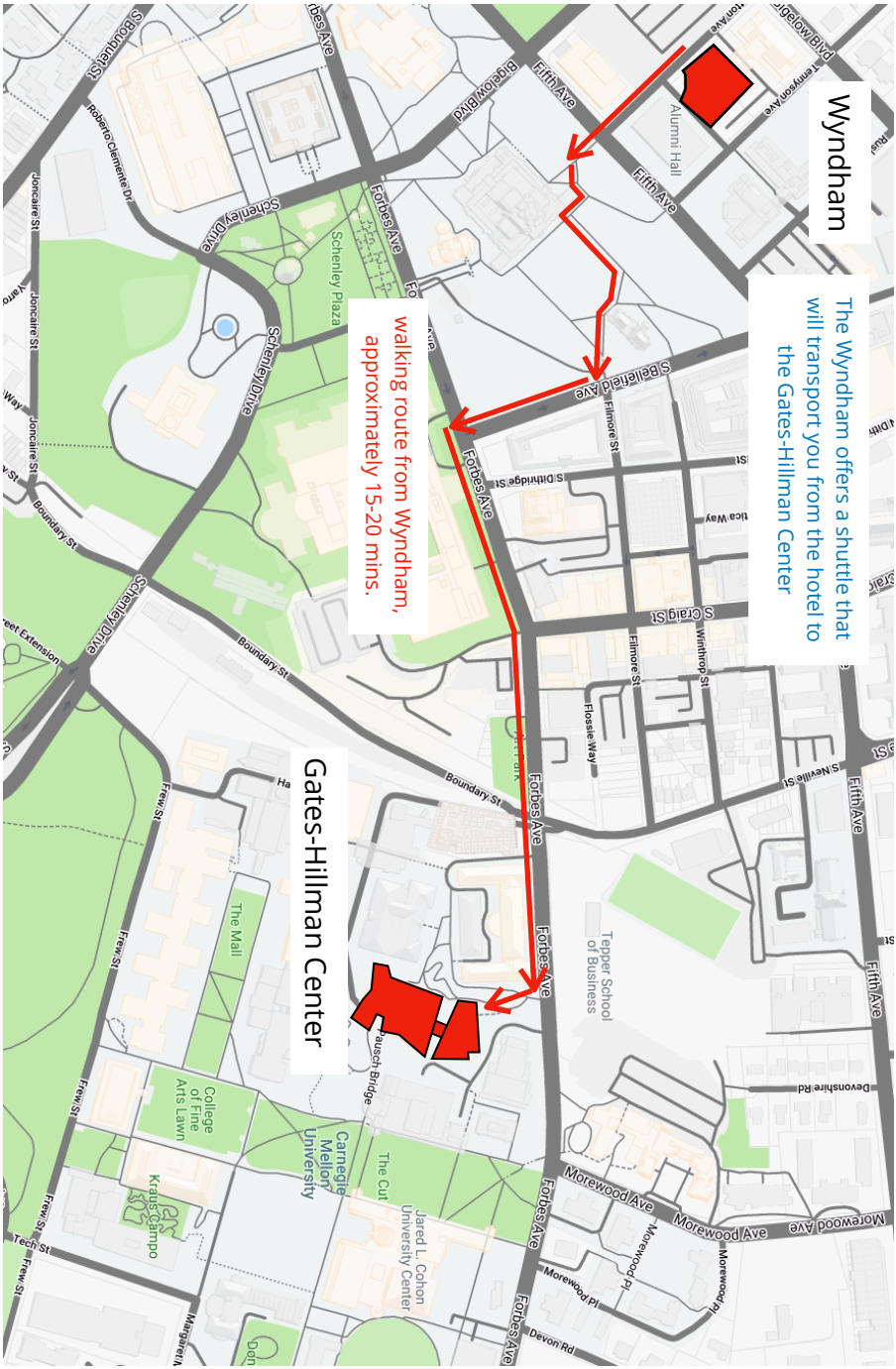
Organizing Committee

Carl Kingsford — Carnegie Mellon University
Tamir Kahveci — University of Florida
Mike Schatz — Johns Hopkins University
Min Xu — Carnegie Mellon University
Dan DeBlasio — Carnegie Mellon University

Made possible with the support of

Grant CCF-1748493

Area Map


Wyndham
The Wyndham offers a shuttle that will transport you from the hotel to the Gates-Hillman Center
walking route from Wyndham, approximately 15-20 mins.
Gates-Hillman Center
Alumni Hall
Fifth Ave
Bigelow Blvd
Forbes Ave
S Bellefield Ave
Schenley Drive
Schenley Plaza
Boundary St
Frew St
The Mall
College of Fine Arts Lawn
Carnegie Mellon University
Kraus Campo
The Cut
Jared L. Cohon University Center
Tepper School of Business
Morewood Ave
Devonshire Rd
Devon Rd
Morewood Pl
Flossie Way
S Craig St
Filmore St
Winthrop St
S Dithridge St
S Neville St
Joncaire St
Roberto Clemente Dr
Tech Trf St
Margaret M

General Information

Conference Photo Release

Photos and videos of attendees will be taken throughout the workshop. These photos may appear on the FAB website, newsletter, conference brochures, social media outlets, or other future FAB promotional material. By virtue of your attendance, you agree to the use of your likeness in such media.

Campus Wifi

Network: CMU-GUEST
Password: 3UDZ2LK9

Social Media

We encourage participants to post about talks (unless the presenter asks otherwise) and continue the constructive discussion on social media using the hashtag:

#FABCMU

Presentation Details

If you will be presenting a talk at FAB, please ensure that your presentation works on the projectors during a break sometime before you talk. Technical staff will be available to help resolve any issues.

The Gates-Hillman Center

All events for FAB will be held on the 6th floor of the Gates-Hillman Center. Signs will be posted for those arriving by foot from Forbes Avenue. Talks and discussions will be in Gates-Hillman Center 6115. The poster session and meals will be held in Gates-Hillman 6121. The opening reception will be held in the Creative Commons (Gates-Hillman 6101).

