LLNL-TR-407449

JGI Computing 5-Year Strategic Plan

D. A. Bader, T. S. Brettin, R. W. Cottingham, P. A. Folta, Y. Golder, S. K. Gregurick, M. E. Himmel, R. C. Mann, K. A. Remington, T. R. Slezak

October 2, 2008

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

JGI Computing
5-Year Strategic Plan

9/30/08

David A. Bader, Thomas Brettin, Robert W. Cottingham, Peg Folta, Yakov Golder, Susan Gregurick, Mike Himmel, Reinhold C. Mann, Karin Remington, Thomas Slezak

A broad range of scientific goals and a similarly diverse set of consumers drive the informatics requirements and computing needs of the JGI. The scope of work in this area encompasses not only the informatics and analysis pipelines in support of the PGF sequence production, but also the integration of data from a variety of sources and sophisticated large scale analyses led by investigators within JGI and driven by the user science community. In laying out a forward looking strategy, the full range of these activities need to be examined together to build a comprehensive program that will serve as a catalyst for the DOE research community.

The science landscape envisioned in the overall strategic plan calls for significantly increasing the throughput of microbial genomes sequenced to cover their phylogenetic space and building a set of finished reference plant genomes to enable DOE relevant science. Additionally, the established impact of microbial communities on global energy cycles and their potential in remediation endeavors, warrant building upon JGI's established expertise in metagenomic analysis. Not only is each of these program areas relevant and exciting in their own right, but they also can and should be undertaken in a way that allows synthesis across domains (e.g. utilize knowledge from sequence of plants and the soil from which they are grown). Both dramatic increases in the scale of genomic data collection and the synergistic potential of integrating data across domains will demand new strategies in the informatics pipeline within the JGI and in the facility's approach to computational analysis and user access to the data in aggregated form.

In addition to a robust and scalable informatics infrastructure, fulfilling the strategic science goals of the JGI will require ongoing investment in usability of the data, to ensure that the data collected will be used to maximal effect. It must be recognized that "usability" will have a different appearance depending on the specific user base, and the JGI has several distinct classes of users it must enable to be successful. For some, rapid and convenient dissemination of the sequence data will be sufficient to enable their external research. For others, JGI hosted analysis tools and collaborative environments will be required to catalyze individual or team research. Finally, and significantly, there are genomic scientists within the JGI, often working closely with external collaborators, who rely on the ability to devise project-dependent and often very large scale customized analyses that result in publicly available tools. A successful strategy will require effort to satisfy each of these user classes, and careful attention to economies of software reuse and extensibility.

There are only a handful of sequencing facilities worldwide that operate at the scale of the JGI's Production Genomics Facility, and these are devoted almost entirely to sequencing driven by biomedical applications. The PGF therefore fulfills a unique and vital role as a resource for genomic studies of DOE relevance. Like the other large-scale facilities, JGI has been carefully following the development of "next-generation" sequencing technologies, and clearly must continue to refresh its instrumentation as advances are made. Critical to advances in sequencing technology are the computational infrastructure advances that are required to turn raw sequence into quality data. This is one area where JGI can leverage the broader sequencing community's investment in technology development, adopting the best practices and software for sequence processing and assembly. JGI can add unique value by further developing annotation pipelines and tools that serve to build an integrated framework where the Institute's complementary science components can be viewed in a larger "systems" perspective than is currently possible.

As technology, tools, and infrastructure advance, JGI is uniquely positioned in its ability complement core PGF expertise with a diverse set of capabilities provided by partners within the broader community that forms the Institute. The coming decade promises radical change in the field, and the ability to quickly recognize developing areas and nimbly build appropriate partner teams will be vital to maximally capitalize on the growing base of data collection capabilities. This dynamic team science environment will require an underlying computational environment that can accommodate innovative ideas and processes. This will be key to the success of the scientific goals of the JGI.

Current Computing at the JGI

Current the JGI computing efforts are distributed among partner labs as well as across the PGF organization. Each computing effort is staffed locally with a few exceptions, and when possible these efforts use common JGI tools. One such example is the extensive use of the IMG by the sequence analysis team at LANL. Still, these distributed activities have resulted in local development teams, data management strategies, compute systems, analysis pipelines, and user interface / environment.

