Editorial for Emerging Computational Methods for the Life Sciences Workshop Special Issue 2012

Judy Qiu
Indiana University
Bloomington, Indiana
xqiu@cs.indiana.edu

Ian Foster
University of Chicago
Chicago, IL
foster@anl.gov

Carole Goble
University of Manchester
United Kingdom
carole.goble@manchester.ac.uk

Abstract

This paper surveys the contents of the special issue on Emerging Computational Methods for the Life Sciences Workshop 2012 with six contributed papers. They cover a rich variety of topics on interface of life sciences and data intensive computation which in detail are parallelization with GPU and multicore of an important tandem mass spectrometry peptide identification tool MyriMatch; enhancement of approximate Bayesian Computation (ABC) to estimate distributions of parameter values for mechanistic model of biological systems such as kinetic rates in a parallel environment; study of high-throughput virtual screening using a parallel version of Autodock4 to screen one-million compounds; GPU version for computation of mutual information and its normalization for biomolecular sequence alignments; a workflow environment to support parameter sweeps with a biomedical application; a browser of proteomic sequences using Multi-Dimensional Scaling to map sequences to three dimensions in a way that approximately preserves distances between sequences.

TABLE OF CONTENTS

1. BACKGROUND

Computing systems are rapidly changing with multicore, GPUs, clusters, volunteer systems, clouds, and grids offering a confusing dazzling array of opportunities. New programming paradigms such as MapReduce and Many Task Computing have joined the

traditional repertoire of workflow and parallel computing for the highest performance systems. Meanwhile the Life Sciences are continuing to expand in data generated with continuing improvement in the instruments for high throughput analysis. This “fourth paradigm” (data driven science) is joined by complex systems or biocomplexity that can build phenomenological models of biological systems and processes. This special issue for Emerging Computational Methods for the Life Sciences Workshop ECMLS2012 [7], juxtaposes these trends seeking those computational methods that will enhance scientific discovery. Within this overall scope, this special issue encouraged researchers to submit and present original work related to the latest trends in parallel and distributed high performance systems applied to Life Science problems.

Important contributions have been provided by Weber et al. [1], by Yang et al. [2], by Elllingson et al. [3], by Hamacher et al. [4], by Cushing et al. [5], and by Stanberry et al. [6]. These contributions focus on:

- Parallelization with GPU and multicore of an important tool MyriMatch that gives highly accurate tandem mass spectrometry peptide identification by multivariate hypergeometric analysis.
- Enhancement of approximate Bayesian Computation (ABC) to estimate distributions of parameter values for mechanistic model of biological systems such as kinetic rates in a parallel environment.
- Study of high-throughput virtual screening using MPI (Message Passing Interface) version of Autodock4 to run a virtual screen of one-million compounds on the Jaguar Cray XK6 Supercomputer.
- Development of GPU version for computation of mutual information (MI) and a normalization procedure to neutral evolution for biomolecular sequence alignments. The MI is computed for two- and three-point correlations within any multiple sequence alignment.
- Presentation of workflow environment to support parameter sweeps with a biomedical application for which 3000 simulations were required to perform a Monte-Carlo study.
- Presentation of a browser of proteomic sequences using Multi-Dimensional Scaling to map sequences to three dimensions in a way that approximately preserves distances between sequences.

2. SPECIAL ISSUE PAPERS

Weber et al. [1] note that GPUs and multicore processors are now pervasive in computational sciences and high-performance computing. Their high arithmetic throughput and memory bandwidth combined with their ever increasing programmability make them suitable for a widening variety of applications. They give a high level overview of Specmaster, a Myrimatch port that can use every OpenCL device available in a machine to identify peptides in tandem MS data. Then they highlight device-specific optimizations for multi-core CPUs and GPUs as well as describing framework for implementing these optimizations while still using a single code-base. They also provide performance results of Specmaster running on four different architectures and compare these numbers to Myrimatch. Finally, they improve on their existing work by showing Specmaster dynamically load balancing on 3 Radeon 7970s and 32 Interlagos 6272 cores as well as comparing the quality of their search results to Myrimatch.

Yang et al. [2] study a difficulty in building a mechanistic model of biological systems coming from determination of correct parameter values. Their paper proposes a novel parameter estimation method to infer unknown parameters, such as kinetic rates, from

noisy experimental observations. Derived from the Approximate Bayesian Computation (ABC) Sequential Monte Carlo (SMC) algorithm, their method predicts the distribution of each parameter rather than a single value via several intermediate distributions. Motivated by the computational intensity of the method, they improve the ABC SMC method in two aspects. First, to increase efficiency, a windowing method is developed to reduce the parameter searching space and an adaptive sampling weight mechanism is introduced to make the intermediate distributions converge to the target distributions in a much quicker manner. Second, to speed up the estimation process, they implement their method in a parallel computing environment to speed up the sampling process.

