

# Pre-application Form

Created Friday, February 21, 2014

Updated Monday, February 24, 2014

## Page 1

### Applicant Contact Information

Applicant Contact Information   First Name:	C. Titus
Applicant Contact Information   Last Name:	Brown
Applicant Contact Information   Primary Phone:	517-505-9237
Applicant Contact Information   Alternate Phone:	517-505-9237
Applicant Contact Information   Email:	ctb@msu.edu
Applicant Contact Information   Position Title:	Assistant Professor

### Institutional Affiliation (United States only):

Institutional Affiliation (United States only):   Institution name:	Michigan State University
Institutional Affiliation (United States only):   Department(s):	Microbiology & Molecular Genetics; Computer Science
Institutional Affiliation (United States only):   Street Address 1:	2215 Biomedical Physical Sciences
Institutional Affiliation (United States only):   Street Address 2:	(No response)
Institutional Affiliation (United States only):   Street Address 3:	(No response)
Institutional Affiliation (United States only):   City:	East Lansing
Institutional Affiliation (United States only):   State:	Michigan
Institutional Affiliation (United States only):   Zip Code:	48824

### With which fields do you most closely identify with?

Please select as many options as you would like.

- Biology
- Bioinformatics
- Computer Science
- Marine Microbiology and Oceanography
- Other, please specify....: Computational Science

### Influential Works

To assist the Foundation in the review of pre-applications, please list up to five references that you find most impactful in the development of Data Science. These may be book chapters, journal/conference papers, or URLs to online work (particularly code or data sets). These references are distinct from, although may somewhat overlap, items on your bio-sketch.

For books, please list author(s), title, year, ISBN, and chapter(s).

For journal/conference papers, please list author(s), title, journal/conference name, year, DOI (if available), and volume(issue):pages.

For online resources, please list author(s), title, brief description (not to exceed 50 words), and URL.

## Reference #1

Online Resource

## Reference #1

Online Resource

Author(s):	Fernando Perez and Brian Granger
Title:	IPython Notebook
Brief Description (50 words):	An electronic notebook application that enables reproducible data analysis; has completely transformed data analysis.
URL:	<a href="http://ipython.org/notebook.html">http://ipython.org/notebook.html</a>
GUID (e.g., DOI, ARC, etc.):	

## Reference #2

Online Resource

## Reference #2

Online Resource

Author(s):	Yihui Xie
Title:	knitr
Brief Description (50 words):	knitr enables literate data analysis in R, just like IPython Notebook enables it in Python.
URL:	<a href="http://yihui.name/knitr/">http://yihui.name/knitr/</a>
GUID (e.g., DOI, ARC, etc.):	

## Reference #3

Online Resource

## Reference #3

Online Resource

Author(s):	Unknown
Title:	GitHub
Brief Description (50 words):	GitHub has enabled collaboration around code and data in tremendously good web-like ways.

---

URL: <http://github.com/>

---

GUID (e.g., DOI, ARC, etc.):

## Reference #4

Books

## Reference #4

Book

---

Author: Tony Hey (Editor), Stewart Tansley (Editor), Kristin Tolle (Editor)

---

Title: The Fourth Paradigm: Data-Intensive Scientific Discovery

---

Year: 2009

---

ISBN #: 0982544200

---

Chapter(s):

## Reference #5

Online Resource

## Reference #5

Online Resource

---

Author(s): Unknown.

---

Title: Amazon Web Services

---

Brief Description (50 words): AWS is the pre-eminent and predominant infrastructure-as-a-service offering in cloud computing, and it has completely changed the way we talk about computing and data analysis.

---

URL: <http://aws.amazon.com/>

---

GUID (e.g., DOI, ARC, etc.):

## Major Accomplishments - C. Titus Brown

I have worked in many research fields over the last 20 years, including digital life, physical meteorology and climate measurements, developmental molecular biology, gene regulatory networks, genomics, evolutionary developmental biology, microbial ecology, and bioinformatics. The theme “better science through superior software” cuts across all of my research: this is the idea that better software development and computational methods can lead to faster and higher quality science. As part of this I mission have invested heavily in open source, open science, open data and open access.

On top of my research and my thematic interests, I have engaged in significant outreach, including regular long-form blogging, active use of Twitter for scientific conversations, and many educational activities that include workshops and development of open educational materials. A few overall highlights are:

- In 1993, I co-authored the first version of the Avida software platform; 20 years later, Avida is a major research platform in evolutionary modeling and education, used by  $\approx 20$  research groups and cited in  $\approx 100$  publications. This was a formative experience because I saw first-hand how engineering a reusable software system enabled research.
- In 1999, I joined Eric Davidson’s lab at Caltech to do my PhD in developmental molecular biology. In Eric’s lab I learned experimental molecular biology, developmental biology, and genomics. This was a remarkably tough transition, as I was already a reasonably experienced computational scientist and was no good at experiments for many years; I ultimately co-authored about a dozen papers from this lab. This experience convinced me that bio/computational collaboration is part of the future of biology.
- In 2010, I developed our first advanced workshop on Analyzing Next Generation Sequencing Data. This workshop has attracted over 500 applicants in 4 years, led to sustained NIH funding, and informed my approach to educating biologists in computation – an approach that has reinforced and driven many of my other activities, including Software Carpentry workshops and teaching graduate students. At the time I was discouraged by colleagues from spending so much time teaching and training, but the course has been a strong net positive for my career, and has also expanded the scope of my research dramatically.
- Since 2008, I’ve led my own research group at Michigan State University. With an excellent group of students and postdocs, I tackled the sequence data deluge; in the process, we built novel theoretical solutions, engineered them at scale, and applied them to real biology research problems. We have implemented our solutions in a widely used software package, khmer, and our solutions have been adopted by several other widely used assembler packages. Our approach relies on streaming algorithms and lossy compression of sequence data, and I believe it provides a theoretical foundation for scalable primary sequence analysis of sequencing data.
- I’ve blogged since 2006 (<http://ivory.idyll.org/blog/>) and maintained it through becoming a parent and a professor. I blog about topics in sequence analysis, open science, academia, education, and computing. The blog has become an important part of my research program, enabling me to participate in deep technical discussion as well as trial broad social perspectives on open science and academia. The blog attracted 150,000 page views and 80,000 unique visitors in 2013.

**Barriers Overcome:** I successfully switched fields multiple times, from evolutionary simulations to physical meteorology to molecular developmental biology to genomics to bioinformatics, and published in each of these areas. I took a faculty position split across two departments (CS and microbiology), neither of which I have been formally trained in; despite this, I have been successful in both. I established a research group in a field new to me, and have published and received many grants in this area. I have devoted substantial effort to controversial approaches, including investing in open science by publicly posting source code, research ideas, preprints, and grants, and my research flourished as a result. In retrospect, none of this should have worked as well as it did!

## Future Research Directions

My interests center on the DNA and RNA sequencing data deluge. My current 2-3 year plan is to follow through on our ultra-efficient streaming and online approaches; we have demonstrated the basic viability of these approaches and developed useful (and well-used) software, but additional work remains to be done on both the theoretical and practical aspects of streaming sequence analysis. I have been funded for much of this follow-through by the NIH. Another nascent effort in the lab is pushing an open science/data/source/cloud agenda in biological sequence analysis; we have developed open protocols for metagenome and transcriptome assembly that run in the cloud, and have now offered to analyze other people's data for free if they agree to open it after an embargo period. We hope to use this to drive the development of an open research ecosystem surrounding metagenome and transcriptome analysis.

**The greatest opportunity that has been enabled by cheap sequencing is the opportunity to understand the function of diverse organisms and diverse biological environments.** Here, we face many technical and scientific challenges: large amounts of inexpensive data, sequence from divergent and poorly studied organisms, insufficient architecture for storing and indexing the data, search and data mining algorithms that do not scale, a variety of metadata that must be correlated with the data, and a pervasive culture of data and source code secrecy. It is tough to address any one of these issues in isolation: for example, it is difficult to motivate data and source code sharing in the absence of good indexing and query systems, and even were sharing data endemic, we do not possess the technical capacity to analyze it all centrally. The innovative and holistic approaches needed to address these challenges are by their nature high risk and unlikely to be supported by traditional government funders.

To tackle these challenges **I propose to build a lightweight graph query system that can be used to “glue” databases and data sets together in a federated, bottom-up manner.** The design and evolution of this system will be driven by tackling a central research challenge in sequencing-era biology: high-throughput determination of the function of unannotated and/or unknown genes. We will build this system in close collaboration between biologically and computationally focused researchers, so that the computational goals and biological goals stay in sync.

The specific **biological focus** will be on two areas that have been particularly enabled by the sequencing glut. The first is microbial community function in environmental data sets, where correlative analysis between deep shotgun sequencing data and environmental metadata, together with metabolic modeling, can begin to enable entirely computational hypothesis refinement. The second area is evolutionary developmental biology, or “evo devo”, where inexpensive mRNAseq is allowing the exploration of gene content in many divergent phyla. Here we can study gene family evolution and connect it with the evolution of developmental body plans, neuronal functionality, and metazoan phylogeny. Our **computational focus** would be on building out a graph database overlay on top of federated databases, and integrating it with exploratory data analysis tools. Graph databases provide a simple abstraction for interconnecting databases and building queries across them. A basic implementation already exists – the pygr project – but it needs to be extended for greater scale and a wider range of data types.

The **impacts** of this project would be many. First, we will provide homology and functional predictions for millions of genes, which will accelerate research in metagenomics and evo-devo alike. Second, we hope to provide proof of concept that a federated graph database designed bottom-up can be effective. And third, we hope to nucleate the creation of additional query backends for the graph database, and publication of additional data sets, as we make cross-dataset queries more available. Any one of these would have a serious impact on data-intensive biology, and all three together could be transformative, in both a scientific and a cultural sense. As with my current research program, I will run training workshops that make use of these methods as a way of generating collaborations and identifying real-world problems worth tackling.

## C. Titus Brown

---

(As of February 2014.)