The Eukaryotic Annotation Pipeline developed and maintained by the PGF Informatics Department, supports animal, plant, algae, and fungi annotation. Through a web-based portal, user communities have access to tools for manual curation of gene structure and function, with access to evidence based on results from large-scale data analysis integrated with high-throughput experimental data, sometimes provided by the external users. The portal has helped build strong user communities, often supporting curation “jamborees” that gather community users together for group curation. It was used to annotate all eukaryotic genomes sequenced at the JGI to date and is responsible for over 80% of web-visitors to JGI. It is unclear what role this system will play in future annotations of plants.

Phytozome is a joint project of the Department of Energy's JGI and the University of California-Berkeley's Center for Integrative Genomics. Developing and using comparative genomics analysis tools, the project is focusing on studies amongst green plants. These studies drive the creation of software in order to better understand the mechanisms of evolution. The Metazome project is similarly is driving the generation of comparative genomics tools by looking at the proteomes of metazoans.

The PGF Genome Biology Program continues active development of computational systems to support microbial sequence analysis that include large-scale data integration, analysis, annotation, and visualization. A manual "cleanup" of each new genome has been credited for high quality of results. Included in this program is: the development of an Integrated Microbial Genome (IMG) system for interpreting newly sequenced genomes and for the analysis of existing genomic sequence data on a comparative level; the development of the IMG/M to analyze the functional capabilities of microbial communities; and support of the IMG Educational Site (IMG/EDU). Issues of scaling are a primary concern. Access to high performance computing at PNNL has been used to augment the PGF systems for re-computing new versions of these systems.

The computational activities of the Phylogenomics Group are done in collaboration with UC Davis and include the development of methods that use phylogenetic trees as a tool in genome sequence analysis.

Partner lab ORNL maintains a mature automated microbial analysis and annotation pipeline system that analyzes all microbial genomes sequenced at the JGI. They maintain a significant local software and hardware infrastructure.

Partner labs LANL and Hudson-Alpha have developed extensive automated finishing pipelines that include new sequencing instruments and require significant software and hardware infrastructure. These efforts are managed locally at the partner institutions, and for the most part resulting software is not re-usable outside the local environment. A few exceptions exist where software has been shared among the partnership.

The PGF production informatics team resides in the Production Sequencing Department and is responsible for the software systems needed to execute high-throughput sequencing. The systems involve information tracking, commonly referred to as LIMS, data processing, data management, sequence project tracking and quality control. These systems support traditional Sanger sequencing as well as the next generation Roche 454 and Illumina Solexa sequencing platforms, which required new computing and storage hardware to support the increased data flows. The senior management at the JGI has recognized staffing issues within the team, primarily due to the new software and hardware required to support the new sequencing technologies. Performance testing of new storage devices, acquisition of new computing resources, and significant increases in data archival resources are requiring different skill profiles.

The PGF Informatics department is also responsible for the JGI web site and the PGF computational systems and data repositories that support the groups above. The department also develops project management tools that enable the User Program to manage sequencing projects as they make their way through the JGI. In addition to the PGF on site computing and storage systems, they are utilizing high performance storage for data archival at NERSC.

Computing Goals

The JGI computational goals are driven by the scientific goals of the associated programs and ultimately DOE mission relevant users. The goals each of the JGI programs have put forward include significant scaling of current analytical capabilities and development of new computational capabilities, while most call for an increase in computing resources to meet their goals. Several programs have synergistic goals. The computing goals take into account the computational needs of each program and identification of areas where shared computing infrastructure, tools, and expertise could be leveraged.

The Computing goals listed below are intended to extend the JGI's foundation of user-focused analytical capabilities while addressing the scaling issues brought on by the increase and complexity of data and the analytical needs of the JGI users.

1. **Advance annotation automation and incorporate additional evidence from newly developed computational prediction methods and alternative experimental results.** The details of implementing this goal vary across JGI science programs and specifics are addressed below. To the greatest extent possible the JGI should use common technologies, tools, personnel, data, and resources to achieve this goal. Common previsions for updating of results periodically as new genomes become available should be developed, as well as common methods to quantify the quality of annotations.

   The Plant Genome Program (PGP) will require the development of new high throughput tools to identify and characterize the functional elements in plant genomes (e.g. PlantENCODE). These tools will rely on the integration of new datasets that will come from the newly sequenced reference genomes, re-sequencing, and various "omics" information. Access to the resulting annotation and its associated evidence should be made available to the varied JGI user base according to their needs.

