Research Statement - C. Titus Brown

The future of biology and biomedicine lies at the intersection of data gathering, hypothesis generation, and hypothesis testing. It is increasingly easy to gather large amounts of relevant data, but **interpretation of that data and connection of that data to downstream hypothesis-driven experimentation is still difficult and rare**. Effectively integrating large scale data analysis with "wet" biology requires the development of new perspectives and new tools, and their integration with current research programs; it also requires biology-immersed scientists who are computationally expert.

My research plans focus on the opportunities at the intersection of computation and biology: in particular, I am interested in enabling and doing data-intensive biology, where I can build tools and approaches that harness large investigator-driven data sets to direct hypothesis-driven experimentation and hypothesis-free interpretation.

I have spent 20 years applying computation effectively to pressing scientific problems. During this time I have worked in evolutionary modeling, data analysis for climate studies, regulatory genomics, developmental gene regulatory networks, and bioinformatics. My graduate and postdoctoral work in evolutionary developmental biology used computation to generate hypotheses in a range of biological systems, including sea urchin, chick, microbes, and marine sediment microbial communities. Moreover, in all these areas I have worked closely with experimentalists to connect hypotheses to experiment. Over the coming decade, I will continue to work on important biological problems using computation and collaboration.

Current research: Data Intensive Biology

As an Assistant Professor at Michigan State University (since 2008), I have focused largely on the opportunities for studying non-model organisms that arrived with the sequencing revolution. Here my work has expanded in biological scope to analyze high-throughput genomic and transcriptomic data from a variety of agricultural animals, environmental metagenomes, and marine animals. To tackle these biological questions, my students, postdocs, collaborators and I have created a novel series of algorithmic approaches to solving problems in *de novo* sequence assembly. These approaches have enabled faster, deeper, and more sensitive analyses of data sets than would otherwise be possible.

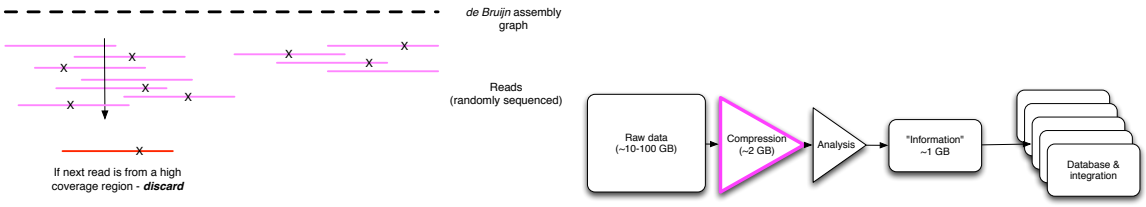

Figure 1: Left: The digital normalization procedure smooths out coverage prior to *de novo* assembly, leading to dramatically decreased computational requirements and (in some cases) better results. See [1] for details. This is a specific example of our general approach towards *lossy compression* of data sets. Right: Lossy compression approaches (like JPEG image compression) decrease data set sizes while conserving information.

Environmental metagenomes: Development of our three primary computational approaches – memory efficient k-mer counting [2], partitioning [3], and digital normalization [1] – was driven

by the need to investigate extremely large metagenomic data sets from agricultural soil. The vast diversity and richness of soil makes soil metagenome analysis one of the hardest sequence analysis problems in bioinformatics, and our approaches have let us achieve assemblies of some of the deepest soil metagenomes ever sequenced [4]. These assemblies enable considerably better functional and biographical annotation, analysis, and investigation of the soil communities than is otherwise possible.

These same approaches have also been applied to other metagenomes, including metagenomes from host-associated microbiota and endosymbionts [5], where they also yield excellent results. Thus my metagenomics work has expanded from soil to include collaborations with Human Microbiome Project researchers and biogeochemical cycling researchers working on marine, lake, and hot spring systems.

Marine genomes and transcriptomes: We have also been working on lamprey and Molgulid (stolidobranch ascidian) genomes and transcriptomes. Here the challenge is similar to environmental metagenomes, in that tremendous amounts of sequencing data must be analyzed before anything biological can be extracted.