Please note: this event is being recorded and publicly live streamed. If you would prefer your talk not be streamed and/or posted publicly please tell the organizers in advance. Without such a request, by presenting you are agreeing to allow your talk to be streamed, recorded and posted online.

Thursday, September 27, 2018

6:00 pm	Opening Reception	Gates-Hillman Center Creative Commons (until 8:00 pm)	

Friday, September 28, 2018

			Page
8:00 am	Breakfast		
8:45 am	Opening Remarks		
9:00 am	Trey Ideker	Towards a structure/function simulation of a cancer cell	5
10:00 am	Saket Navlakha	Algorithms: a future language of biology?	11
10:30 am	Coffee Break		
10:45 am	Fabio Vandin	Finding patterns in cancer genomes: algorithms and challenges	12
11:15 am	Ana Ligia Scott	Protein-protein docking, protein-ligand docking and QM/MM calculation considering functional movements: advantages, challenges and problems	13
11:45 am	Jinbo Xu	Fast and accurate ab initio folding by deep learning	14
12:15 pm	Lunch Break		
1:15 pm	Lightning Talks	See pages 23-25 for abstracts	
1:45 pm	Panel Discussion	See page 35 for details	
2:15 pm	Olgica Milenkovic	Higher-order and constrained clustering for genomic data analysis	15
2:45 pm	Laxmi Parida	Algorithms in biology: a shift in paradigm	16
3:15 pm	John Kececioglu	Parameter inference and parameter advising in computational biology	17

Friday, September 28, 2018 (continued)

			Page
3:45 pm	Poster Session	See pages 29-34 for abstracts	
4:45 pm	Nadia El-Mabrouk	Future of algorithms for gene family evolution prediction	7

Saturday, September 29, 2018

			Page
8:00 am	Breakfast		
9:00 am	Zhiping Weng	The temporal landscape of recursive splicing during Pol II transcription elongation in human cells	18
9:30 am	Gregory Johnson & Rory Donovan-Maiye	Building a Babel fish for biology (and making it useful for biologists)	19
10:00 am	Lightning Talks	See pages 26-28 for abstracts	
10:30 am	Coffee Break		
10:45 am	White Paper Preparation		
11:15 am	Christina Boucher	Scalable data structures: challenges and advancements	20
11:45 am	Kathy Tzeng	Challenges in genomics medicine	21
12:15 pm	Lunch Break		
1:15 pm	S. Cenk Sahinalp	Algorithmic approaches for tumor phylogeny reconstruction via integrative use of single cell and bulk sequencing data	22
1:45 pm	Gaudenz Danuser	Inferring causality in molecular pathways from image fluctuations	9
2:45 pm	Closing Remarks		

Towards a structure/function simulation of a cancer cell

Friday
9:00 am
10:00 am

Recently we and other laboratories have launched the Cancer Cell Map Initiative (ccmi.org) and have been building momentum. The goal of the CCMI is to produce a complete map of the gene and protein wiring diagram of a cancer cell. We and others believe this map, currently missing, will be a critical component of any future system to decode a patient's cancer genome. I will describe efforts along several lines:

1. Coalition building. We have made notable progress in building a coalition of institutions to generate the data, as well as to develop the computational methodology required to build and use the maps.
2. Development of technology for mapping gene-gene interactions rapidly using the CRISPR system.
3. Causal network maps connecting DNA mutations (somatic and germline, coding and noncoding) to the cancer events they induce downstream.
4. Development of software and database technology to visualize and store cancer cell maps.
5. A machine learning system for integrating the above data to create multi-scale models of cancer cells.

In a recent paper by Ma et al., we have shown how a hierarchical map of cell structure can be embedded with a deep neural network, so that the model is able to accurately simulate the effect of mutations in genotype on the cellular phenotype.

Keynote

Trey Ideker
University of California, San Diego

Speaker Profile

Dr. Ideker is a Professor of Medicine at UC San Diego. He is the Director of the National Resource for Network Biology, the San Diego Center for Systems Biology, and the Cancer Cell Map Initiative. He is a pioneer in using genome-scale measurements to construct network models of cellular processes and disease. The goal of the Ideker Lab is to develop a new type of medicine based on knowledge of the complete physical and functional wiring of the cell. They are developing new ways of mapping these wiring diagrams directly from genome-scale measurements (genomic, proteomic, and metabolomic) and for using these maps to translate the increasingly complex data gathered from patients to predict disease outcomes and develop better treatments.

Future of algorithms for gene family evolution prediction

Friday
4:45 pm
5:45 pm

Genes are the molecular units of heredity holding the information to build and maintain cells. They are the main target for understanding biological mechanisms, identifying genetic variation and designing appropriate gene therapies. During evolution, they are mutated, duplicated, lost and passed to organisms through speciation or Horizontal Gene Transfer (HGT). Genes originating from the same ancestral copy are called homologs . Homologous genes are orthologs if their parental origin is a speciation, paralogs if it is a duplication and xenologs if it is a HGT. Inferring gene relations is a prerequisite for functional prediction purposes. Tree-based methods consist in reconstructing a phylogenetic tree for the gene family, and then inferring the nature of internal nodes (duplication, speciation or HGT) from reconciliation with a species tree. Accuracy of gene relation inference depend on the type of considered datasets (sequences , trees, gene orders, etc) and on their accuracy. It also depends on the accuracy of the considered evolutionary models and of to prediction tools. For example, reconciliation is based on the assumption that each gene family evolves independently through single gain and loss events. Although this hypothesis holds for genes that are far apart in the genome, it is not appropriate for genes appearing grouped into syntenies, which are more plausibly the result of a concerted evolution.

In this presentation, I will exhibit various algorithmic challenges related to gene tree reconstruction and gene relation inference handling various sources of information, various sources of errors in the datasets, various models of evolution, and exhibit future directions that can be explored to unify the various algorithmic tools into a single robust framework for gene tree reconstruction.

