AM7: ISMB2006 Tutorial

Title: Exploring Computational Biology with a Massively Parallel High Performance Computing Environment

Topic Area:

- Sequence Analysis 30%
- Transcriptomics 10%
- Systems Biology (including Pathways and Networks) 20%
- Other: Computing Systems and Parallel Programming 40%

Main Presenter:

- Dr.
- Kirk E. Jordan
- IBM
- 1 Rogers Street
  Cambridge, MA 02142
- kjordan@us.ibm.com
- 617-693-4581
- 617-693-5532
- http://www-3.ibm.com/software/info/university/people/kjordan.html

- Professor 2 years; Adjunct Professor 6 years; Developed numerous application oriented tutorials; Presented numerous invited talks especially in the Life Sciences; Led many workshops averaging 4 per year for last 15 years.
- **Previous Tutorials:**
  1. A Massively Parallel High Performance Computing Environment for Computational Biology, ISMB, Detroit, MI, 2005 – Good
  2. Is Blue Gene a System for Computational Biology & Chemistry?, Sanibel Symposium, St. Simons Island, GA, 2005 – Good
  3. IBM Blue Gene - A Massively Parallel High Performance Computing Environment for Computational Science and Engineering, APAC05, Gold Coast, Australia, 2005 – Good

Second presenter:

- Prof.
- Srinivas Aluru
- Iowa State University
- Iowa State University
  Dept. of Electrical and Computer Engineering
  3227 Coover Hall
  Ames, IA 50011
- aluru@iastate.edu
- 515-294-3539
- 515-294-8432
- http://vulcan.ece.iastate.edu/~aluru

- Professor for 12 years. Recipient of teaching awards as faculty and earlier as graduate student; IEEE Distinguished Visiting Speaker from 2004 to 2006. Consistently receives high ratings from students.

- **Previous Tutorials:**
  1. Indexing methods for biological sequences, International Conference on Management of Data, December 2005.
  2. Opportunities and challenges in computational biology, Supercomputing Conference, 2002.
  3. Opportunities and challenges in computational biology, International Conference on High Performance Computing, 2001.
  4. Opportunities and challenges in computational biology, International Conference on High Performance Computing, 2000.

Other contributors to the tutorial presentation:
Charles DeLisi & team, Gyan Bhanot, Barbara Butler

50-word abstract:
The complexity of biological systems demand both advanced computer architectures and innovative approaches to exploit them. We give an overview of the IBM Blue Gene system with a few biological examples of break through results. Through a detailed description of the Maize Genome Assembly, we show how to exploit this system.

Tutorial level: Introductory to Intermediate

Prior knowledge required:
Participants that desire to understand how to use massively parallel processing in their area of research is assumed. An introductory knowledge of how to implement applications on a cluster is of value. The computational biologists who use parallel computing in their on own work will gain more from this tutorial but it is not essential to be able to do parallel programming. Some familiarity with parallel programming models and MPI (the Message Passing Interface) would be useful.

Suitability of this tutorial for ISMB:
The biological community has found computation to be an aid in gaining insight and understanding of the complexity of biological systems. Massively parallel systems like the Blue Gene machine, now available, bring a new tool to the computational biologists. We are beginning to see new science through this tool as we will point out. However, the computational biologist often needs to rethink the problem to exploit massive parallelism that exists in nature. This tutorial will arm the academic and industrial researchers, graduate students, and computational scientists with the knowledge to exploit this tool in their research and by doing so, have the opportunity to make break throughs in their problems.

Profile of Presenter 1 – Dr. Kirk E. Jordan
Kirk E. Jordan has a Ph.D. in Applied Mathematics from the University of Delaware. Throughout his career of more than 25 years in high performance and parallel computing, he has applied computational scientific techniques in many application areas, most recently in the areas of systems biology, biomedical imaging and simulation of biological systems. At IBM, he was the team lead for IBM's Healthcare and Life Sciences Strategic Relationships and Institutes of Innovation Programs over seeing and supporting many innovative research projects in computational biology. Currently, Jordan is Emerging Solutions Executive in IBM's

Deep Computing Group where he is responsible for overseeing development of applications for IBM's advanced computing architectures, investigating and developing concepts for new areas of growth for IBM especially in the life sciences involving high performance computing, and providing leadership in high-end computing and simulation in such areas as systems biology and high-end visualization. In this position, he has conducted many tutorial and workshop sessions on the use of IBM's high performance technologies in these and other application areas. Over the past year, he has led 7 tutorials and workshops, 3 in the area of computational biology. In addition, Dr. Jordan is a Research Affiliate in MIT's Department of Aeronautic and Astronautics and holds several leadership positions in the Society for Industrial and Applied Mathematics (SIAM) including Vice President for Industry.