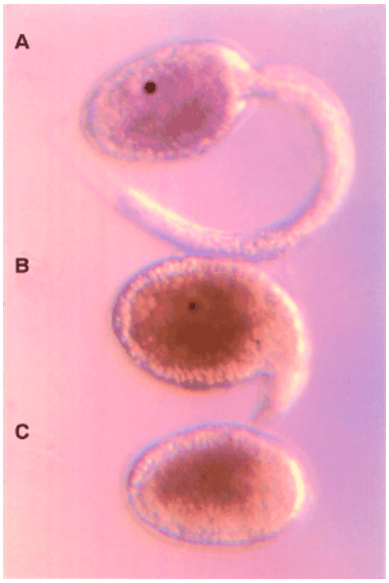

Figure 2: Two Molgulid ascidians have divergent tadpole and chordate-like development (a and c). We generated hybrids and analyzed developmental dynamics at the transcriptome level.

In the case of the lamprey, somatic recombination of the liver tissue we sequenced rendered the genome-based gene set incomplete [6]. We therefore generate an additional 5 billion transcriptome reads from over 60 tissues in an attempt to generate a better transcriptome. These data were impossible to analyze on any existing computer with any existing program. Applying digital normalization let us readily generate a transcriptome superior to the reference-based gene set, which we are in the process of writing up for publication. The insights from this transcriptome have already been applied in research on the evolution of olfaction [7] as well as pheromonal response [8], and we are collaborating with groups that study spinal cord regeneration, biliary atresia, and the mechanisms of pheromone response in lamprey. Without digital normalization, this transcriptome could not have been assembled.

Transcriptome and genome analysis have also yielded mechanistic insights into the nature of tail loss in the Molgulid ascidians. This group of sea squirts has repeatedly and independently undergone larval tail loss, and from comparison of tailed, tail-less, and hybrid transcriptomes we now believe that tail loss occurs after differentiation but prior to morphogenesis; this has been confirmed by direct imaging, which shows that tail cells are specified but fail to extend. One of the mechanisms behind this appears to be heterochrony of metamorphosis genes, which in the tail-less species turn on early and may activate resorption of the tail. Our computational approaches enabled rapid assembly and analysis of these highly polymorphic and divergent genomes and transcriptomes.

Agricultural genomes and transcriptomes: We have been collaborating with a number of agricultural researchers to improve the state of the chicken and cow genomes and transcriptomes. In particular, we have just been funded by the USDA for a collaboration with the St. Louis genome sequencing center to integrate Sanger, 454, Illumina, and PacBio data to improve the chicken genome assembly[1]. We are also collaborating with a USDA lab on improving the chicken transcriptome and analyzing Marek's disease resistance lines [9], and are working with another lab at MSU on cattle mRNAseq to determine genomic mechanisms of paratuberculosis resistance.

One major achievement in this area was in enabling the assembly and analysis of the *H. contortus* genome [10]. *Haemonchus contortus* is a parasitic nematode that has significant agricultural impact because it infests ruminants. The genome presented a number of challenges, including substantial polymorphism, bias in the data from genome amplification, and extensive repeat structure; the first good assembly from Illumina data was completed due to the use of digital normalization.

Development of computational tools: Our focus has been less on developing novel approaches than on enabling data-intensive biological investigations by whatever means necessary. While this has led to several novel algorithms, we have also spent a considerable amount of time building tools that can be used by many. These tools are in increasingly wide use: digital normalization has only been published as a preprint, but has been cited in 10 or more publications in the past year, and we estimate that it is being used by several hundred research groups. It is clear that our approach to designing and building software is an effective way to advance the field of biology.

