Laboratory and Research Experience/Other Employment

Begin with current or most recent employment. Please include employer, dates employment started and ended, position, and nature of work. (Max 4000 chars) – QUESTION: I have ~3k chars so far, should I expand on my experiences?

Example:

-- Laboratory Experience example --

Los Alamos National Laboratory, summer 2011, Lab assistant, Research

Naval Research Laboratory, 5/2010–8/2010, Research Assistant, Immunosensor Research

-- Student Employment example --

Michigan State University, Engineering Department, 12/2010 - 5/2011, Lab Technician, Research.

University of Pennsylvania, 5/2010 - 8/2010, Research Assistant, Developed stochastic models of neurons.

Broad Institute of Harvard and MIT, 11/2010 – 8/2011, Associate Computational Biologist. Working with Dr. Jill Mesirov and Prof. Todd Golub, I developed REVEALER, an algorithm that integrates genomic and functional data to infer new associations. A pathway may be broken in many ways, but produce the same phenotypic output. Given a known phenotypic output, REVEALER finds candidate explanations by removing samples with known breaks in the pathway and uncovering new events via a mutual information metric. I presented this at the conference Intelligent Systems for Molecular Biology, another publication is in review, and three more are in preparation.

MIT Department of Brain and Cognitive Sciences, 1/2010 – 6/2010, Research Assistant. In Prof. Sebastian Seung's lab, I pursued computational image analysis. I analyzed neuron orientation in the rabbit retina inner plexiform layer (IPL) and found that neuron fragments in this 3D image lie along the axis of information transmission, and do not significantly traverse the direction perpendicular to of signal flow.

MIT Computer Science and Artificial Intelligence Laboratory, 1/2009 – 12/2009, Research Assistant. I worked in Prof. David Gifford's lab on two projects in yeast metabolic genome networks and T-cell receptor (TCR) sequences. For the yeast metabolome, I hypothesized methods of information content in graphs can inform synthetic lethality of a gene, and learned Java to implement the network in JGraphT. My metrics indicated purely genomic methods do not fully explain biological interactions. For the TCR project, I created random TCR sequences and tested the minimum frequency a particular sequence must occur for detection by statistical

methods. I found that even a slight (2x) increase in frequency was detectable, a promising result for diagnosis. This project was an exciting convergence of bioinformatics and clinical medicine as its implications allow health professionals to screen an individual's vaccination and antigen exposure history.

Howard Hughes Medical Institute (HHMI) Janelia Farm (JF) Research Campus (RC), 6/2008 – 8/2008, Summer Scholar. Admitted to the competitive JF Summer Scholars program, I worked with Dr. Eddy. I used Hidden Markov Models to create a robust null model for HMMER, sequence homology software developed in the Eddy lab. I completed this project in two months in Python and then rewrote it in two months in C, the language of HMMER.

Harvard Medical School, 6/2007 – 9/2007, Research Assistant. I worked in Prof. Bulyk's lab where I analyzed homeodomain transcription factor binding sites by performing protein-binding microarray experiments, resulting in a 2008 publication in the journal *Cell*. I was fascinated by how a computer could interpret the minutely polka-dotted microarrays and transform them into binding motifs, which inspired me to further explore quantitative analysis of biology. Resulted in a publication in the journal Cell.

Academic Awards and Honors

Include undergraduate and graduate honors (if applicable). (Max 4000 chars)

Example:

Highest Honors, U of Michigan, 2007 - 2010

National Merit Scholar, 2007

Dean's List, U of Michigan, 2007, 2008, 2009, 2010
Graduate
- University of California Regents Scholarship (2011)
Undergraduate
- Bernard M. Gordon-MIT Engineering Leadership Program (2009)
- Cold Spring Harbor Undergraduate Research Program (Summer 2009, declined to continue in Gifford lab)
- Howard Hughes Medical Institute Janelia Farm Research Campus Summer Scholar (2008)
- Cold Spring Harbor Undergraduate Research Program (Summer 2008, declined for HHMI JFRC above)

Extracurricular Activities

Include technical societies and service organizations.

Example:

Men's Fencing Captain, 2010

Volunteer for Habitat for Humanity

American Physical Society

-- Teaching --

- Creating bioinformatics modules to reinforce biology concepts in high school AP Biology (2011-present)
- Mentor for We Teach Science, weekly algebra tutoring to an 8th grader in San Jose, CA (2011-present)
- Mentor Scientist for Science Club for Girls, taught "Body Maps" (anatomy) curriculum to 2nd graders (2011)
- Volunteer math tutor at Mission Hill School, assisting 7th grade students with homework (2009-2011)
- Calculus Tutor in MIT Department of Mathematics (2007)