Nadia El-Mabrouk
University of Montreal

Speaker Profile

Nadia El-Mabrouk is full professor at the Computer Science Department and member of the "Centre de Recherche Mathématiques" at the University of Montreal. She has a longstanding experience in developing algorithms for comparative genomics and especially genome rearrangements, gene tree reconstruction and orthology/paralogy relations between genes. She has organized two RECOMB Comparative Genomics Workshops in Montreal. This year, she has been chairing the Population Genomics and Molecular Evolution track at ISMB. She is involved, each, year in the program committee of some of the most prestigious computational biology conferences such as RECOMB, ISMB, WABI or APBC. Her research appears in a variety of computer science, bioinformatics and life science journals, among them IEEE/ACM, Molecular Biology and Evolution, Bioinformatics, Nature Scientific Reports or BMC-Genomics.

Inferring causality in molecular pathways from image fluctuations

Saturday
1:45 pm
2:45 pm

Keynote

One of the major limitations, if not the biggest limitation, in the study of complex molecular pathways is adaptation of the system to experimental perturbation. Here we define 'complex' as pathways with functional redundancy between pathway components and with feedback and feed-forward interactions. While experimental perturbation of one pathway component may lead to a phenotype, it is impossible to interpret the difference between phenotype and wildtype in terms of the function the targeted component fulfills in the unperturbed system. What the phenotype shows is how the system performs without the perturbed component in place. In strict terms the limitation of perturbation approaches apply equally to genetic perturbations, which lead to long-term adaptation, and acute perturbation approaches, which often cause short-term adaptation. Nonetheless, system perturbation is the gold standard in dissecting cause and effect relations in molecular pathways.

Inspired by the field of econometrics, where predictive models are built entirely from passive observation of financial fluctuation time series, my lab is developing a novel mathematical framework to determine causality in molecular pathways. Biologically, we are particularly interested in pathways that regulate cell morphogenesis. These pathways are organized with relations between components that are distributed not only in time but also in space. Accordingly, our development of causal inference has been paralleled by the development of a quantitative imaging workflow to extract meaningful fluctuation series from live cell movies at the appropriate time and length scales.

After ten years of work, we have finally arrived at the point the approach begins to unveil surprising insights of pathway organization. In this overview talk I will take cell protrusion as a prime example of a system of molecular pathways that feature complexity as defined above. I will introduce the mathematical and computational concepts based on which we can now accurately delineate the functional hierarchy between signaling and mechanical processes driving cell protrusion events without a single perturbation experiment; and I will demonstrate validation of the approaches by select experiments.

Gaudenz Danuser
University of Texas Southwestern

Speaker Profile

Gaudenz Danuser is currently appointed as the inaugural chairman of the Lyda Hill Department of Bioinformatics au UT Southwestern Medical Center (UTSW) in Dallas. He also holds the Patrick E. Haggerty Distinguished Chair in Basic Biomedical Science and is a Scholar of the Cancer Prevention Institute of Texas (CPRIT). Before moving to UTSW, Danuser directed research laboratories at ETH Zurich (2002 – 2003), at The Scripps Research Institute in La Jolla (2003 – 2009), and at Harvard Medical School (2009 – 2014).

Trained as an engineer (geodetic and electrical engineering/computer science), he entered the field of cell biology as a postdoctoral fellow in the Program for Architectural Dynamics of Living Cells at the MBL in Woods Hole. Since then, he has focused his research on the question how chemical and mechanical signals integrate in the regulation of cytoskeleton dynamics and membrane trafficking.

Currently, his lab's main interest is focused on understanding the roles shape regulation play in the enhanced proliferation and survival of the metastatic cell, including how shape cues may confer drug resistance. To address these questions the lab develops innovative quantitative imaging methods to experimentally probe these processes and uses mathematical modeling to compile the data in mechanistic systems analyses. He is a devoted teacher in areas of computational cell biology, cellular biophysics, and the theory of measurement applied to cell biology both at the institutional and international level.

Algorithms: a future language of biology?

Friday
10:00 am
10:30 am

Saket Navlakha
Salk Institute for Biological Studies

Biological systems — be it brains or plants — must solve problems to survive, and the strategies these systems have evolved can be viewed as algorithms. I will argue that moving forward, computer scientists should help make algorithms a preferred language for describing problem-solving and decision-making strategies in biology. I will present one example of our work towards this end, on how the olfactory neural circuit in the fruit fly evolved a variant of a common computer science method (called locality-sensitive hashing) to assess odor similarity. I will conclude by describing a few open problems and challenges in pursuing this research direction.

This is joint work with Sanjoy Dasgupta and Charles F. Stevens.

Talk

Finding patterns in cancer genomes: algorithms and challenges

Fabio Vandin
University of Padova

Friday
10:45 am
11:15 am

Sequencing technologies now allow measuring different features of a cancer genome at an outstanding level of detail for an unprecedented number of tumors. The resulting datasets provide an exceptional opportunity to gain insight into cancer development and progression by identifying patterns of mutations across different tumors. However, the identification of reliable patterns poses severe computational and statistical challenges, due to the high dimensionality of the problem. In this talk I will discuss some recent work on the development of algorithms to identify reliable and significant patterns from measurements of a large collection of tumors, including patterns associated with clinical or functional measures, and highlight some of the current and future challenges.

Talk

Protein-protein docking, protein-ligand docking and QM/MM calculation considering functional movements: advantages, challenges and problems

Friday
11:15 am
11:45 am

Ana Ligia Scott
Federal University of ABC (Brazil)
University of Pittsburgh

Considering functional movements and large or local conformational in protein-protein docking, protein-ligand docking and QM/MM calculations is essential and mains a great challenger to the programs and protocol proposed in literature. A efficient a precise protocol to considerer these kind of conformational changes can improved our ability of simulation.

Our group form Federal University of ABC (Brazil) have worked with this kind of problem trying to propose new protocols and applying them to molecular systems involved with diseases. This talk proposed share our experience and discuss: what had been developed nowadays, problems, advantages and challenges.

