Report on NSF Workshop on Future Computing Platforms to Accelerate Next-Gen Sequencing (NGS) Applications  
Boston, MA, May 19th 2013

PIs & Co-PIs: Viktor Prasanna, Yogesh Simmhan (USC); David Bader, Jason Riedy, Richard Vuduc (GATech); Manish Parashar (Rutgers)  
Project Manager: Shel Swenson (USC & GATech)

This report is the outcome of the the Workshop on Future Computing Platforms to Accelerate Next-Gen Sequencing (NGS) Applications, held in Cambridge, MA on May 19th 2013, collocated with IPDPS 2013, and supported by the National Science Foundation (NSF) as part of a Scientific Software Innovation Institutes (S2I2) conceptualization project [NSF 11-589].

Motivation and Objectives

Solving scientific grand challenges requires effective use of cyberinfrastructure (CI). Recent advances in computing technology can dramatically accelerate innovation to solve complex problems of societal importance. These Future Computing Platforms (FCP) include Field Programmable Gate Arrays (FPGAs), General Purpose Graphics Processing Units (GPGPUs), multi-core and -threaded processors, and Cloud computing platforms. These emerging, potentially disruptive, platforms assist not only traditional high-performance computing (HPC)-based science but also data-intensive science in such areas as biological science, social science, and security.

The “Software Infrastructure for Accelerating Grand Challenge Science with Future Computing Platforms” project (future-compute.usc.edu), is a multi-institutional effort focused on supporting sustainable software for data-intensive and inter-disciplinary problems that may be enabled by graph algorithms. We target biological sciences, cybersecurity, and social networking based on their under-served needs and demonstrated possibilities. Applications made possible by high-throughput sequencing technologies is one of our primary target application areas in this conceptualization phase.

The workshop invited domain scientists from the NGS-enabled application areas of biology and computer science researchers. The workshop included more than a dozen interdisciplinary scientist from biology and CS including representatives from iPlant, Galaxy, and BigData for NGS applications, as well as designers and developers of several widely used NGS-enabled software applications. The aim of this workshop was to identify software infrastructure needs within the NGS applications community and develop a plan of action for our proposed center in this area. The workshop had three main objectives:

- Identify specific grand challenge problems in this area with the best opportunities for future computing platforms.
- Determine how graph algorithms and analytics can best be leveraged to advance NGS applications, and identify algorithms of broad use and high-impact for NGS to accelerate on FCP.
- Identify support mechanisms an S2I2 should provide domain researchers to accelerate scientific progress.
Workshop Activities and Outcomes

The workshop began with an overview of the potential of future platforms, followed by invited participant presentations of candidate NGS applications (including genome mapping and assembly, transcriptome assembly, haplotype assembly, and metagenomic clustering and assembly). Participants decided on two main areas of focus (assembly and systems biology) for the afternoon break out session where they would explore in more depth how graph abstractions and FCP could accelerate progress in specific NGS-enabled applications. The afternoon breakout sessions, produced potential use cases, identified what an FCP center could do that would help domain researchers utilize accelerator technologies, and other applications or groups could benefit from these efforts.

The insights gained from the afternoon break out session can be summarized as follows:

- Participants identified use cases (de novo transcriptome assembly, polyploidy assembly, and systems biology)
  - which are important NGS-enabled applications,
  - for which accelerated algorithms in these applications would benefit a large community of users,
  - and for which graph problems are a critical component.

- Participants agreed that for accelerated algorithms to be broadly used in the BIO community, outreach and training will be key. (The problems change rapidly and there is a more diverse community of users – both of which necessitate tighter integration with domain.)

During the break out session, one group discussed a variety of assembly problems, including transcriptome and polyploidy assembly, all of which can involve constructing and analyzing large graphs. Assembly for polyploid organisms, as of the time of this workshop, is a relatively unexplored problem. Creating efficient and accurate polyploid assembly software would mean addressing challenges presented by heterozygosity, transposon repeats, and a more complex error-correction problem. Such software would benefit a broad community of users particularly those studying flowering plants (including important food and biofuel crops) where polyploidy is common. We present the de novo transcriptome assembly use case below as an example of the potential impact of an FCP center on NGS-enabled applications and of potential synergies with other current efforts in this area.

Another breakout group discussed NGS-enabled applications in systems biology, where the graphs are often used to model the interactions between components of a biological system. Such networks can be used to study the effect of genetic mutations, how these systems work under different environmental conditions, and alternative pathways for producing particular compounds. For example, by modifying a pathway, can we improve vitamin content or yield in critical food crop organisms? Or could the insertion of some gene into plants increase resistance to pests or increase their utility in carbon sequestration?