Elllingson et al. [3] present the current state of high-throughput virtual screening. They describe a case study of using a task-parallel MPI (Message Passing Interface) version of Autodock4 to run a virtual high-throughput screen of one-million compounds on the Jaguar Cray XK6 Supercomputer at Oak Ridge National Laboratory. The paper includes a description of scripts developed to increase the efficiency of the predocking file preparation and postdocking analysis. A detailed tutorial, scripts, and source code for this MPI version of Autodock4 are available online at http://www.bio.utk.edu/baudrylab/autodockmpi.htm

Hamacher et al. [4] present a massively-parallel implementation of the computation of (co)evolutionary signals from biomolecular sequence alignments based on mutual information (MI) and a normalization procedure to neutral evolution. The MI is computed for two- and three-point correlations within any multiple sequence alignment. The high computational demand in the normalization procedure is met efficiently with an implementation on Graphics Processing Units using NVIDIA's CUDA framework. In particular, the normalization of the MI for three-point 'cliques' of amino acids or nucleotides requires large sampling numbers in the normalization that is achieved by using GPUs. GPU computation serves as an enabling technology here insofar as MI normalization is also possible using traditional computational methods or cluster computation but only GPU computation makes MI normalization for sequence analysis feasible in a statistically sufficient sample and in acceptable time given affordable commodity hardware. They illustrate a) the computational efficiency and b) the biological usefulness of two- and three-point MI by applications to the well-known protein calmodulin and the Variable Surface Glycoprotein (VSG) of Trypanosoma brucei which are subject to involved evolutionary pressure. Here, they find striking coevolutionary patterns and distinct information on the molecular evolution of these molecules that question previous work that relied on inefficient MI computations.

Cushing et al. [5] note that task farming is often used to enable parameter sweep for exploration of large sets of initial conditions for large scale complex simulations. Such applications occur very often in life sciences. Available solutions enable to perform parameter sweep by creating multiple job submissions with different parameters. This paper presents an approach to farm workflows employing service oriented paradigms using the WS-VLAM workflow manager which provides ways to create, control and monitor workflows applications and their components. They present two service oriented approaches for workflow farming: task-level whereby task harness acts as services by being invoked on which task to load, and data-level where the actual task is invoked as a service with different chunks of data to process. An experimental evaluation of the presented solution is performed with a biomedical application for which 3000 simulations were required to perform a Monte-Carlo study.

Finally, **Stanberry et al.** [6] note that modern biology is experiencing a rapid increase in data volumes that challenges analytical skills and existing cyberinfrastructure. Exponential

expansion of the Protein Sequence Universe (PSU), the protein sequence space, together with the costs and complexities of manual curation creates a major bottleneck in life sciences research. Existing resources lack scalable visualization tools that are instrumental for functional annotation. They describe a new visualization tool using multi-dimensional scaling (MDS) to create a 3D embedding of the protein space. The advantages of the proposed PSU method include the ability to scale to large numbers of sequences, integrate different similarity measures with other functional and experimental data, and facilitate protein annotation. They apply the method to visualize the prokaryotic PSU using sequence alignment scores. As an annotation example, they use an interpolation approach to map the set of annotated archaeal proteins into the prokaryotic PSU. Transdisciplinary approaches akin to the one described in this paper are urgently needed to quickly and efficiently translate the influx of new data into tangible innovations and groundbreaking discoveries

3. ACKNOWLEDGEMENTS

We would like to thank the authors for contributing papers on their research on latest trends in data intensive technologies and applications for this special issue, and thank all the reviewers for providing constructive reviews and in helping to shape this special issue. Finally we would like to thank the editors of Concurrency and Computation: Practice and Experience for providing us an opportunity to bring this special issue to the research community.

4. REFERENCES (typesetters: please add correct Concurrency and Computation: Practice and Experience citations for refs 1-6)

1. CPE-12-0265.R1 *Specmaster: an OpenCL-based Peptide Search Engine for Tandem Mass Spectrometry*, Weber, Rick; Jenkins, David; Peterson, Gregory

2. CPE-12-0233.R1 *An iterative parameter estimation method for biological systems and its parallel implementation*, Yang, Xian ; Guo, Yike; Guo, Li

3. CPE-12-0264.R1 *Accelerating Virtual High-Throughput Ligand Docking: current technology and case study on a petascale supercomputer*, Elllingson, Sally; Dakshanamurthy, Sivanesan; Brown, Milton; Smith, Jeremy; Baudry, Jerome

4. CPE-12-0231.R1 *Using Graphics Processing Units to Investigate Molecular Coevolution*, Hamacher, Kay; Waechter, Michael; Jaeger, Kathrin; Thuerck, Daniel; Weissgraeber, Stephanie; Widmer, Sven; Goesele, Michael

5. CPE-12-0234.R1 *Applying Workflow as a Service Paradigm to Application Farming*, Cushing, Reginald; Koulouzis, Spiros; Belloum, Adam; Bubak, Maria

6. CPE-12-0177.R1 *Visualizing the Protein Sequence Universe*, Stanberry, Larissa; Higdon, Roger; Haynes, Winston; Kolker, Natali; Broomall, William; Ekanayake, Saliya; Hughes, Adam; Ruan, Yang; Qiu, Judy; Kolker, Eugene; Fox, Geoffrey

7. Emerging Computational Methods for the Life Sciences Workshop ECMLS2012 of ACM HPDC 2012 conference at Delft The Netherlands http://salsahpc.indiana.edu/ECMLS2012/