Talk

Fast and accurate ab initio folding by deep learning

Jinbo Xu
Toyota Technical Institute at Chicago

Friday
11:45 am
12:15 pm

Ab initio folding is one of the most challenging problems in Computational Biology. Recently contact-assisted folding has made some progress on this problem, but it requires accurate inter-residue contact prediction, which by existing co-evolution methods can only be achieved on some proteins with a very large number of sequence homologs. To deal with proteins without many sequence homologs, we have developed a novel deep learning (DL) method for contact prediction by concatenating two deep residual neural networks (ResNet). The first ResNet conducts convolutional transformation of 1-dimensional protein features to capture sequential context of one residue and the second conducts convolutional transformation of 2-dimensional features to exploit higher-order residue correlation and global information. Experimental results suggest that our DL method doubles contact prediction accuracy of pure co-evolution methods on proteins without many sequence homologs and that our predicted contacts can fold many more proteins than ever before while running on a single workstation. Our DL method also works well on membrane proteins and inter-protein contact prediction even if trained by single-chain non-membrane proteins. See [1] and [2] for technical details.

1. S Wang, S Sun, Z Li, R Zhang, and J Xu. *Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model.* 2017. **PLOS Computational Biology.** doi.org/10.1371/journal.pcbi.1005324
2. S Wang, Z Li, Y Yu, and J Xu. *Folding Membrane Proteins by Deep Transfer Learning.* 2017**. Cell Systems.** doi.org/10.1016/j.cels.2017.09.00

Higher-order and constrained clustering for genomic data analysis

Friday
2:15 am
2:45 am

Olgica Milenkovic
University of Illinois, Urbana-Champaign

Clustering is one of the most frequently used machine learning technique in bioinformatics, and it has found wide-spread applications in areas such as single-cell RNA data analysis and reverse engineering of biological networks. Despite the apparent maturity of the field, many interesting challenges have recently emerged in the context of clustering on hypergraphs, clustering with side-information and clustering with must-link and cannot-link side constraints. We will discuss a number of algorithmic solutions for the above clustering paradigms, provide a brief description of their theoretical underpinnings, and present several new applications of these learning tools in genomic and biological data analysis.

This is a joint work with Eli Chen, Pan Li, and Chao Pan.

Talk

Algorithms in biology: a shift in paradigm

Laxmi Parida
IBM T. J. Watson Research Center

Friday
2:45 pm
3:15 pm

Since the turn of this century, the field of biology has been inundated with data, first genomic, then transcriptomic, epigenomic and so on, and now moving towards even single cell omics. Each mode presents its unique challenges. String algorithms with layers of statistical checks and balances have successfully served the class of problems with NGS omic-reads: similarly graph algorithms on "relationship" information and phylogeny algorithms on "evolution" models. At the risk of over-generalizing and over-simplifying, I believe these approaches have been based on concrete models, albeit driven by the data. The results in the last decades are a testimony to the efficacy of these approaches.

However, propelled by the successes in the big data field with ML/AI tools, there is a growing impatience or urgency to make sense of the biological data that is rapidly accumulating. Topological Data Analysis (TDA) takes a refreshingly different view of data that is not necessarily geometric, providing a means to explore multi-way relationships, at arbitrary depths, within the data. Using the example of metagenomic data, I will describe an application of TDA to address the problem of accurate organism identification. This is made possible by an abstract mapping of the underlying problem to a filtration of certain simplicial complexes, on which TDA can be applied as a "black-box".

Parameter inference and parameter advising in computational biology

Friday
3:15 pm
3:45 pm

John Kececioglu
University of Arizona

Many scenarios in computational biology involve an optimization model that is solved to find the best reconstruction of an unobserved phenomenon, such as when assembling a genome sequence, inferring an evolutionary tree, predicting the folded state of a protein or RNA molecule, or computing a multiple sequence alignment. Key to such models is the objective function being optimized, which often has many free parameters that must be tuned. Setting these parameter values is crucial, as they can radically affect the accuracy of the reconstructions output by tools that optimize such models. We discuss two forms of the problem of automatically finding good parameter settings for such models: (1) parameter inference, which finds the best default parameter setting that yields reconstructions of highest average accuracy over a collection of benchmark training examples; and (2) parameter advising, which selects a good parameter choice from a collection of possible settings to yield a reconstruction of highest estimated accuracy for a user's specific input data. We present general and efficient algorithms for parameter inference and parameter advising, and report concrete results for the context of protein multiple sequence alignment.

This is joint work with Dan DeBlasio and Eagu Kim.

The temporal landscape of recursive splicing during Pol II transcription elongation in human cells

Zhiping Weng
University of Massachusetts Medical School

Saturday
9:00 am
9:30 am

Recursive splicing (RS) is an evolutionarily conserved process of removing long introns via multiple steps of splicing. It was first discovered in *Drosophila* and recently proven to occur also in humans. The detailed mechanism of recursive splicing is not well understood, in particular, whether it is kinetically coupled with transcription. To investigate the dynamic process that underlies recursive splicing, we systematically characterized 342 RS sites in three human cell types using published time-series data that monitored synchronized Pol II elongation and nascent RNA production with 4-thiouridine labeling. We found that half of the RS events occurred post-transcriptionally with long delays. For at least 18–47% RS introns, we detected RS junction reads only after detecting canonical splicing junction reads, supporting the notion that these introns were removed by both recursive splicing and canonical splicing. Furthermore, the choice of which splicing mechanism was used showed cell type specificity. Our results suggest that recursive splicing supplements, rather than replaces, canonical splicing for removing long introns.

Talk

Building a Babel fish for biology (and making it useful for biologists)

Saturday
9:30 am
10:00 am

Gregory Johnson & Rory Donovan-Maiye
Allen Institute for Cell Science

Our knowledge of biology is largely determined by the collection of all observations we can take of a system. These measurements take many forms (images, amounts of gene products, protein structures, etc) and can be destructive to the sample, mutually exclusive to each other, noisy, incomplete, expensive, and take years to gather. These measurements co-vary with each other across data types, as well as across samples and conditions.

An open area of research is concerned with how to fuse disparate data types and create unified representations of biological systems (cells, ecosystems, etc). Such a representation could manifest itself as a model which intakes any subset of observations of a particular sample, including experimental conditions and predicts all other possible measurements that could be acquired from that sample with corresponding degrees of certainty.

