Major Achievements

The primary focus in my varied career has always been on doing good computation in the service of science, which has led in some odd directions:

Avida, 1992-1997

In 1993, as my freshman summer project, I wrote the first version of the Avida platform for digital evolution. This is an environment in which simple self-replicating computer programs evolve and adapt. Since 1993, the Avida platform has become a major model for investigating evolutionary principles. Of particular note, the BEACON NSF STC at MSU is partly centered around using Avida for research in evolution. There are now several dozen researchers working with Avida. Avida has helped legitimize the role of bottom-up computational models in evolutionary research.

Genomics, 1999-2007, Caltech

In graduate school, I worked on sea urchin genomics in the Davidson Lab, learning molecular biology and embryology for experimental regulatory analysis of a spatiotemporally transcribed gene. This is where I learned bioinformatics and wrote several tools for sequence analysis. Significant outputs during this time include a Web site for regulatory analysis, and a strategy for pre-assembly analysis of sea urchin genome content. My postdoc was in chick genomics, doing similar work. In both labs, my tools helped turn the purely ”wet” 2 year process of finding and analyzing cis-regulatory regions for a gene, into a 3 month summer project.

Marine (meta)genomics, 2005-, Caltech and MSU

In 2005, I initiated a collaboration with Victoria Orphan, analyzing 454 data from her bead enrichment work on marine sediment. This culminated in a PNAS paper, and also led to a collaboration with Shana Goffredi, sequencing on an Oceanospirillales endosymbiont in a species of Osedax, a bone-eating annelid. With Dr. Orphan, I discovered and resolved the 454 replicate problem before others, but we didn’t discuss this outside of Methods. With Dr. Goffredi, we recently assembled a nearly complete endosymbiont genome from MDA shotgun sequencing of host+symbiont DNA.

Metagenomics and next-gen sequence analysis, MSU, 2008-

I arrived at MSU just as Illumina sequence began to really flow, and my current work is largely motivated by the need to intelligently analyze extremely large amounts of metagenomic and mRNAseq data. My flagship project has been the Great Prairie metagenome, in collaboration with Jim Tiedje and JGI. We have now resolved the unprecedented challenges involved in assembling 300 Gbp+ of Illumina sequence from an agricultural soil environment. In the process we have also provided generic solutions for scaling and improving assembly of meta-omes, single-cell genomes, transcriptomes, and highly polymorphic and repeat-rich eukaryotic genomes.

Open access to tools and training

Since 2010, I have run a two-week intensive summer course on next-generation sequence analysis for biologists. This course is attended by 24 students each year, with 168 applicants in 2012. All of the course materials are freely available. The course joins with my own computational research in help democratize sequence analysis by making tutorials, compute infrastructure, data sets, software tools, and publications freely available, easy to use, and open access. In 2012 and 2013 the course is NIH funded (R25). Note, in August, I am participating in the STAMPS course.

Current Research

My current research is driven by the need to generate high quality hypotheses from extremely large metagenomic data sets.

For the past three years, we have been working on how to analyze short-read shotgun sequencing data from soil and marine metagenomes and MDA samples. We initially tried analyzing single reads but found that the reads were too short for robust homology-based analysis of rare genes. This led us to *de novo* assembly, a notoriously challenging computational problem; for diverse metagenomes, with large amounts of data and uneven coverage, it is a grand challenge.

We have developed two basic approaches that solve the major problems specific to metagenome assembly. The first approach, partitioning, breaks the sample into disconnected components based on transitive read connectivity. These components can then be assembled independently, with parameters chosen for specific genomic features. Using computational spike-ins we recover 98% or more of input genomes using this process, while scaling overall assembly by a factor of 20 or more.

Our second approach, digital normalization, is transformative. Digital normalization relies on the observation that for the uneven abundance distributions characteristic of metagenomes and MDA samples, the vast majority of shotgun sequencing data will be from high abundance organisms and is redundant. Digital normalization does locus-specific downsampling based on de Bruijn graph structure, while retaining low-abundance sequences; it is a converse of the level-set approaches pioneered by the Banfield Lab and JCVI. We have found that we can discard 95% or more of sequencing data and yet get *improved* assemblies on single cell MDA data relative to e.g. Velvet-SC.

Using these novel approaches, we have been able to assemble 85% or more of the genome of an endosymbiont of a bone-eating worm from an enriched MDA sample generated by Shana Goffredi (a collaborator). This improves the assembly done by standard practices by more than a factor of 2. We have also assembled vast amounts of data from several incredibly complex ecosystems, with our current best achievement being the assembly of a 320 Gbp agricultural soil sample, yielding 3.5 Gbp of genomic contigs $>$ 300bp.

