Designing Cyberinfrastructure for Future Users

Matthew J. Bietz & Charlotte P. Lee
University of Washington, USA
[mbietz/cplee]@u.washington.edu

Abstract. Scientific information infrastructures are expected to operate over long time scales, but this creates challenges for the design of those infrastructures. This paper uses the example of cyberinfrastructures for metagenomics research to illustrate some of the issues that can arise when scientists attempt to use legacy cyberinfrastructures to answer new research questions. New science brings new forms of data, new analysis tools, and the need to recontextualize existing data. Cyberinfrastructure design is complicated by the difficulty of predicting future user requirements. We discuss three strategies for addressing these issues that are emerging in the metagenomics domain.

1 Introduction

Information infrastructures operate over relatively long time scales, but this creates certain challenges for the designers of these infrastructures. Often the infrastructure is expected to persist through funding cycles, changes in technologies, the coming and going of people involved in the project, and larger social and policy changes (Edwards, et al., 2007; Ribes & Finholt, 2007). One particularly difficult challenge is that as the infrastructure evolves, the user base may change. As users change their focus or new users arrive, they present a new set of requirements and infrastructure needs. Here we use the emergent field of metagenomics research to illustrate some of the challenges that arise when scientists begin to use existing information infrastructures to answer new research questions.

Metagenomics, sometimes called population genomics or environmental genomics, is a "new science" that allows scientists to study the genetic

composition of populations of microorganisms to understand biological diversity, microbes' functional roles, and microbial impacts on and adaptations to their environments. Metagenomics is an interdisciplinary approach, using the analysis of genetic sequence data to answer questions in fields as diverse as environmental remediation, cancer research, drug discovery, marine microbiology, and power generation (National Research Council (U.S.). Committee on Metagenomics: Challenges and Functional Applications, 2007).

Metagenomics is enabled by new laboratory methods, advances in sequencing technologies, and cutting edge information infrastructures. In the past, geneticists and genomicists had to isolate individual organisms and grow them in the laboratory in order to study their DNA. However, it has been estimated that less than 0.1% of the world's microorganisms are amenable to culturing in the laboratory. New techniques and technologies have been developed that make it possible to bypass this culturing step while significantly lowering the cost of DNA sequencing. These changes give scientists access to a wealth of genetic information from organisms that previously could not be studied. Metagenomics also makes it possible to ask new questions about the relationships among organisms and their relationship to their environment.

The field of metagenomics provides an interesting case study in part because of its rapid growth. Indeed, the term was only coined in 1998 (Handelsman, et al., 1998), and by mid-2005, nine major metagenomic sequencing projects had been completed (Chen & Pachter, 2005). Interest in these techniques is growing: for example, the Metagenomics 2008 conference attracted more than 250 participants, and the NIH is embarking on a major project to study the human microbiome [http://nihroadmap.nih.gov/hmp/]. Here we use the emergence of metagenomics to demonstrate some of the ways that the introduction of a new community of scientists with new research questions and new information needs can challenge existing information infrastructures.

2 Our Study

This research reports on an ongoing ethnographic study of the development of cyberinfrastructure to support metagenomics research. This study includes both an in-depth examination of one particular cyberinfrastructure development project, and a broad survey of information infrastructures serving metagenomics researchers. We have conducted thirty-three interviews with metagenomics researchers, computer scientists, bioinformaticists, and others involved in the development of metagenomics cyberinfrastructures. We have also conducted over 100 hours of formal and informal observation, including attending development meetings, laboratory meetings, workshops and conferences. Interview transcripts and field notes were analyzed using a grounded theory approach (Glaser & Strauss, 1967).

3 Cyberinfrastructures for DNA Sequence Data

Scientific cyberinfrastructures are distributed enterprises supported by advanced technological infrastructures such as supercomputers and high-speed networks. In the genetic sciences, scientists have long recognized the importance of sharing sequence data, and have developed significant infrastructures for doing so. GenBank, for example, has been collecting and distributing DNA sequence data since 1982 (National Center for Biotechnology Information). GenBank is only one of many infrastructures that provide storage of DNA sequence data and facilities for analyzing and visualizing that data.