   Automated microbial annotation is more mature and several pipelines exist within the JGI. Migrating toward a common extensible system that can meet the varied needs of all JGI programs, as well as the broader JGI user base that support DOE Programs, should be considered. This could be accomplished by setting JGI annotation standards that could be met by multiple partners, allowing the computational workload to be partitioned across the resource base without resulting in overlapping annotation systems. Annotation needs to evolve from being a project-specific function to one that is focused globally across the JGI mission-space to the extent possible.

   Metagenomic annotation is well established at the JGI and has an active user community. While the IMG suite of tools is considered a leader in this area, metagenomics analysis is widely viewed as primitive and an active area of research. Collaboration on novel analytical methods should be encouraged to address the deconvolution and identification of genomic annotation and individual species within the community. Multiple issues of scale are of major concern here, particularly in the area of manual genome “cleanup”. Adoption of new methods and continued investment in scaling issues and automation should be a high priority.

Within the Phylogenetics Group, we see opportunities for a tighter integration with the IMG system, providing a more powerful, decentralized global annotation functionality.

While the JGI plans to increase their analytical capabilities, it is unclear if they intend to extend their current user base to include the broader DOE mission relevant public by providing support for community analysis and annotation. The microbial program strategic plan encourages “Community/wikified annotation curation” as a desired future direction. The PGP would benefit by expanding the community based for their reference genomes. The JGI could become a leader in establishing a social networking facility for researchers by providing open access to the datasets, annotation, and tools for analysis that enable basic annotation and comparative genomics. The JGI should evaluate if there is a viable model for partners like Google and Oracle to be interested in providing computing and database expertise in a way that could revolutionize how biological information is stored, processed, integrated, and shared. And a social networking interface (e.g. “Genebook?”) could be layered with this to provide an improved way for worldwide science collaborations to be formed and managed? The JGI and its partners should explore leadership ideas that could change the way science is done and position the JGI as a community enabler.

2. Scale computational infrastructure and algorithms to support all JGI needs. Rapid increases in data volume and complexity generated by advances in sequencing technology and inclusion of high throughput experimental results make it imperative that the JGI consider architecture appropriate computing and software tools.

The DOE has been developing ultra-scale computing and data management systems and capabilities that could be accessed and applied to help achieve the JGI objectives. These systems include the large number of processors with small memory per node machines as well as capabilities that support large data-intensive applications that require large memories do not decompose well onto small-memory clusters. The JGI should consider the use of commodity “cloud” computing and storage, and explore partnerships with commercial entities to advance the JGI goals in the most effective way. By not attempting to own all its computing resources, the JGI will be freer to respond in a nimble fashion to take advantage of novel, yet appropriate advances in computing hardware as they becomes available from service providers. Computational jobs that support the JGI’s analysis pipelines could be very well suited to a “cloud” infrastructure, whereas time-critical jobs such as base calling and quality control will likely need to be done locally. The use of cloud computing for annotation and comparative genomics studies will undoubtedly be a reality in the near future. JGI is in a position to drive standards for biological data processing. A byproduct of such standards will be better integration of the JGI partners into a common computing framework for biological research.

New algorithms and tools will need to be developed or obtained to support the increase of data and alternate computing architectures. Primary areas for advancement include large genome assembly, metagenomic analysis, SNP phenotype modeling, genotype-to-phenotype association, phylogenetic reconstruction, and large-scale comparative genomics. Application of high-performance strategies on appropriate computing architectures can be used to address these scaling issues.

The creation of a computer hardware test bed is critical at this time. A test bed would focus on new hardware such as the emerging high memory machines, hybrid compute platforms (cells, GPUs, FPGAs), and advances in parallel and distributed file systems. Scaling the production line of next generation sequencers will require us to rethink our choices and utilization of computer hardware platforms. This genomics test bed will also facilitate the use of external "cloud" computing by allowing hardware and software components to be tested and debugged locally at a small scale and then ported to external clouds for scaling to production needs.

Similar scaling is also needed for data management, information integration, and visualization capabilities. The JGI cannot afford to create all the required software but must apply business logic appropriately for each make/buy decision. This should be a JGI-wide decision as opposed to a project-local selection.

The JGI is considering analysis only projects, managed by a CSP-like program. Metrics to match the number and scope of projects to be supported with available resources will be needed. Extending the current project management tools should provide this capability. Novel approaches to fund the required computing resources for analysis-only projects should also be considered.