Our focus in this work has been to lay a solid foundation on which other researchers can stand. Most of the published metagenome assembly approaches rely on computational approximations that reduce the sensitivity of results, and it is generally unclear how to choose and evaluate parameters; this is something we tackled in conjunction with our novel algorithms. We now have a widely applicable set of tools and computational protocols that can be used to assemble metagenomic data, evaluate and compare the assemblies, and inform future sequencing efforts.

Embarking on this research at the start of my faculty position was, frankly, stupid: we were entering a well-populated field with famous problems and a long publication history, and, moreover, a field in which I had little background. We had no expectation of success, but persevered because of the pressing biological need; I think this is why we have been successful in finding solutions.

Future research

I plan to continue to focus on analyzing and interpreting sequence data in support of understanding microbial communities. With the rise of next-gen sequencing, it is increasingly clear that we desperately need good computation, both for data analysis and downstream integration into experimental science and modeling. The good news is that sequencing is finally generating enough data to inform, build, and constrain interesting models; the bad news is that we are ill prepared to build these models and integrate them with data and biology. Advances in sequencing and data analysis, annotation, and data-driven model building, are all extremely important for understanding how microbial communities assemble, evolve, and are maintained [Brown and Tiedje, 2011].

Introduction

Ensemble sequencing of complex marine communities is now a standard way of examining both metabolic potential and metabolic activity. Our data generation capacity has increased dramatically, to the point where *sensitivity of sampling* is less of a concern than *sensitivity and interpretation of analysis*. I propose to address two significant problems: metagenome assembly, and metagenome interpretation.

Assembling infinite data

We are close to being able to generate essentially infinite sequence data at virtually no cost. Illumina HiSeqs can generate 600 Gbp a week, and will likely continue to be the dominant technology for deep sampling of complex microbial communities for the next 5 years. While longer read technologies are important, it is unlikely that the sampling depth of these technologies will increase enough to do sensitive, quantitative analysis of metagenome and metatranscriptome data.

To robustly detect low-abundance organisms at (say) one in a 100,000 dilution with shotgun sequencing, approximately 5 Tbp of shotgun sequencing is needed. Shockingly, we will actually be able to generate this volume of data in the near future! However, assembling this data is essentially impossible with current approaches.

There are significant challenges in analyzing and assembling large amounts of short-read data. Our technical goal is to make assembly of arbitrarily large metagenomic data sets simple and straightforward. In addition to the scaling problem – current approaches try a holistic approach, which is impossible for massive data sets – we are building tools capable of dealing with significant strain variation, such as we expect to see in viral and phage populations.

The two approaches we have developed, however, are capable of solving the basic scaling problem with relatively little additional engineering [Brown et al., 2012, Pell et al., 2012]. Digital normalization provides a lightweight single-pass approach for discarding data that has already been seen, while partitioning can be done progressively as data arrives: in particular, we can take advantage of the uneven abundance distribution in most complex ecosystems to ”assemble out” the most abundance organisms first, and then discard any further such data.

Digital normalization and partitioning also provide a theoretical basis for integrating multiple different data types: e.g. long reads can be used to link partitions across repetitive sequence, and digital normalization can eliminate redundant short-read data and error-correct long-read data. For scaffolding and extracting complete genomes, the Armbrust Lab pipeline could be integrated with our current tools [Iverson et al., 2012].

We therefore propose to implement ”infinite assembly” for metagenomes and metatranscriptomes, including error correction of reads and integration of different sequencing types such as 454, PacBio, Ion Torrent, and Nanopore, as well as whatever new technologies arise. Our goal is to make ”push button” assembly available on lightweight rental compute infrastructure such as the Amazon cloud, while providing an open set of tools, techniques, and algorithms. These approaches should also work well on on metatranscriptomic data, and we will provide both reference-free and reference-based analysis of metatranscriptomes.

In addition to providing significant leverage in meta-omics, digital normalization also provides a principled approach to assembling sequences from single-cell MD-amplified genomic DNA. In fact, in the digital normalization paper we re-analyze a SAR324 data set and show that we get approximately 10% more conserved sequence out of it than the current best approach [Brown et al., 2012, Chitsaz et al., 2011]. Our approach also dramatically decreases the computational requirements for the assembly process. We plan to extend our contig assembly work to scaffolding assembly, which should improve contiguity of MDA genomes; the Iverson et al. work may provide a path.