Data in these databases is submitted by the scientists who conduct the DNA sequencing and analysis. The field has strong norms around data sharing, backed up by a commitment by journal publishers not to publish analyses of genetic data unless the data is made public and submitted to GenBank or other databases (Marshall, 2001). While the databases may have their own underlying architectures, data sharing among the scientists and databases is supported by a strong standard called FASTA, which specifies a uniform file format for representing sequences using individual letters to stand for amino acids (National Center for Biotechnology Information). Many of these systems also provide tools like the Basic Local Alignment Search Tool (BLAST) which allow scientists to compare new genetic sequences with those in the database (Altschul, et al., 1990).

4 New Questions for Old Infrastructures

Metagenomic analyses use the same sequence data that is used in other genetics-based fields, and tools like BLAST are still useful to compare new genetic sequences to sequences generated by other scientists. Metagenomicists need some of the same basic functionality provided by infrastructures like GenBank. But at the same time, these infrastructures become more valuable when the design of the tools and databases have a good fit to the scientific questions being asked (Bietz & Lee, 2009). In this section, we discuss the ways that new metagenomic questions challenge the design of cyberinfrastructure.

4.1 New Data and Tools

Metagenomics and its associated laboratory techniques bring a new set of data storage and analyses requirements to existing cyberinfrastructure. One of the consequences of new DNA sampling and sequencing technologies is that DNA sequencing has become relatively inexpensive. While sequencing costs were around $10 per base pair in 1990 (Powledge, 2003), today researchers pay a few cents per thousand base pairs. The amount of DNA sequence data being produced

is overwhelming, to the extent that data storage and computation requirements are outpacing Moore's law (Dooling, 2009).

In addition to simply having more data, metagenomics also assumes a different unit of analysis. Rather than focusing on the gene or even whole genome of an organism, metagenomicists work at the level of a community or population of microorganisms. Many existing sequence databases cannot easily represent this level of relationships among data.

One of the key focus areas in metagenomics is the relationship between microbes and the environment, but studying these relationships requires scientists to also collect contextual "metadata" that describe where the samples were found, including location, temperature, pH, etc. Most genetic and genomic databases were not designed to handle this level of data complexity.

Along with this new data, scientists need new tools to analyze and visualize the data. For example, a common question in marine metagenomics involves understanding how ocean temperature affects the diversity of the local microbiome. Not only would this require temperature data, but also the ability to query it, include it in analyses, and create visualizations around it. This kind of question would be almost impossible to answer with the data structures and tools provided in cyberinfrastructures created for traditional genetics and genomics researchers.

4.2 Recontextualizing Existing Data

Metagenomicists bring new data to existing infrastructures, but they also want to ask their new questions about old data. Often to ask a new question requires putting the old data into the new metagenomic context. For example, even if metadata were not stored in the database originally, there may be sources (like the publication record) that could be used to populate new fields in the database. However, reformatting data or retrospectively adding metadata are expensive tasks, especially when the work may need to be done again for the next group of scientists who pose a new question.

Another issue arises in that new metagenomic data may change the interpretation of legacy data. As metagenomic data is added at a phenomenal rate to theses database, the computational problems are becoming immense. One database developer told us:

> So you do need to go back from time to time and do all [the analyses] from scratch…. So the problem there is that we need to do periodic updates and periodic updates are every three months…. Now if new data is coming at an increasing pace, we are already at the point where even really big infrastructures and big computer clusters cannot really support all that.

Beyond these issues of computational power, scientists are also refining and expanding theory. In genetics and genomics, for example, scientists are finding that some prior assumptions about how genes operate, the role of "non-coding"

regions of DNA, and evolutionary processes can necessitate a reconsideration of old data and interpretations.

5 Difficulty of Requirements Prediction

One question that arises is why these systems were not designed originally to support these new questions. If we accept the history told by many metagenomics researchers, metagenomics is a "logical progression" from genetics and genomics, and these future needs could have been predicted.