Profile of Presenter 2 – Prof. Srinivas Aluru

Srinivas Aluru is a professor and associate chair for research in the department of Electrical and Computer Engineering at Iowa State University. He also chairs Iowa State's Bioinformatics and Computational Biology Ph.D. program that currently enrols 60 Ph.D. students. His research interests in computational biology include computational genomics and high performance computational biology on large scale parallel computers. He recently edited a comprehensive Handbook of Computational Molecular Biology, which provides the first comprehensive treatment of the field of computational molecular biology. Prof. Aluru has 12 years of teaching experience and is a well regarded speaker. He co-chairs the IEEE International workshops in High Performance Computational Biology and routinely serves on NSF and NIH panels. He is an active researcher with 18 journal and conference publications in reputed forms in 2004 and 2005 alone. He played significant roles in many conferences including program chair, program vice chair, workshop chair, tutorials chair, and served on numerous program committees in parallel processing and computational biology. He is an IEEE Distinguished Visiting Speaker from 2004 to 2006. His computational biology work received best paper awards at IEEE Computational Systems Bioinformatics conference in 2005 and IEEE Parallel and Distributed Processing Symposium in 2006.

Tutorial Outline:

Computation is playing an ever increasing and vital role in biology creating demand for new machines. Vendors strive to meet demands with advanced computer architectures such as IBM's Blue Gene machine. In this tutorial, we will give an overview of the Blue Gene architecture. We will briefly describe both the hardware and software architecture and the central philosophy behind the development of the Blue Gene that makes it easy to use on ultrascalable problems. We will emphasize the key features that allow thousands of processors to work together on a user's problem. We will present the programming model used on Blue Gene. We will explain ways to take advantage of the Blue Gene nodes and their associated networks. We hope to provide a foundation for attendees to begin to think about problems and how to design and implement them so they will scale out and take full advantage of the computational power in Blue Gene.

Once we have presented a basic understand of the architecture, our goal will be to show how Blue Gene is impacting bioinformatics through several examples. We will describe briefly some solutions done on Blue Gene in such as areas as protein folding, transcription factor binding sites, and systems biology to give demonstrate to the

audience the wide applicability. We will try to illustrate the ease of use of the systems through remote demonstration if Internet facilities are available. Depending on the partition size, we will demonstrate some simple scaling up to the number of processors available on some simple problems pertinent to the audience.

Through the computational power of Blue Gene, scientists will tackle problems that to date they had not considered. This will happen in some ways we are starting to see today but there are other approaches we yet can not predict. For this reason, we will discuss alternative approaches to design of computation of the problem that might spark others imagination. We will also show by example some problems that one might not think would be suitable for the Blue Gene architecture. We will discuss in these examples which have been run on Blue Gene systems, actual performance results. More importantly, we will try to point how having unprecedented number of processors changes how one approaches the computational problem.

The tutorial will proceed to go in depth on one application area, Maize Genome Assembly. We will describe a massively parallel framework for genome assemblies on the Blue Gene, and its application to the ongoing maize genome sequencing project. We will show how to harness the power of massively parallel Blue Gene to carry out genome assemblies at a significantly rapid pace of hours instead of days and weeks. We will discuss the applicability of this framework to solve other large-scale computational genomics problems including EST clustering, SNP identification, and selected problems in comparative genomics.

Detailed Outline:
The tutorial will be split into three major sections with some demonstration sessions as we proceed.

- First, we would give a general overview of Blue Gene, what Blue Gene is about – lots of processors. Some detail on the hardware architecture, detailing the compute core comprising the Blue Gene compute nodes, the i/o nodes, front-end node and service nodes. The five different networks will be discussed and the use of these networks will be described. The hardware and software design philosophy will be briefly described, in other words, we will explain why this system in little over year is already helping to generate breakthroughs in different sciences areas including computational biology and bioinformatics. If an internet connection is available, we demonstrate by example some of the features that will be discussed in order to give the attendees a feel for the machine. 1 hour in duration

- The second major section of the tutorial will focus on several application areas of interest to the computational biology community. These may include a discussion on how Blue Gene is being used in the transcription factor binding site identification, protein-protein interactions modelling, and high dimensional modelling of intracellular pathways. The key aspect of the particular examples discussed will be the biological problem is mapped onto the massively parallel systems and the solutions exploits the various features of the architecture to drive home the point that such solutions would not easily be implemented on a standard cluster. Where applicable, we will include some discussion on performance of the system. In some cases, a performance

comparison is hard to describe because similar problems can not be run on other systems. 1 hour in duration

- In the last major section to give the attendees a good understanding of the opportunity for breakthrough science in their respective fields, we will go into greater details on one specific application, the Maize Genome Assembly project using the PaCE Software Architecture. In particular, the Parallel Clustering Phase of the PaCE Software Architecture will be described as it is implemented on Blue Gene. We will describe in detail the $32M NSF/DOE/USDA maize genome sequencing project and how Blue Gene is being used to generate assemblies for this project. We will discuss gene enrichment sequencing using Methyl Filtration and Hi-C_ot techniques, and implications of gene enrichment to genome assembly. We will demonstrate how the framework can be fine tuned to solve several large-scale problems in computational genomics. 1 hour in duration

- We would close this tutorial with an open discussion with the attendees. Part of the discussion would be what problems the attendees might have that could use the utlrascalabilty of Blue Gene. We would describe ways the community could get started and opportunities for potential collaborations. If both time and connectivity allow, we would investigate some short runs of various codes available on Blue Gene that would be of interest to the audience. .5 hour in duration