Collaborations: We are by necessity extremely collaborative! The projects outlined above are in collaboration with a dozen labs and many institutions, including Janet Jansson at LBL, Billie Swalla of UW Seattle, Jennifer Morgan of Woods Hole MBL, Ona Bloom of the Feinstein Institute, Paul Sternberg of Caltech, Shana Goffredi of Occidental College, and Wes Warren of St. Louis. This is an addition to collaborations with five different research groups at MSU and the neighboring USDA ADOL facility. My preferred approach to collaboration is to co-advise graduate students or postdocs with my collaborators; this allows the student or postdoc to be embedded in the biology, while providing them with deep access to our computational expertise.

Education, outreach, and teaching related to research: I believe that computational education is perhaps the biggest challenge facing biology in the next 20 years, and I am heavily involved in bioinformatics and computational education. In particular,

- I am the founding director and organizer of a 2-week intensive summer workshop on Analyzing Next-Generation Sequencing Data, now in its fifth year (and funded through 2016 by the NIH). (See: bioinformatics.msu.edu/ngs-summer-course-2014)
- I participate regularly in teaching and training activities elsewhere, including the Embryology and Microbial Ecology courses at Woods Hole/MBL.
- I am on the advisory board of Software Carpentry, which seeks to educate scientists on efficient, effective, and correct approaches in computational science.
- I routinely run workshops and training sessions for our local NSF Center, BEACON, and the MSU High-Performance Compute Center.

All of the materials we develop for workshops and courses are available on the Web under an unrestricted Creative Commons license and attract several hundred thousand visitors a year[2]. One

[1]Note: this grant has not yet been officially awarded as of Nov 1, 2013, due to the government shutdown.

[2]http://ged.msu.edu/angus/

notable aspect to this work has been the use of cloud computing for teaching, which has enabled many to reuse our materials.

Open science and scientific reproducibility: I am a strong advocate of open science, open source, open data, open access, and the use of social media in research as a way to advance research more broadly. I write a fairly popular blog (ivory.idyll.org/blog/) that is regularly featured on MSU's "Spartan Ideas" feed, and I participate in Twitter conversations about research, diversity, and education. All of my senior-author papers are available as preprints; all of our source code is published on github; and I am well known in the areas of open science and scientific reproducibility, and have contributed to several publications and blog posts in this area [11]. As these topics have gained visibility over the last two years, I have been interviewed by an increasing number of scientific news outlets (see ged.msu.edu/press.html). I have also been invited to present at the NAS, NSF, and NIH on these topics, and I am a newly appointed member of the NSF iPlant Collaborative Scientific Advisory Board.

Future Research

In the short term, I will continue working on several problems that are directly downstream of my current research. These all relate to open challenges in next-generation sequencing, including (1) efficient assembly and exploration of ever-larger data sets; (2) detection and resolution of variants in complex mixtures, including strain variation and polymorphism; and (3) engineering implementations of everything at large scale. This is ongoing research funded by the NIH through 2016; we know how to do it and it is no longer a theoretical or methodological challenge, merely a (difficult and important!) research-level engineering problem.

In the long term, there are three major challenges in data intensive biology that I would like to tackle.

First, **linking genotype to phenotype**. This is a cross-cutting theme of the current sequencing bonanza, and we are sorely lacking in the methods and mechanisms to do this effectively. I am actively working with researchers in environmental metagenomics, agricultural genomics, and biomedical metagenomics to "connect the dots" between metadata, genomic information, and phenotypic information, but a much bigger effort must be mounted once primary sequence analysis is no longer computationally difficult. The chief challenges here are computational (integrating large, heterogeneous data sets without common keys), statistical (teasing out correlations without drowning in false positives), and "ecological" (building a sustainable ecosystem of correct and reusable tools). We must make progress on all of these problems in order to connect data effectively to multiscale models and hypothesis-driven experiment. However, this must be done in domain- and project-specific ways, which requires computationally expert domain specialists like myself.

Second, the related question of **genes of unknown function**. Functional gene annotation is a catastrophe, even in model organisms, where between 30% and 60% of genes have only inferred function; these annotations can be transferred into agricultural genomes, but should not be transferred blindly. In environmental metagenomics, it is worse: typically 60-80% of genes have no close homology and no strongly inferred function. We need a combination of biology method development, database building, computational investigation in correlation with observed phenotypes (see the first challenge, above), and mature and usable software capable of exploring and navigating homology relationships and functional assignments. As above, this must be done in domain- and project-specific ways, which requires close collaboration between experimental and computational biologists.