> The concept was simple: Take seawater and capture all the microorganisms swimming in it on filters with microscopic pores, isolate the DNA from all the captured organisms simultaneously.... Rather than focusing on the hunt for one particular type of life, we would obtain a snapshot of the microbial diversity in a single drop of seawater-a genome of the ocean itself. This was, to me, a straightforward extension of work that had started with the EST method and led to the whole-genome shotgun approach, then the first genome of an organism in history, and then of course to the human genome. (Venter, 2007, p. 345)

While this version of the origin of metagenomics creates a compelling narrative, it does not recognize two important features of these scientific changes. First, as science has "progressed" through these phases, it has not left old questions behind. There are still scientists who are studying the functions of individual genes, and there are still scientists who are studying the genomes of individual organisms. Metagenomics has not supplanted these fields. In fact, it is essential for metagenomicists that research continues in genetics and genomics:

> It would help us tremendously in doing metagenomics if we had a wide range of reference genomes.... The NIH is funding 400 complete genomes of microbes that live in humans. And again, these are to give us standards and to allow us to interpret metagenomic data more rigorously. So first of all, as far as I'm concerned, we've only begun to sequence. We need to sequence - whole genome studies need to go on to expand the opportunities in studying evolution and getting many specific genes and models for human disease and for understanding biology.

Not only are genetic and genomic studies important for metagenomics, new metagenomic techniques are also changing the way geneticists and genomicists do their work. For example, shotgun sequencing not only allows for the sequencing of populations of microorganisms, it also makes it possible to sequence genomes from organisms that could not be cultured in a laboratory.

Secondly, the progression from genetics through genomics to metagenomics is logical only in retrospect. The development of metagenomics was by no means a foregone conclusion, and scientists found that they had to work hard to convince their peers that these techniques were valid. One scientist explained the difficulties she experienced finding a venue in which to publish her work this way:

> Not only has there been this distrust between the two fields, the genomics and the traditional fields—I think it's becoming more acceptable—but now metagenomics has come in too. So

> we're not just talking about sequencing entire genomes, we're talking about populations of genomes and defining what's there based solely on sequence similarities to those genomes.
>
> So what I've - I'm taking a huge leap here. I'm saying I have these 50,000 sequences. They're very distantly related to these sequences from [other] genomes. I know nothing about their physiology. I don't know what they infect. I don't know their reproductive lifecycle. I don't know anything about them. I'm just giving them a name based on the history of those sequences. So I think I'm taking an even farther leap.... And I think we try not to tread too heavily upon people's toes. We don't want people to think we're trying to take over their fields and that these approaches are the end all to the field.

Traditional approaches to identifying microbes rely on direct examination of microbes' physiology, pathogenesis, and reproduction, this scientist found that using only metagenomic techniques was not readily accepted by peer reviewers. Even though some scientists see metagenomics as a "straightforward extension" from earlier techniques, this new way of looking at microbial populations was not predicted by early geneticists and genomicists, the science is not without its detractors, and it is not entirely clear how these techniques will unfold into the future.

These observations highlight significant challenges for the development and maintenance of cyberinfrastructure. As science changes over time, scientists will need different things from cyberinfrastructures. While some research questions will persist, others will change and new research questions will be asked. A new science like metagenomics brings new questions and new communities of scholars with different ways of understanding the world. The requirements for information infrastructures develop and change as the science and communities change. Just as it is impossible to predict with any certainty how a scientific field will develop, it is equally impossible to predict all future information infrastructure requirements.

6 Infrastructure Adaptation

So far we have focused on the challenges that a new science can pose for an existing scientific information infrastructure. Determining the best methods to address these challenges remains an open question, but the metagenomics field provides examples of three different approaches.

One strategy that has been adopted has been to create work-arounds for existing infrastructures to adapt them to new uses and questions. For example, GenBank does not provide much support for the contextual metadata that is key to metagenomics approaches. Metagenomics researchers have begun to add metadata to free-text comment fields, sometimes using a "structured comment" that mimics a table of fields and data. This provides the benefit that it can be used immediately and without much disruption to the existing infrastructure, and allows the new metagenomics researchers to store contextual data without affecting how other geneticists or genomicists use the system. On the other hand, work-arounds like

these are often difficult to use, lack standardization, and often do not provide full integration of the new science.