-- Professional --

- Developing biomedical research at MIT/Skolkovo Institute of Technology in Russia (Summer 2012)
- Co-Chair for ISMB Student Council Symposium 2012
- Member, International Society of Computational Biology (ISCB)
- Member, Society of Women Engineers

-- Extracurriculars --

Graduate

- Organizing "Miss Representation" documentary screening (Fall 2011 – Winter 2012)
- Fellowship critiquing circle (Fall 2011)
- Department-wide holiday gift exchange (2011)
- Math tutoring of high school and college students (paid, Sept 2011 - present)

Undergraduate

- Lightweight Men's Crew (Coxswain, 2006 – 2007)
- Everett Moore Baker House Executive Committee (Social Chair, 2008)
- Leadershape (2008)
- Kappa Alpha Theta Women's Fraternity (Zeta Mu chapter, 2007-2010)
- DanceTroupe (Dancer, Choreographer, Publicity Chair, 2007-2011)

Program of Study

Science/Engineering

Mathematics

Computer Science

Other Planned Courses

Research Statements

Field of Interest

"Computational science" involves the innovative and essential use of high-performance computation, and/or the development of high-performance computational technologies, to advance knowledge or capabilities in a scientific or engineering discipline. Please describe (in no more than 300 words) your personal research interest, paying particular attention to how computational science will help you make significant contributions to your chosen research area.

Current *in vitro* models of tumors are wrong. Results from cell lines are often extrapolated to a whole class of cancers, but less than 10% of tumor cells even survive *in vitro* and monoclonal lines do not represent tumor heterogeneity. Certainly, there are successful *in vitro* cases, such as PLX4032 designed to target BRAF's V600E mutation, but most drugs designed for a cell line are unsuccessful in clinical trials. Before a drug even reaches the clinic, it is unclear why applying a drug to monoclonal cell lines results in resistance, and whether survival was driven by intercellular interactions. Relying on statistics and success "most of the time" is an antiquated view – we now have the tools to understand the mechanisms behind an ineffective drug.

I want to shatter current acceptance of monoclonal cell lines as usable models of tumors. I will develop an *in silico* model of *in vitro* heterogeneity of HER2 positive breast cancer (BRCA) cell line SK-BR-3 in response to Trastuzumab (Herceptin, HER2 positive drug) by analyzing single-cell RNA-Sequencing (RNA-Seq) data, high-resolution snapshots of a cell's transcriptome which robustly determine alternative splicing isoforms. I hypothesize the heterogeneity of a cell population can be modeled by extracting network information from many single-cell transcriptomes. First, I will develop RNA-Seq assemblers tailored to the very small DNA sample amount of a single cell, less than a nanogram**.** Second, I will analyze differential transcripts between pairs of neighboring cells, accounting for differences in cell cycle stage using documented cell cycle genes. I expect to see a spectrum of active transcription in secretory pathways of neighboring respondent and healthy survivor. Using these differentially expressed genes and known interaction networks, I will develop a network of cell-cell interaction between adjacent cells in response to Herceptin. Without computation, this research would be nearly impossible to complete, as isolating new splice forms and comparing differential expression by eye is tedious and difficult to replicate.

Program of Study

The fellowship program of study requirement is designed to give you a breadth of competency in fields outside your own that will enhance your ability to perform computational science research. Please describe (in no more than 300 words) how you expect that the courses listed in your planned program of study outside your chosen discipline will contribute to your own research in the future. Describe why you chose these courses and how they will impact your research plans. QUESTION: 396/300 words – what should I cut?

As the computational collaborator in biomedical research, it is my job to be an expert in all things algorithmic and statistical, and my program of study gives me that knowledge through courses in computer science and mathematics. The 'Machine Learning' course will expose me to computational methods for both unsupervised learning such as clustering and supervised learning such as classification, both of which are key to bioinformatics. 'Algorithm Analysis,' 'Computational Complexity' and 'Optimization and Algorithmic Paradigms' will show me how to pinpoint bottlenecks and assess efficiency of the algorithms I develop, and as I hope to eventually create patient-side algorithms, a fast algorithm is important. Biomedical data is huge, and I am excited to take 'Mining Massive Data Sets' to both learn algorithms optimized for large data and understand the consequences of multiple hypothesis testing with these voluminous sets. As many questions in bioinformatics are conditional, the 'Bayesian Statistics' class will ensure that I am grounded in probability theory. 'Randomized Algorithms' will help me mimic biological networks as many interactions are derived from random, Brownian motion. Finally, I used mutual information methods in developing the REVEALER algorithm and am eager to immerse myself in the course 'Information Theory.'