### PROFESSIONAL PREPARATION

**Reed College**, Portland, OR; Mathematics; B.A., 1997

**California Institute of Technology**, Davidson Lab (graduate student);  
Developmental Biology; PhD., 2007

**California Institute of Technology**, Bronner-Fraser Lab (postdoc);  
Developmental Biology and Bioinformatics; 2007-2008

### APPOINTMENTS

**Assistant Professor**, Microbiology & Molecular Genetics / Computer Science and Engineering  
Michigan State University, 2008-present.

### HONOURS AND AWARDS

Burroughs-Wellcome Fund Computational Biology Fellowship (1999-2004).  
Withrow Award for Teaching Excellence in Computer Science (2008-2009).  
Woods Hole Marine Biological Laboratory Summer Fellow (2013).  
Michigan State University / College of Natural Science Teacher-Scholar Award (2013).

### SYNERGISTIC ACTIVITIES

1. Personal/professional blog at: [ivory.idyll.org/blog/](http://ivory.idyll.org/blog/). A few selected posts: “Our approach to replication in computational science”, “Thoughts on Assemblathon 2”, “The future of khmer (2013)”. 150,000 visitors, 80,000 unique (Feb 2013-Feb 2014).
2. Course director, 2010-present, Next-Generation Sequence Analysis for Biologists, KBS, MSU. Open materials at [ged.msu.edu/angus/](http://ged.msu.edu/angus/). Over 500 applicants in four years.
3. iPlant Collaborative Scientific Advisory Board member. iPlant Collaborative is a large NSF BIO Cyberinfrastructure project focused on genomics and phenomics research.
4. Software Carpentry Advisory Board member.
5. Member of the Python Software Foundation.

### PAPERS AND PROJECTS

*Full publication list at: <http://scholar.google.com/citations?user=O4rYanMAAAAJ>*

*khmer: k-mer counting and filtering FTW*. Software project, at <http://github.com/ged-lab/khmer/>. Michael R. Crusoe, Greg Edverson, Jordan Fish, Adina Howe, Eric McDonald, Joshua Nahum, Kaben Nanohy, Jason Pell, Jared Simpson, C. S. Welcher, Qingpeng Zhang, and C. Titus Brown. BSD license. Estimated  $\approx$  200 users; 32 GitHub stars (93-99%ile); 56 GitHub forks (97-100%ile)

*Assembling large, complex environmental metagenomes*. Howe AC, Jansson J, Malfatti SA, Tringe SG, Tiedje JM, **Brown CT**. preprint arXiv:1212.2832. (2 cit.) In review, PNAS.

*A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data*. **Brown CT**, Howe AC, Zhang Q, Pyrkosz AB, Brom TH. preprint arXiv:1203.4802. (15 cit.)

*Scaling metagenome sequence assembly with probabilistic de Bruijn graphs*. Pell J, Hintze A, Canino-Koning R, Howe A, Tiedje JM, **Brown CT**. Proc Natl Acad Sci USA, published online before print July 30, 2012, doi: 10.1073/pnas.1121464109. (41 cit.)

*Exploring the future of bioinformatics data sharing and mining with Pygr and Worldbase* Lee C, Alekseyenko A, **Brown CT**. in *Proceedings of the 8th Python in Science conference (SciPy 2009)*, G Varoquaux, S van der Walt, J Millman (Eds.), pp. 62-67. (4 cit.)

OTHER WORKS

*Best practices for scientific computing.* Wilson GV et al. PLoS biology 12 (1), e1001745. (25 citations)

*Reproducible Bioinformatics Research for Biologists.* Preeyanon L, Pyrkosz AB, and Brown CT. Chapter in *Implementing Reproducible Computational Research*, V. Stodden and R. Peng, ed. Forthcoming in Dec 2013; (link)

*A genomic regulatory network for development.*

Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C, Otim O, **Brown CT**, Livi CB, Lee PY, Revilla R, Rust AG, Pan Z, Schilstra MJ, Clarke PJ, Arnone MI, Rowen L, Cameron RA, McClay DR, Hood L, Bolouri H. Science. 2002 Mar 1;295(5560):1669-78. PMID: 11872831. (1083 cit.)

*Earthshine observations of the earth's reflectance* P. R. Goode, J. Qiu, V. Yurchyshyn, J. Hickey, M.-C. Chu, E. Kolbe, C. T. Brown, and S. E. Koonin Geophys. Res. Lett. 28, 1671 (2001). (95 cit.)

*Evolutionary Learning in the 2D Artificial Life System "Avida"*

Adami C, Brown CT. Proc. of "Artificial Life IV", MIT Press, p. 377-381 (1994). (198 cit.)

COLLABORATORS

J. Barrick (UT Austin), J. Bell (MSU), H. Cheng (MSU), S. Goffredi (Caltech), C. Lee (UCLA), L. Mansfield (MSU), P.W. Sternberg (Caltech), B.J. Swalla (UW), J.M. Tiedje (MSU), S.G. Tringe (JGI), J.Jansson (JGI)., W. Warren (WUSTL), E. White (USU), G.V. Wilson (Mozilla).

Graduate advisor: Eric H. Davidson (Caltech); Postdoctoral: Marianne Bronner (Caltech)