A second approach involves modifying or extending existing infrastructures to support metagenomic data. This seems to be happening in systems like IMG (Markowitz, et al., 2008) and The SEED (Overbeek, et al., 2005), which have extended their systems to include new metagenomics tools and support for metagenomics data. While this provides greater integration than the work-arounds, there can still be a disconnect between legacy data and new tools.

A third approach, taken by projects like CAMERA (Seshadri, et al., 2007), creates new infrastructure from scratch specifically to support the new science. While this may provide the best fit to the scientific questions, it can also be a very expensive option, and can make it more difficult to use legacy data and tools. Splitting off from existing infrastructures may also reinforce the separation between communities that may benefit from greater interaction.

7 Conclusion

Scientific information infrastructures that persist over long time scales must respond to the emergence of new science. New science brings with it a new set of research questions, data and tools, scientific communities, and ways of understanding legacy data. There is a need for a deeper understanding of how developers of cyberinfrastructure can manage the evolution of user needs and requirements, and to understand when it is better to extend an existing information infrastructure and when it is necessary to create new infrastructure. The introduction of metagenomic approaches in molecular biology highlights the dynamic nature of both the human and technological aspects of cyberinfrastructure.

8 References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990): 'Basic local alignment search tool', Journal of Molecular Biology, vol. 215, no. 3, Oct 5, pp. 403-410.

Bietz, M. J., & Lee, C. P. (2009): 'Collaboration in metagenomics: Sequence databases and the organization of scientific work' ECSCW 2009: Proceedings of the Eleventh European Conference on Computer Supported Cooperative Work, London: Springer-Verlag.

Chen, K., & Pachter, L. (2005): 'Bioinformatics for whole-genome shotgun sequencing of microbial communities', PLoS Comput Biol, vol. 1, no. 2, Jul, pp. 106-112.

Dooling, D. (2009): Maximizing utility of genome sequence data. Paper presented at the The Biology of Genomes. from http://www.politigenomics.com/2009/04/cshl-biology-of-genomes-2009.html

Edwards, P. N., Jackson, S. J., Bowker, G. C., & Knobel, C. P. (2007): 'Understanding infrastructure: Dynamics, tensions, and design', from http://www.si.umich.edu/~pne/PDF/ui.pdf

Glaser, B. G., & Strauss, A. L. (1967): The Discovery of Grounded Theory: Strategies for Qualitative Research. New York: Aldine de Gruyter.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998): 'Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products', Chem Biol, vol. 5, no. 10, Oct, pp. R245-249.

Markowitz, V. M., Ivanova, N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D., et al. (2008): 'IMG/M: a data management and analysis system for metagenomes', Nucleic Acids Research, vol. 36, pp. D534-D538.

Marshall, E. (2001): 'Bermuda Rules: Community Spirit, With Teeth', Science, vol. 291, no. 5507, 16 February 2001, pp. 1192.

National Center for Biotechnology Information: 'FASTA Format Description', Retrieved 14 April, 2009, from http://www.ncbi.nlm.nih.gov/blast/fasta.shtml

National Center for Biotechnology Information (April 2, 2008): 'GenBank Overview', Retrieved February 23, 2009, from http://www.ncbi.nlm.nih.gov/Genbank/index.html

National Research Council (U.S.). Committee on Metagenomics: Challenges and Functional Applications (2007): New science of metagenomics: Revealing the secrets of our microbial planet. Washington, D. C.: National Academies Press.

Overbeek, R., Begley, T., Butler, R., Choudhuri, J., Chuang, H., Cohoon, M., et al. (2005): 'The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.', Nucleic Acids Research, vol. 33, no. 17, pp. 5691-5702.

Powledge, T. M. (2003): 'How many genomes are enough?', The Scientist, vol. 4, no. 1, pp.

Ribes, D., & Finholt, T. A. (2007): 'Tensions across the scales: Planning infrastructure for the long-term' Proceedings of the 2007 International ACM Conference on Supporting Group Work, New York: ACM, pp. 229-238.

Seshadri, R., Kravitz, S. A., Smarr, L., Gilna, P., & Frazier, M. (2007): 'CAMERA: A community resource for metagenomics', PLoS Biology, vol. 5, no. 3, pp. e75.

Venter, J. C. (2007): A Life Decoded: My Genome: My Life. New York: Viking.