Recovery of strain variants and analysis of very polymorphic phage and viral genomes are extremely challenging, bioinformatically. There are several anecdotal reports of assemblers either collapsing significant strain variation, or simply discarding such genomes. Digital normalization seems to enable better recovery of core genomes.

Some additional goals of our work are, first, to develop protocols for generating high quality assemblies; and second, to develop *evaluation* tools and protocols. Given the increasing interest for the role that microbial consortia play in the environment, de novo metagenome sequencing and assembly will continue to be an active area of research, yet there has been little in the way of broad evaluation of metagenome assemblers across many environments; we hope to initiate an Assemblathon- or GAGE-like effort in this area, as it is desperately needed [Earl and et al., 2011, Salzberg et al., 2012].

Community modeling to understand diversity

Functional investigation and controlled perturbation of complex microbial communities is notoriously difficult, which makes it virtually impossible to analyze ecological associations without the use of modeling techniques. Several recent efforts have used network models of interactions to analyze, investigate, and predict taxonomic structure, but these efforts have relied on tag sequencing to observe communities, and are therefore limited in their explanatory power [Steele et al., 2011, Larsen et al., 2012a, Larsen et al., 2012b]

As whole-metagenome shotgun sequencing and analysis continues to advance in capability, we will be generating deep spatiotemporal profiles of both microbial taxonomy and function. In addition to basic data analysis challenges (above), we face the challenge of understanding the functional interactions of these communities and predicting novel functions based largely on sequence and variation in presence and expression of that sequence – i.e. metagenomic and metatranscriptomic data. The goal will be not just to catalog known genes, but to constrain the set of possible models based on the data.

At present, we have no “null model” for how complex communities assemble, are maintained, and evolve. The Paradox of the Plankton suggests that due to the competitive exclusion principle, rich communities should not persist in oligotrophic environments. So why do such diverse, uneven marine communities exist in the deep sea?

Results from several recent models suggests that multiple, independent stable commensal communities may form around the utilization of a limited set of resources, even in a well-mixed environment [Cooper and Ofria, 2003] (Ostman and Adami, unpublished). In this situation, each commensal community would be collectively and independently optimized in their ability to metabolize available resources. The models suggest that community members will compete on their ability to metabolize rare resources, while availing themselves of more abundant resources as possible. These commensal communities can be identified by their collective resource consumption, and are resistant to niche invasion by individual species from other stable communities.

Intriguingly, these models suggest that highly diverse microbial ecosystems could form and be maintained without significant spatial structure, temporal variation, or syntrophic interaction. Since we *see* plenty of highly diverse microbial ecosystems in practice, the question of whether or not these models apply is important for understanding resource usage within natural systems.

Even more interesting is the question of whether we can *use* such models to help in investigation and understanding of microbial consortia, based on modeling observational data. Specifically, (1) is there a metagenomic or metatranscriptomic ”signature” of this kind of community interaction, and how can we maximize our ability to accept or reject this as a null hypothesis, e.g. through spatiotemporal sampling strategies? (2) what kinds of interactions are spurred by added environmental or community complexity, what meta-omic signatures would be present, and can we distinguish them from the null model? (3) How do stochastic fluctuations affect these communities and modify the community structure?

This kind of deep interaction between data analysis, modeling, and hypothesis generation is central to other fields with vast amounts of data, e.g. experimental particle physics. Developing a similar ability to interpret metagenome data and extract the signal of interesting interactions may be key to understanding complex microbial communities.

Other projects: Sequencing Ice Cores

In collaboration with Buford Price, Steven Giovannoni, and Richard Lenski, I have been discussing how best to study apparent Prochlorococcus species present in ice cores, estimated to be 30-40k years old. These bacteria appear to be Prochlorococcus based on morphology, but we have not yet investigated them genetically or genomically. We are particularly interested in examining the gene content and phylogenetic placement of these ancient organisms relative to modern-day Prochlorococcus, given the changes in ocean conditions over the last 40,000 years; this could shed light on rates of adaptation and genetic divergence underlying adaptation. Because there are only a few cells present in each layer of the ice cores and grown these bacteria may be difficult, we plan to apply culture-independent sequencing. Downstream bioinformatic challenges include recovering sufficient contiguity from the samples, as well as developing approaches for doing phylogenies on metagenomic samples.

Other projects: Euprymna scolopes

As part of the Cephalopod Genomics Consortium, Spencer Nyholm and I have initiated a collaboration to sequence, assemble, and annotate the genome of *Euprymna scolopes*, the Hawaiian bobtail squid that hosts a *Vibrio fischeri* consortium. As with the Osedax work with Shana Goffredi (see Current Research), this is an opportunity to examine host-associated marine microbes at a genomic level.