For example, in cell biology, subcellular organization can be related to gene expression through fluorescence in situ hybridization experiments. If a unified model existed, an image of a tissue sample showing fluorescently labeled nuclei and mitochondria could be provided, and the model would output both the possible loci of all RNAs in the cell, their absolute expression levels and images of the predicted locations of all other subcellular structures. The model would also report the level of certainty and possible ranges of these values. One could imagine that the certainty in some of these predicted outputs would increase if the scientist also reported the cell types and other conditions of the sample. Such a model could be extended to all possible data types (western blots, mass spec, regulatory network state, natural language descriptions, etc) intaking k measurements and outputting n predictions, where k is between 0 and n data types.

Here I will present how such a model could be constructed and it's desired properties, how it relates to existing models in biological and non-biological domains, and an overview of existing barriers for implementation.

Scalable data structures: challenges and advancements

Christina Boucher
University of Florida

Saturday
11:15 am
11:45 am

Burrows Wheeler transform (BWT) redefined how we view sequence alignment; yet, there exists several challenges in applying it elsewhere in bioinformatics. Here, we go over several challenges that exist in it's application and present solutions. The solutions, in effect, are contemporary data structures, such as the generalized compressed suffix array, and the wavelet tree. In this talk, we will go over these data structures and describe their use in biological sequence analysis.

Talk

Challenges in genomics medicine

Saturday
11:45 am
12:15 pm

Kathy Tzeng
Optum, UnitedHealth Group

With the advent of Next Generation Sequencing (NGS) technologies, the cost and time required to sequence human genomes have significantly reduced. Genomics information is increasingly a determinant of clinical diagnosis, prevention and treatment. While genomics medicine is gaining momentum, there are still significant challenges in need of novel algorithms in various areas, from association of genomics data with other data sources to genomics data representation, interpretation and management. Selected areas will be discussed in this presentation.

Talk

Algorithmic approaches for tumor phylogeny reconstruction via integrative use of single cell and bulk sequencing data

S. Cenk Sahinalp
Indiana University

Saturday
1:15 pm
1:45 pm

Recent technological advances in single cell sequencing (SCS) provide high resolution data for studying intra-tumor heterogeneity and tumor evolution. Available computational methods for tumor phylogeny inference via SCS typically aim to identify the most likely perfect phylogeny tree satisfying infinite sites assumption (ISA). However limitations of SCS technologies such as frequent allele dropout or highly variable sequence coverage, commonly result in mutational call errors and prohibit a perfect phylogeny. In addition, ISA violations are commonly observed in tumor phylogenies due to the loss of heterozygosity, deletions and convergent evolution.

In order to address the above limitations, we introduce two combinatorial techniques: the first one is B-SCITE, a MCMC based combinatorial approach that infers trees of tumor evolution from combined single-cell and bulk sequencing data. Using a comprehensive set of simulated data, we show that B-SCITE systematically outperforms existing methods with respect to tree reconstruction accuracy and subclone identification. We note that B-SCITE obtains high fidelity reconstructions even with a modest number of single cells.

Next we introduce PHISCS, a new combinatorial technique that aims to minimize a linear combination of potential false negatives (due to e.g. allele dropout or variance in sequence coverage) and potential false positives (due to e.g. read errors) among mutation calls, as well as the number of mutations that violate ISA - to define the optimal sub-perfect phylogeny. PHISCS ensures that several lineage constraints imposed by the use of variant allele frequencies (VAFs, derived from bulk sequence data) are satisfied. We express our formulation both in the form of an integer linear program (ILP) and - for the first time in the context of tumor phylogeny reconstruction - a boolean constraint satisfaction problem (CSP) and solve them by leveraging state-of-the-art ILP/CSP solvers. Our formulation is the first to integrate SCS and bulk sequencing data under the finite sites model.

The future of data science education: delivering approachable, reusable algorithms and workflows!

Friday
1:15 pm
1:45 pm

Ben Busby
NCBI

Over the past three years, NCBI has run or been involved in 29 data science hackathons! In many of these hackathons, participants assemble into teams of five or six to work collaboratively for three days on pre-scoped projects of general interest to the bioinformatics community. On average, about 80% of teams produce an alpha or beta working prototype, and approximately ten percent ultimately publish a manuscript describing their work. Some of these can be found at biohackathons.github.io. Reusability, user centered design and increasing both ease of use and scientific information content are focused on at these events.

NCBI and other parts of NLM and NIH are also involved in other programs pertaining to project-based data science education. These include the NIH data science mentorship program, the visiting bioinformatician program, and the microbial metagenomics discovery challenge. As the amount of publicly available data funded by NIH increases, workflows from these programs are increasingly brought to bear on these datasets. They will be discussed with an eye toward scaling up biomedical data science training worldwide.

Steering Biological Adaptation Strategically

Tuomas Sandholm
Carnegie Mellon University

Friday
1:15 pm
1:45 pm

Living organisms react to challenges through evolution and adaptation. This has proven to be a key difficulty in developing therapies, since the organisms develop resistance. I proposed the wild idea of steering evolution/adaptation strategically—using automatically generated multi-step steering plans. I will briefly summarize the directions that we are pursuing for generating such plans:

1. Using computational game theory algorithms for solving (typically incomplete-information) multistage game models
2. Deep reinforcement learning, stochastic optimization, and opponent exploitation techniques
3. In the biological context, the opponent (e.g, disease) has a systematic handicap because it evolves myopically. This can be exploited by computing trapping strategies that cause the opponent to evolve into states where it can be handled effectively.

Potential application classes include therapeutics at the population, individual, and molecular levels (drug design), as well as cell repurposing and synthetic biology.

Different parts of this research are joint work with different students and collaborators: Christian Kroer, Gabriele Farina, Jonathan Li, Tyler Lovelace, Penelope Morel, and Jim Faeder.

