C. Titus Brown / ctb@msu.edu
Career statement

My overall focus is on connecting biological sequence data to biological function. This has been my research focus for the last ten years, and continues to be my career focus.

Broadly speaking, connecting data to function in biology encompasses a immense range of the biological sciences, computer science, and computational science. Genome and gene sequences connect to much of biology, and placing genomes and genes in their proper context requires broad and expert knowledge of areas ranging from molecular biology to microbial, plant, and animal physiology to developmental biology. The computational aspects of sequence analysis are also omnipresent, ranging from a basic understanding of a number of pattern matching and database search algorithms to database curation, algorithmic scalability, data management, statistics, software development and scientific software accuracy, and user interface development. However, if one truly wants to move the field forward, an even broader understanding must be reached: in particular, how biologists actually use (and need to use) tools, what biological approaches mesh with which computational analysis techniques, and the dividing line between "research questions" and automatable scientific workflow. This, in turn, connects with computational science education and how we can effectively educate new interdisciplinary researchers.

Increasingly, my research at MSU is focusing on the intersection of next-generation sequencing, analysis scalability, and cloud computing. Next-generation sequencing platforms produce increasingly large amounts of data (200 Gb/2 day run/$10k) and barely existed when I accepted the position at MSU, yet now the data from these platforms is ubiquitous in biology. Nearly every project in my lab, and all of my collaborations, involves working with this data. Because of the large volume of this data, questions of how to scale existing approaches become critical, and the field as a whole is starting to move towards cloud computing – the on-demand rental of compute resources from commercial providers. I have established myself as a leader in cloud computing in sequence analysis and have recently been invited to a number of conferences to speak on this subject.

This focus on next-generation sequence data analysis is evident in my papers from my first two years – all of which involve analyzing this kind of data – and the majority of the papers we are working on now are either on using this data to answer biological questions, or on new algorithmic approaches to scaling data analysis.

My most exciting computational project involves the development of an entirely new approach to sequence assembly based on a simple probabilistic data structure, an extension of a Bloom filter, that lets us scale sequence assembly by several orders of magnitude. The development of this novel technique was driven by a collaboration with Dr. Tiedje and the DOE Joint Genome Institute involving the deep

metagenomic sequencing of mixed microbial populations. We are currently working on several papers on this technique, intended for both computational and biological journals. Some of my work in this area has been funded by an NSF research grant.

Another computational collaboration that has already led to one paper (submitted, with W. Li) and an NIH proposal (in resubmission now) is work on the sea lamprey. We are working on the basic genome analysis, as well as a novel approach analyzing evolutionary relationships with other vertebrate genomes, and a range of sequence-level analyses underwriting many different biological projects in Dr. Li's lab. As part of this collaboration, we are developing a range of tools to deal with sequence from non-model organisms, which requires a variety of new techniques due to poor reference sequence and poor annotations.

A third computational collaboration, with Dr. Hans Cheng at the USDA Avian Disease and Oncology Lab, involves applying variation analysis and transcriptome analysis to the study of Marek's Disease Virus in chicken. As with the other collaborations, we are developing a range of new computational techniques to deal with issues ranging from data size to biological inference from poorly assembled reference sequence. This research is funded partly by a USDA grant with Dr. Cheng.

These three projects are specific examples of my primary *computational* research goal for the next few years, which is to build generally usable tools for scaling computational analysis of sequence data to a broad range of biological problems. This includes the creation of new tools, the integration of existing tools, and the development of user interfaces and deployment strategies for making these tools usable by non-computational biologists. The USDA recognized the importance of this general approach by funding a four-year, $700k grant for exactly this purpose.

This goal of building useful and usable tools for computational analysis of biological sequence data also intersects with my mentoring, teaching and outreach activities.

I have a medium-sized lab with six graduate students and three postdocs. I co-advise one graduate student with Charles Ofria and another with James Tiedje, and I also share one postdoc with James Tiedje. In most cases, the graduate students are not trained in the area in which they are doing research: I attract the computer scientists that want to do biology, and the biologists that want to do computational science. This results in an extremely creative group that is rarely content with applying just a single approach to any problem, and it has been a pleasure to mentor them and bring them to interact with each other. One notable success story is a Fisheries and Wildlife Master's student, Jiarong Guo, who entered my lab, learned to program and do sequence analysis on metagenomic data in a year, graduated, and is now a PhD student shared between my lab and James Tiedje's lab.

During my time at MSU I have developed four new courses: one undergraduate course, which is focused on effective programming of Web interfaces and the architectural issues underlying them; and three graduate-level courses, all of which

are centered on the appropriate and effective use of computation to answer biological questions. The undergraduate course, which connects with research issues of software architecture, testing, and user interface, has been extremely popular and led to a Withrow Teaching Award the first time I taught it. Two of the three graduate level courses (a Summer 2010 intensive course/workshop on next-generation sequence analysis, and my current Fall 2010 BEACON course on Computational Science for Evolutionary Biologists) use a strongly guided strategy to introduce biologists to concepts and practice in computational scientist, and also have been heavily subscribed (although note, only 3 students took the Summer 2010 course for credit; an additional 20 students took it as a research workshop). My fourth course, offered in Spring 2009 and Spring 2010, is a seminar course on a wide-ranging set of topics in biological data analysis, that is attended by both CSE and biology graduate students.

In all of my graduate level courses, I focus on critical thinking and the identification of appropriate computational strategies for the end goal of addressing biological problems. This approach has proven to be quite popular, with attendance of my seminar nearly doubling in the second year that I taught it; other professors have told me that the course has been effective in raising the level of their students' research.

My outreach activities have been similarly focused on raising the bar on computational science in general, and computational biology in specific, with the significant side goal of raising my research profile internationally. For example, I write a well-subscribed blog which covers topics in computational biology, software engineering and testing, scaling, open science, and data management; this blog has brought me the attention of a number of bioinformatics researchers and cloud computing companies. It has also been instrumental in generating invitations to give talks at universities and conferences and generating potential collaborations. Largely as a result of this activity, I am now on the editorial board of a new BMC journal, "Open Research Computation", focused on open computational science.

I am also heavily involved in outreach in computational science education, where I have developed a site (http://ged.msu.edu/angus/) containing educational materials for training students in next-generation sequence analysis using the cloud. I am extending this site now to contain materials for running Avida and LTEE resequencing analyses as part of the BEACON course that I am teaching.

My service at MSU has largely consisted of colloquium committee work (in both CSE and MMG), where I have especially helped bring in an array of computational biology and genomics speakers to CSE, MMG, and SATE. I am also the faculty organizer of the Systems Biology conference (Oct 23, 2010), which brought 6 internationally known scientists to MSU.

While all of these research, teaching, and outreach activities are quite far flung, they all address a common goal: using computation to make sense of biological sequence.

As this is one of, if not *the,* central themes of the next decade in biological research, my activities fit well with the departmental, collegial, and university goals in hiring me: to become a leader in research and teaching, to catalyze collaborations within and without MSU, and to help train the next generation of students and researchers in interdisciplinary techniques, tools, and thought processes. The central challenge for me is to maintain enough focus to make substantial and identifiable progress in one or two areas of computational biology, while still moving my broader research, education, and outreach efforts forward. My preliminary progress in obtaining research grants, getting invited to conferences, and (soon) submitting papers for publication, suggests that I am making some progress in this effort.