3. **The JGI should take a systems approach to create the most efficient and effective computing enterprise while utilizing the varied resources of the JGI.** The JGI has been successful in forming independent focused teams to address the computing needs of different user communities and projects, but the ability to scale along these lines will come at a high cost. Analyzing the varying requirements of the JGI Programs and users against the available resources across the JGI will provide opportunities to strengthen, simplify, and extend their computing capabilities. Developing an "informatics enterprise vision and plan" that is more than the sum of the parts of current individual projects can be achieved by selective use of software reuse, functional consolidation and/or partitioning, and redundancy. Growing businesses across the world are grappling with the same issues and in study after study these best practices have been proven. (e.g. http://harvardbusinessonline.hbsp.harvard.edu/hbsp/hbr/articles/article.jsp?ml_action=get-article&ml_issueid=BR0807&articleID=R0807J&pageNumber=1&ml_subscriber=true&uid=24499457&aid=R0807J&rid=24587452&eom=1)) There is no reason that this approach will not work in a large-scale production sequencing setting, as demonstrated by the Broad Institute's informatics program.

Optimization at the institutional level will provide the opportunity for the JGI computing teams to work together to achieve the greater goals of the JGI. Tapping the best practices and expertise of all the computing groups at the JGI will result in a superior outcome. Software reuse can help create a scalable, extensible s/w base that cuts across programs to increase maintainability of software systems as a whole. It can be used to create common robust systems that can be distributed across the organization to avoid single points of failure. In other instances elimination of redundancy could free up resources that can address other needs. The JGI Informatics Dept. is beginning to address some of these basic issues within the Project Management Office at the PGF by eliminating redundancy of project status in various databases and defining a "user" that all components will be using when citing metrics.

One specific area to target is the new Synthetic Biology program. The JGI should leverage and expand existing JGI infrastructure, tools and expertise to support their new synthetic production of biological parts. This would be expected to include a database to support the parts registry and repository, and the interfaces for parts characterization and input that will derive from existing annotation from the Plant, Microbe, Metagenomics and Resequencing Programs. In addition, a system for the design and assembly of specific synthetic organisms may be needed, first electronically and then driving the assembly line.

4. The JGI should collaborate with the broader bioinformatics community on the development of common methods to manage, use, and view the integration of observed variation within the associated reference sequence annotation to provide a basis for understanding function and functional changes.
Currently a significant portion of the JGI's sequencing capability is being applied to genomic resequencing efforts relevant to bioengery. The Plant and Microbial strategic plans indicated the need to understand variation as a key component. Development of a Synthetic Biology program will rely heavily on the capture and display of variation as well. Incorporation of previously developed human-based tools could be an initial start, however it is not clear they are directly applicable and these tools are not considered mature.

5. **The JGI should maintain a robust automated production informatics system that supports the evolving state-of-the-art sequencing technology as one of its highest priorities.** Sequence production will remain at the heart of the JGI and the informatics to support it is of primary importance. The JGI, along with the rest of the world, is adapting to the latest sequencing technology advancements. These latest technologies are still in their infancy and are evolving rapidly. They bring a significant change to the informatics infrastructure needed to support them, primarily in terms of scaling, automation, and data quality assessment. The coming years will likely bring even newer game changing technologies that will undoubtedly alter the informatics landscape again.

Production Sequencing Informatics should be tightly coupled with both the Sequence Production Department, as well as the New Technologies organization that initially tests these technologies. Informatics development and test environments should be created to address the impacts of the new technologies and ensure support for the production of accurate, high quality sequence data that is readily accessible to projects and users long before the systems move into production. The impacts on the computing infrastructure can play a meaningful role in the determination of technology adoption and schedule. It is vital that investments in the evolution of the production sequencing software infrastructure keep pace with the adoption of new sequencing technologies.

6. **The JGI should be an active participant in the rapidly growing spectrum of community standards** (see: www.mibbi.org - MIBBI: Minimum Information for Biological and Biomedical Investigations). Of particular importance to the JGI will be the meta data associated with metagenomic data. Standards are vital because the sheer data volume will soon make it infeasible for everybody to store local copies of all information they wish to use. It is also important because users wish to query across the idiosyncrasies of multiple current silos of information at independent sites. At the heart of these effort will be a program to ensure that the JGI is an integral part of the Genomics:GTL data and computing environment (http://genomicsgtl.energy.gov/compbio/index.shtml). This environment includes the capabilities and resources created by the Bioenergy Research Centers, other ongoing GTL consortia, and the emerging dynamic GTL Knowledgebase.