Additionally, classes in biology, bioengineering, and bioinformatics will show me the current state of biomedical research, the instrumentation used to measure the biology, and the bioinformatics algorithms used to analyze these measurements. 'Advanced Molecular Biology,' 'Human Genetics' and … will teach me the detailed complexities of DNA and the subtle changes that can lead to disease. 'Biotechnology and Drug Development' and 'Applied Gene Technology' will expose me to current best practices in the biomedical instrumentation that creates the data I will use. 'Bioinformatics Models and Algorithms,' 'Computational Genomics,' TA'ing 'Computational Systems Biology,' 'Representations and Algorithms for Computational Molecular Biology,' 'Data Driven Medicine,' 'Modeling Biomedical Systems,' 'Smart Health through Effective Design,' and 'Translational Bioinformatics' will teach me the state-of the art methods for analyzing biological data.

I am now doing a Masters in one year—half the expected time—in Bioinformatics at the University of California, Santa Cruz, where I am deepening my understanding of computational biology. I hope to do my PhD at an institution with a strong computer science program and medical school, as I want to collaborate directly with physicians and develop new methods for analyzing patient samples to minimize invasiveness and maximize information gain.

High-Performance Computation and Research

The goal of this program is to support doctoral students in pursuit of novel scientific or engineering discoveries through the use of high-performance computing resources. There are many reasons to consider migrating a simulation to, or hosting large data sets in a high-performance computing environment. Some common motivations discussed in the DOE report "A Science-based Case for Large-scale Simulation"("SCaLeS", vol. 1, 2003; vol. 2, 2004; http://www.pnl.gov/scales/) are:

1. Better resolve the full, natural range of length or time scales in a model.

2. Accommodate physical effects with greater fidelity.

3. Allow the model degrees of freedom in all relevant dimensions.

4. Better isolate artificial boundary conditions or better approach realistic levels of dilution.

5. Solve an inverse problem, or perform data assimilation.

6. Perform optimization or control.

7. Quantify uncertainty.

8. Improve statistical estimates.

9. Operate without models.

For more information on these topics see the SCaLeS report cited above. By explicitly discussing one or two of the topics above or a different one of your own choosing describe in 300 words or less how migrating to a high-performance computing environment would advance your research beyond what is possible with a modest sized cluster. QUESTION: 221/300 words – expand here?

Please clearly indicate which of the above topics or your own topic you will discuss.
To advance the *in vitro* cell-cell interaction project I proposed to the next level of modeling tumor intercellular interactions, the *in vitro* scale, a high performance cluster is necessary to allow the model degrees of freedom in all relevant dimensions. While a modest sized cluster can handle the 2x20,000 differential genes between two individual cells, calculating interactions between the millions of cells even in a small biopsy becomes astronomically computationally expensive. Assuming cells are 2500 microns cubed on average, there are 4E8 cells in a small 1cm biopsy, and assuming each cell is a cube and interacts with only 8 other cells, that's 3.2E9 cell-cell interactions. Including the 20,000 differential gene comparisons, this amounts to calculating 6.4E13 variables, no small feat.

Besides differential expression, I want to incorporate other, epigenetic parameters to accommodate physical effects with greater fidelity. Epigenetic effects become even more important when applying a drug, as most drugs do not attack the DNA itself, but inhibit a protein coded by the DNA. Measuring this specific effect is difficult in biology and is typically found by

observing downstream actions. How can this effect be encoded into a genetic network? Typically, known protein-protein interaction networks are used to supplement gene expression, but when a drug is applied, changes in gene expression are indirect compensatory mechanisms that are difficult to detect.

List of publications:

Please include a list of publications authored or co-authored by the applicant.

Botvinnik OB, Tamayo P, Mesirov JP. Discovery of Genomic Features as Alternative Causes of Activation/Association with Functional Readouts or Phenotypic Differences: the REVEALER algorithm. (in preparation)

Galili N, Tamayo P, Botvinnik OB, Mesirov JP, Brown G, Raza A. Gene expression studies may identify lower risk myelodysplastic syndrome patients likely to respond to therapy with ezatiostat hydrochloride (TLK199). (in preparation)

Botvinnik OB, Tamayo P, Mesirov JP. Single-sample GSEA compares gene set enrichment of multi-sample experiments. (in preparation)

Wood KC, Konieczkowski DJ, Johannessen CM, Boehm JS, Tamayo P, Botvinnik OB, Mesirov JP, Hahn WC, Root D.E, Garraway LA, Sabatini DM. Miniaturized functional screening reveals genetic modifiers of therapeutic response in melanoma (in review)

Botvinnik OB, Tamayo P, Mesirov JP. Discovery of novel candidate oncogenic activators with REVEALER. Intelligent Systems for Molecular Biology Conference. Vienna, Austria (2011)

Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Peña-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET, Khalid F, Zhang W, Newburger D, Jaeger SA, Morris QD, Bulyk ML, Hughes TR. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell (2008) PMID: 18585359

Completed Courses

Description of 6.874

Additional comments

If necessary, include any additional comments regarding your application. (Max 2400 chars)