The FCP team presented an overview of a proposed model for center organization and management (e.g. scope, potential activities, resource management, and measures of success), and solicited feedback from participants as to what features would best support progress in NGS-enabled applications. Participants agreed that for accelerated algorithms to be broadly used in the BIO community, outreach and training will be key. Further, that the primary target for these interactions should be software developers who were intern tightly connected with the bio community. By creating and supporting a strong user (developer) community the software and incorporation of FCP will become self sustaining. Suggestions as to how to maintain this developer community is to bring them together on a regular basis via, for example, yearly training workshops and hackathons.
Use case: de novo transcriptome assembly

Until recently, transcriptome assembly from RNA-expression data required mapping sequence reads to the genome of the organism itself or that of a closely related organism. However, assembling transcriptomes directly from sequence reads without the use of a reference genome (de novo assembly) is of critical importance as the cost of high-throughput sequencing techniques diminishes expanding sequencing efforts to an ever broader array of organisms. The current methods for de novo assembly involve several graph problems including graph building, path finding, and path compression all of which can be computationally expensive. Accelerating these approaches using FCP, e.g. massively parallel platforms designed to handle graph problems, could benefit a large and diverse community. For example, those studying the resistance of noxious weeds to herbicides\(^1\), efforts to increase the nutritional content of non-model plants\(^2\), and humanitarian projects aimed at increasing the productivity of food crops\(^3\). In general, such advances could impact any group studying gene expression in non-model organisms.

Assemblers which could be accelerated include Trinity, Velvet/Oasis, Trans-Abys, and SOAP. The majority of participants agreed that Trinity was the currently the best target for development of FCP implementations, but given the rate at which assemblers come in and out of fashion, development of extensible implementations of the core graph algorithms is key. This relatively fast turnover in cutting edge techniques together with a diverse community of NGS-enabled application areas also requires tighter integration of FCP technology experts and algorithm developers with the bioinformatics community. Efforts like the Galaxy Project and iPlant Collaborative could be instrumental in connecting domain users with accelerated applications, and participants recommend that a FCP center pursue these potential synergies.

---

\(^1\) [1KP project, www.onekp.com/agriculture.html](http://www.onekp.com/agriculture.html)

\(^2\) e.g.,


\(^3\) C4 rice project, [http://c4rice.irri.org](http://c4rice.irri.org)
Appendix A: Workshop Agenda

9:30 - 10:00 Set-up (load presentations), coffee and tea
10:00 - 10:15 Welcome and Introductions (Viktor Prasanna)
10:15 - 10:45 FCP Presentation (David Bader)
10:45 - 11:45 Participant Presentations (Facilitator: Jason Riedy)
11:45 - 12:30 Identify use-cases and goals for break-out sessions (Facilitator: Shel Swenson)
   Use-cases will be specific data-intensive problems with the best opportunities for graph abstractions and future computing platforms – ideally algorithms of broad use and high-impact for NGS to accelerate on FCP.
12:30 - 1:30 Lunch (provided)
1:30 - 3:00 Break out sessions (Facilitators: Shel Swenson and Yogesh Simmhan) For each problem identified at the end of the morning session, determine:
   * What an FCP center could do that would help
   * Why a center vs additions to existing efforts?
   * Potential timeline
   * What other applications may benefit from these efforts?
   * Outline a report
3:00 - 3:20 Break (snacks and refreshments provided)
3:20 - 4:30 Discuss proposed center model (Facilitator: Yogesh Simmhan)
4:30 - 5:00 NSF Report Generation (Viktor Prasanna)
## Appendix B: Participant List

<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
<th>NGS-enabled Applications Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Srinivas Aluru</td>
<td>Iowa State University</td>
<td>PI $2M NSF/NIH BigData for NGS applications</td>
</tr>
<tr>
<td>Mark Borodovsky</td>
<td>Georgia Institute of Technology and Emory University</td>
<td>NGS applications software design e.g., GeneMark gene prediction software suite</td>
</tr>
<tr>
<td>Thomas Brutnell</td>
<td>Donald Danforth Plant Science Center</td>
<td>iPlant Ultra Hightthroughput Sequence Pipeline lead</td>
</tr>
<tr>
<td>Ananth Kalyanaraman</td>
<td>Washington State University</td>
<td>HPC NGS applications software design/development</td>
</tr>
<tr>
<td>Ben Langmead</td>
<td>Johns Hopkins University</td>
<td>NGS applications software design/development, e.g., Bowtie, Bowtie2, Crossbow</td>
</tr>
<tr>
<td>Ion Mandoiu</td>
<td>University of Connecticut</td>
<td>NGS applications software design e.g., ViSpA viral assembler</td>
</tr>
<tr>
<td>James Taylor</td>
<td>Emory University</td>
<td>Galaxy Project PI (web-based platform for bioinformatics research more than 80,000 users)</td>
</tr>
<tr>
<td>Jaroslaw Zola</td>
<td>Rutgers University</td>
<td>CoPI $2M NSF/NIH BigData for NGS applications</td>
</tr>
<tr>
<td>David A. Bader</td>
<td>Georgia Institute of Technology</td>
<td>FCP S2I2 PI</td>
</tr>
<tr>
<td>Manish Parashar</td>
<td>Rutgers University</td>
<td>FCP S2I2 PI</td>
</tr>
<tr>
<td>Viktor Prasanna</td>
<td>University of Southern California</td>
<td>FCP S2I2 PI</td>
</tr>
<tr>
<td>Jason Riedy</td>
<td>Georgia Institute of Technology</td>
<td>FCP S2I2 CoPI</td>
</tr>
<tr>
<td>Yogesh Simmhan</td>
<td>University of Southern California</td>
<td>FCP S2I2 CoPI</td>
</tr>
<tr>
<td>Shel Swenson</td>
<td>University of Southern California and Georgia Institute of Technology</td>
<td>FCP S2I2 Project Manager</td>
</tr>
</tbody>
</table>