Algorithmic and machine learning challenges in digital pathology

Friday
1:15 pm
1:45 pm

Aly Azeem Khan
Toyota Technological Institute at Chicago

Deep neural networks (DNNs) have demonstrated a robust and durable approach to solving complex image analysis tasks, especially when specific visual features may be difficult to articulate a priori. The extremely nuanced and fine-grained nature of features that distinguish certain pathological classifications has been a long-standing challenge in visual assessment of histopathology images. While DNNs have already been applied to several complex tasks in digital pathology, these models have often been trained and evaluated on only a specific subset of anatomically-restricted cancers. The ability to effectively use DNNs and make generalized predictions on related but different pathologies remains a hard and open problem for digital pathology.

In this workshop, I will present three open problems for developing generalizable and interpretable models in digital pathology, including using domain adaption, unlabeled side-information in training, and guided backpropagation. I will use examples from industry and clinical pathology to explain how solving these problems could help to improve patient care. In particular, I will highlight an example from cancer immunology where effective use of these models may help predict microsatellite instability, which is a clinically actionable genomic indication for cancer immunotherapies and occurs in a wide range of malignancies, including colorectal, uterine, and gastric cancers.

For more than half a century, manual evaluation of histopathology slides by experienced pathologists has remained the standard for identifying potential clinically actionable pathologic features for different cancers. I will make the case for using emergent deep learning models to predict relevant pathologies directly from hematoxylin and eosin stained histopathology slide images. My particular focus on developing a generalizable and interpretable framework for modeling histopathology slide images holds the potential for substantial clinical impact, including broadening access to clinical testing and augmenting clinical decision-making for pathologists.

Building computational models that connect macroscale anatomy to nanoscale molecular function

Navid Farahani
3Scan, Inc.

Saturday
10:00 am
10:30 am

Studying and modeling networks generally yields powerful features and insights, which can help scientists predict behaviors at a systems level. The trillions of cells present in complex multicellular organisms, like human beings, interact and cooperate from embryonic development through adult life, forming a dynamic cellular network. Cellular functions and interactions are crucial to network modeling, but have been historically studied one at a time, precluding complex computational models.

We will review convergent technological advancements in robotics, computing, sequencing and multi-scale imaging, which are enabling contemporary researchers to extract massive amounts of tissue, cellular, and ultrastructural data from complex biological systems. For example, precise and systemic determination of the relative spatial positions of cells and their transcriptomes is now a reality thanks to rapid progress in the field of spatial transcriptomics. By studying how structure and function interact in dynamic cellular microenvironments, we will inevitably discover novel cellular niches, preferential cellular or proteomic interactions, and more importantly, provide crucial context in our understandings of disease etiology, pathogenesis, and prognosis.

Emerging computational approaches such as multi-view representation learning, image-to-image translation, and hemodynamic modeling demonstrate the potential to bridge the rubicon between macroscale anatomy and nanoscale molecular function. We propose that interdisciplinary collaboration in this space will catalyze the synergistic energy required to build in silico models of biologic processes, potentially yielding new medicinal pathways and identification of novel disease targets.

Alignment using prior information

Saturday
10:00 am
10:30 am

Ben Langmead
Johns Hopkins University

There is established and growing interest in alignment algorithms that are informed by prior information. An example is spliced alignment of RNA-seq reads, where information about known splicing patterns is used to construct a new reference that is "augmented" to include spliced sequence. Another example is the rapidly growing area of graph alignment, where information about genetic variants in the population is used to augment the genome.

While this is commonly done, it's also poorly studied, and several questions have gone unaddressed in published literature. Can we construct alignment algorithms that use prior information in a way that goes beyond presence/absence? Given many possible variants to include, how do we pick the "best" set to include? Is it always better to include more variation, or is there a penalty as well as an advantage to making more variation visible to the algorithm?

I will discuss these issues in the context of both spliced alignment and graph alignment. I will discuss some software that is already published — e.g. Rail-RNA for spliced alignment — but also some preliminary studies — e.g. FORGe for prioritizing genetic variants for graph genomes — as well as some new work.

Open computational problems in HIV epidemiology

Niema Moshiri
University of California, San Diego

Saturday
10:00 am
10:30 am

It is believed that Human Immunodeficiency Virus (HIV) was introduced to the United States in the 1970s. During the 1990s, HIV spread rapidly throughout the United States. Due to advances in medicine, numerous treatments have been developed to suppress HIV in patients (though no cure exists), and due to intervention efforts by epidemiologists via frequent affordable screening and treatment dispersion, the epidemic has become largely contained in the developed world. However, the virus is still rampant in much of the developing world, namely in Sub-Saharan Africa. Due to recent advancements in sequencing, it is now fairly common to sequence viral samples from patients when they receive treatment, and the obtainment of these sequences opens the door for a wide range of computational applications. In my flash talk, I will present some open problems in the field of HIV epidemiology (original biological/epidemiological problems as well as their formal computational problem formulations), and I will discuss various methods that exist that attempt to solve the aforementioned problems. I hope to help define the future of algorithms in biology by presenting an overview of the computational problems that exist in HIV epidemiology so computational researchers will be able to develop novel algorithms to solve these problems in the years to come.

Comparative genomics meets topology: a novel view on genome median and halving problems

Friday
3:45 pm
4:45 pm

Pavel Avdeyev
The George Washington University

Genome median and genome halving are combinatorial optimization problems that aim at reconstructing ancestral genomes by minimizing the number of evolutionary events between them and the genomes of extant species. While these problems have been widely studied in past decades, their known algorithmic solutions are either not efficient or produce biologically inadequate results. These shortcomings have been recently addressed by restricting the problems solution space. We will show that the restricted variants of genome median and halving problems are, in fact, closely related and have a neat interpretation in terms of embedded graphs and polygon gluings. Namely, we introduce the following problem: for a given embedded graph G, find the shortest sequence of surgeries (operations, which cut the surface along two edges of G and glue the resulting four sides in a new order) that results in an graph G' such that it has the maximal possible number of connected components and each of its connected components is embedded into a sphere. Further, we demonstrate its advantages for solving genome median and halving problems in some particular cases.