Third, the meta-problem of **opening up data, software, and science.** Most data is most useful in the context of other data; software that cannot be read, understood, modified, repurposed, and redistributed is severely limiting; and open science maximizes serendipitous discovery. We know that we can maximize our ability to progress in basic research on medical, environmental, and societal challenges by opening up data, software, and science. Yet biology has been strangely conservative in this regard, due to a combination of culture, lack of incentives, and lack of training. I have explored some promising paths with my work (open source, blogging, preprints) as have some others, but we do not yet have a sense of what will scale to many. There are great opportunities, moreover: grant agencies are desperate to leverage the vast amount of data and software they are funding, and we have many individual examples of how opening up science, data, and source can accelerate both individual research and larger research agendas. I am well placed to help tackle these issues and am already embarking on some initiatives to incentivize open data in transcriptomics and metagenomics.

I am already working on these three major challenges with many of my collaborators, since they are obvious next steps for most biologically driven sequencing investigations. I am looking forward to the opportunity to focus in on these challenges as our current work matures.

References

[1] Brown C, Howe A, Zhang Q, Pyrkosz A, Brom T (2012) A reference-free algorithm for computational normalization of shotgun sequencing data. In revision for PLoS One; Preprint at http://arxiv.org/abs/1203.4802.

[2] Q Z, Pell J, Canino-Koning R, Howe A, Brown C (2013) These are not the k-mers you are looking for: efficient online k-mer counting using a probabilistic data structure. In review at PLoS One; Preprint at http://arxiv.org/abs/1309.2975.

[3] Pell J, Hintze A, Canino-Koning R, Howe A, Tiedje J, et al. (2012) Scaling metagenome sequence assembly with probabilistic de bruijn graphs. Proc Natl Acad Sci U S A 109: 13272-7.

[4] Howe AC, Jansson J, Malfatti SA, Tringe SG, Tiedje JM, et al. (2012) Assembling large, complex environmental metagenomes. In review at PNAS; Preprint at http://arxiv.org/abs/1212.2832.

[5] Goffredi S, Yi H, Zhang Q, Klann J, Struve I, et al. (2013) Genomic versatility and functional variation between two dominant heterotrophic symbionts of deep-sea osedax worms. Accepted for publication in Intl Society for Microbial Ecology, Oct 2013.

[6] Smith J, Kuraku S, Holt C, Sauka-Spengler T, Jiang N, et al. (2013) Sequencing of the sea lamprey (petromyzon marinus) genome provides insights into vertebrate evolution. Nat Genet 45: 415-21, 421e1-2.

[7] Chang S, Chung-Davidson Y, Libants S, Nanlohy K, Kiupel M, et al. (2013) The sea lamprey has a primordial accessory olfactory system. BMC Evol Biol 13: 172.

[8] Chung-Davidson Y, Priess M, Yeh C, Brant C, Johnson N, et al. (2013) A thermogenic secondary sexual character in male sea lamprey. J Exp Biol 216: 2702-12.

[9] Subramaniam S, Johnston J, Preeyanon L, Brown C, Kung H, et al. (2013) Integrated analyses of genome-wide dna occupancy and expression profiling identify key genes and pathways involved in cellular transformation by a marek's disease virus oncoprotein, meq. J Virol 87: 9016-29.

[10] Schwarz E, Korhonen P, Campbell B, Young N, Jex A, et al. (2013) The genome and developmental transcriptome of the strongylid nematode haemonchus contortus. Genome Biol 14: R89.

[11] Wilson G, Aruliah D, Brown C, Chue Hong N, Davis M, et al. (2013) Best practices for scientific computing. Accepted at PLoS Biology, Oct 2013; Preprint at http://arxiv.org/abs/1210.0530.