   The PGP informatics should strive to work in close cooperation with established efforts in NSF's Plant Genome Program, such as the PlantGDB (http://www.plantgdb.org/) and many others to ensure that appropriate plant community standards are used or developed in concert with the wider plant community.

Realizing the potential for leadership in computing and informatics within this five-year planning period will require an evaluation of how this effort should be led and managed. The necessary expansion from organism-centric projects and “jamboree” community collaborations to truly cross-cutting and world-leading capabilities in JGI data analytics, including effective integration and coordination with the rest of the GTL program and other efforts outside the DOE, will require leadership and management support singularly focused on these goals and committed to their accomplishment.

JGI Computing Roadmap

There has never been an attempt to develop a strategic plan that encompassed all JGI computing needs. We encourage the JGI to take the time now to develop such a plan by involving key members of the PGF informatics department, each science program, representative users, and each partner lab.

Near-term, within 1 year. Much of the first year will focus on analysis and planning, while meeting current milestones.

Annotation Automation

- Determine how to organize annotation as a global service across all JGI missions
    - Hold an internal workshop that reviews each individual annotation effort, and analyzes how the collection as a whole can be optimized in terms of cost, quality, and functionality
        - Initiate development of Common extensible automated microbial annotation
    - Establish working group to assess and improve the quality of annotation
    - Determine if community annotation will be a goal for the JGI, launch a feasibility study to assist the decision that evaluates partnerships with Oracle, Google, etc.

Scaling Infrastructure

- Determine the options for adopting cloud computing
- Work with DOE to determine how best to utilize their HPC capabilities, both in terms of access to cycles and support for porting applications to those systems
- Determine the information integration infrastructure needed to accomplish the JGI's mission goals
- Determine computational needs for standard analytical capabilities that will assist the PMO in managing analysis only projects

Enterprise solutions

- Establish an Internal Informatics Working Group with representatives of all major computing team, to develop and execute a detailed JGI Computing Plan, enable networking and sharing of best practices (see annotation workshop above) and have a voice in decision making on computing level topics.
    - Develop detailed informatics plans and annual goals for each of the major science, user and production activities based on estimates of technology change, sequence produced, usage, and analysis measures driving these activities
- Establish External Informatics Advisory Committee to review detailed plans, assess accomplishments, and update strategic goals and roadmap
- Determine the minimal informatics support required for a viable synthetic biology pilot
- Determine the investments required for scalable algorithms for handling short-read assembly and metagenomic analysis needs (build or buy decisions)

Variation

- Form a working group of those involved in variation analysis to share current and planned methods for analysis, management, and visualization of variation data. Leveraging the group to understand what exists externally and what should be developed internally

Production Informatics

- Grow team to support current, but prepare for new generations of sequencing technologies
- In collaboration with the Informatics Dept., create a test bed to prepare for new technologies

Standards

- Contribute to establishment of DOE open source software policy similar to open data policy
- Identify possible standards for JGI consideration

Mid-term, 1-3 years The focus on execution of the plans from the previous years.

Annotation Automation

- Implementation of common extensible automated microbial annotation
- Introduce "plug and play" annotation components
- User and data access systems provide annotation quality and links to evidence and original sources
- Begin implementing community annotation

Scaling Infrastructure

- Build Phase 1 of information integration infrastructure
- Implement cloud computing and infrastructure to access DOE HPC
- Develop or import key algorithm needs to address scaling issues

Enterprise solutions

- Pilot the use of social networking techniques to organize the JGI communities around new paradigms of data access and interaction
- Extend informatics infrastructure to demonstrate computationally supported experimental discovery driving synthetic biology success

Variation

- Measurable improvement in annotation quality

Production Informatics

- Pilot next generation systems to support emerging sequence technologies

Standards

- Implement agreed upon standards

Long-term, 3-5 years. Focus is completion and assessment of planned systems and development of next 5 year plan

Annotation Automation

- Asses success of JGI's global service annotation efforts implemented over the last 1-4 years, make adjustments as necessary to address current bottlenecks

Scaling Infrastructure

- Build phase 2 of information integration infrastructure
- Assess success of cloud and DOE HPC computing, make adjustments as necessary

Enterprise solutions

- If successful, extend social networking techniques to support broader DOE user base
- Extend the informatics infrastructure needed for a scalable synthetic biology program
- Development of next 5 year plan

Variation

- Develop of adopt advanced methods to address analysis, management, and visualization of variation data

Production Informatics

- Implement next generation systems to support emerging sequence technologies

Standards

- Drive standards within DOE community in key areas