Poster

Building an automated bioinformatician

Dan DeBlasio
Carnegie Mellon University

Friday
3:45 pm
4:45 pm

Modern scientific programs have large numbers of tunable parameters that impact the output, and in turn the resulting downstream analysis. Finding the correct parameter choice efficiently often requires having background knowledge about the software and the application domain; because of this most users rely on the default parameter choices. These default are set to work well on average, but the most interesting problems are rarely "average".

We have developed a framework called parameter advising [1] to make input-specific parameter choices with very little impact on wall-clock time in most cases. This framework is very general and we have successfully applied it to multiple sequence alignment [1] and transcript assembly [2].

These two domains have provided separate methods for generating the components of a parameter advisor and have given us insight on how to reapply it to new problems. Our goal is to eventually build a toolbox of methods for producing an advisor that can be re-used when applying advising to new domains.

1. DeBlasio, D and Kececioglu, J. *Parameter Advising for Multiple Sequence Alignment.* 2017. **Springer International Publishing**. Volume 26 of Computational Biology Series.
2. DeBlasio, D and Kingsford, C. *Automatically eliminating errors induced by suboptimal parameter choices in transcript assembly.* 2018. **bioRxiv**, 342865

Poster

Learning models in the data-poor but experiment-rich regime

Friday
3:45 pm
4:45 pm

Rory Donovan-Maiye & Gregory Johnson
Allen Institute for Cell Science

Learning robust and interpretable simple models where the number of features is much greater than the number of samples (P >> N) is a well studied problem and has been addressed by methods such as sparse linear models. However, the situation is much less straightforward when data is not iid, as is often the case in biology. In particular, when data is collected in batches where conditions are largely similar but differ in small but not insignificant ways (e.g. male/female cohorts, different levels of perturbations, neighboring time-points, etc), to overcome the endemic data-poor scenario one might desire to leverage information *across* non-identical samples to improve models *within* a sample.

For example, consider a situation where we want to learn the how the genes driving the expression of a certain target gene change over time and under exposure to different levels of a drug. An experiment is conducted consisting of single-cell RNA-seq data collected at ten different time points and ten levels of drug exposure, where for each of the 100 conditions, 500 cells are sequenced and counts for 20,000 genes are produced. Treating each of the 100 conditions independently and learning independent models for each leaves structural information on the table, and is subject to highly noisy predictions. However, if we assume the the models at neighboring time-points and drug-exposure levels won't be *too* different, we can start taking principled approaches to leverage data across conditions to refine our models within each condition.

I will present the work we've done building a tool to learn sparse linear models in situations like this, where we modify the elastic net objective function to respect sparsity but also penalize sparse solutions for being too different from solutions in neighboring conditions. Given the modified objective function, we derive a coordinate descent algorithm to solve the problem efficiently. The formalism we work in generalizes in a straightforward manner to conditions as nodes and similarity as edges on an arbitrary weighted graph.

Tracing the evolution of cis-regulatory elements across the mammalian phylogeny

Irene Kaplow
Carnegie Mellon University

Friday
3:45 pm
4:45 pm

The Genome 10K Project (G10K) is sequencing hundreds of mammalian genomes, enabling us to compare diverse species whose most recent common ancestors lived hundreds of millions of years ago. Many phenotypes evolved through gene expression, so they differ across species due to differences in cis-regulatory elements (CREs) that affect transcription. Since many of these phenotypes are tissue-specific, we are studying CRE strength in one tissue relative to another.

As a proxy for CREs, we are using DNA accessibility from ATAC-seq, an inexpensive assay that allows us to measure the open chromatin in tens of thousands of cells. Unfortunately, most tissues from most of the G10K species are impossible to obtain, so novel methods for using ATAC-seq from multiple tissues from a small number of species to impute CRE tissue-specificity in other species are needed. We are developing methods for multiple components of this task, including improving methods for identifying orthologs of CREs and training machine learning models to predict CRE tissue-specificity from DNA sequence.

As a proof-of-concept, we trained a convolutional neural network (CNN) to predict whether a CRE in mouse is brain-specific relative to liver or vise versa. Our current CNN achieves >90% AUROC and >80% AUPRC on chromosomes in the validation set and on human regions whose orthologs are not CREs in the same mouse tissue. In addition, we used deepLIFT to identify important nucleotides in each brain-specific example and found that nucleotides in motif hits of known brain transcription factors (TFs) tend to be more important than the nucleotides in motif hits of other TFs (Wilcoxon rank-sum $p = 2.662 \times 10^{-7}$). We are now using our new mapping method to identify the orthologs of mouse CREs in the other G10K mammals and our CNN to predict the tissue-specificity of CREs at these orthologous regions.

Scallop enables accurate assembly of transcripts through phase-preserving graph decomposition

Friday
3:45 pm
4:45 pm

Mingfu Shao
The Pennsylvania State University

Transcript assembly is the fundamental computational problem of reconstructing the full-length expressed transcripts from the (short) RNA-seq reads. Transcript assembly is crucial for transcript quantification and differential expression analysis and also plays a central role in revealing tissue-specific splicing patterns and understanding the regulation of gene expressions. Recently we have developed a new reference-based transcript assembler, called Scallop, published at Nature Biotechnology, 2017. Scallop significantly improves reconstructing multi-exon transcripts and lowly expressed transcripts: on 10 human RNA-seq samples, Scallop produces 34.5% and 36.3% more correct multi-exon transcripts than two leading assemblers StringTie and TransComb, while such improvement reaches 67.5% and 52.3% for lowly expressed transcripts. Scallop obtains such improvements through a novel graph decomposition algorithm that preserves all phasing paths while producing a parsimonious set of transcripts and minimizing coverage deviation. Scallop is freely available at

github.com/Kingsford-Group/scallop.

Scallop has been downloaded more than 800 times in a few months.

Poster

Characterizing protein-DNA interactions at individual genomic loci using high-resolution functional genomics assays

Naomi Yamada
The Pennsylvania State University

Friday
3:45 pm
4:45 pm

High throughput functional genomics sequencing assays enable us to profile regulatory activities genome-wide. High-resolution functional genomics sequencing assays such as ChIP-exo and ATAC-seq use nuclease digestion with or without chromatin immunoprecipitation (ChIP). Dense mapping of nuclease cleavage sites identifies small regions protected from digestion by the presence of a DNA-bound protein. Most approaches to analyzing high-resolution assays merely catalog the sites that are enriched for sequencing reads. However, analysis of the sequencing read distribution shapes created by the nuclease can potentially enable greater levels of biological insight by identifying the proteins that bind to DNA or the modes by which they bind.

For example, the ChIP-exo protocol precisely characterizes protein-DNA crosslinking patterns by combining ChIP with 5' to 3' exonuclease digestion. Since different regulatory complexes will result in different protein-DNA crosslinking signatures, analysis of ChIP-exo tag enrichment patterns should enable detection of multiple protein-DNA binding modes for a given regulatory protein. We have recently developed a mixture model-based approach to detect multiple DNA-protein interaction modes from a single ChIP-exo experiment by using both ChIP-exo read patterns and DNA sequence motifs. Our method can thereby characterize whether a profiled protein is interacting with individual sites directly or indirectly via protein-protein interactions with other regulators.

Future challenges in this domain include enabling integrative analysis across multiple high-resolution functional genomics experiments. We aim to develop computational methods that identify the positional organization of protein-DNA complexes from hundreds of ChIP-exo experiments. Furthermore, we seek to develop methods to detect site specific changes in nuclease footprinting shapes and read enrichment levels across different biological conditions. A major challenge in exploiting read pattern shapes from high-resolution data includes the high variation in per-base read counts. Read distribution analysis of assays that profile broader biological activities such as chromatin accessibility (e.g. ATAC-seq) leads to even less reliable read profiles than protein-DNA binding assays because of their lower read coverage per genomic position and cleavage biases associated with the nucleases. Possible solutions include applying probabilistic approaches to modeling read distribution shape. Despite the problems, exploiting read distribution patterns is a key to gaining a detailed understanding of gene regulatory architecture in a given cell type, and will thus be a fruitful application area for future algorithmic development.

Computational Biology Department

The Computational Biology Department is proud to be a sponsor of FAB 2018.

CBD offers programs at all academic levels

- B.S. in Computational Biology **(new in 2017!)**
- M.S. in Computational Biology
- Joint CMU-Pitt Ph.D. Program in Computational Biology (CPCB)
- Lane Fellows Postdoctoral Program

Find out more at

cbd.cmu.edu

Carnegie Mellon University

School of Computer Science

Preparing the next generation of computational biologists

Mike Schatz
Moderator — Johns Hopkins University

Friday
1:45 pm
2:15 pm

A panel of researchers in various areas of computational biology will share not only the current state of training but also how we will need to adapt our current methods to prepare the students of today to be the researchers of tomorrow. What are the essential skills that need to be added to our current curricula to support the future of algorithms in biology? What are the current topics that will need to be removed in order to make time, and which ones are too fundamental to be altered? All attendees are invited to participate in the discussion and ask questions of the panel.

Things to do in Pittsburgh

We encourage you to explore the many cultural and culinary attractions that make Pittsburgh unique. Here we have listed a small number of possible activities that FAB participants may enjoy before and after the meeting.

Eat

The Porch — Wood-fired pizza and seasonable selections with a park view.
221 Schenley Drive, corner of Forbes and Bigelow

Union Grill — Classic American style offerings with large portions.
413 S Craig St, on Craig just north of Forbes

Primanti Brothers — Famous Pittsburgh sandwiches.
3803 Forbes Ave, between Oakland Ave and Bouquet St
46 18th St, original location in the Strip District

Casbah — Mediterranean food and wine bar
229 S Highland Ave, in the Shadyside neighborhood

Independent Brewing Co./Hidden Harbor — Modern American/Latin Am.
1708 Shady Ave, just south of Forbes Ave in Squirrel Hill

Mineo's Pizza — Highly rated New York Style pizza.
2128 Murray Ave, between Hobart and Phillips, Squirrel Hill

Church Brew Works — Church turned brewery with classic Pittsburgh faire.
3525 Liberty Ave, in the Lawerenceville neighborhood

Plus many others, CMU and the Wyndham are in Oakland, and nearby neighborhoods include Squirrel Hill and Shadyside.

Explore

Carnegie Museums of Art and Natural History
4400 Forbes Ave, between Craig and Belfield

Cathedral of Learning Tour and Nationality Rooms
4200 Fifth Ave, between Belfield, Forbes, and Bigelow

Soldiers and Sailors Memory Hall
4141 Fifth Ave, between Bigelow and Thackeray

Schenley Park / Phipps Conservancy
1 Schenley Drive, just south of Oakland

Point State Park / Fort Pitt Museum
601 Commonwealth Pl, Downtown

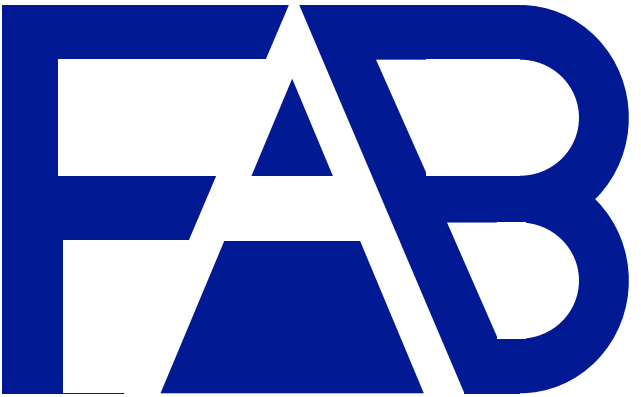
Andy Warhol Museum
117 Sandusky St, just north of Downtown

Duquesne Incline
1197 W Carson St, just south of Downtown

More extensive lists are linked from: fab2018.cbd.cmu.edu/attending

Notes:

Notes:



Workshop on the Future of Algorithms in Biology
September 28-29, 2018 — Carnegie Mellon University
fab2018.cbd.cmu.edu