Bringing Innovations in Systems and Analytics to the Bedside: Design of the CompGen Machine

S. S. Banerjee[1], A. P. Athreya[1], Y. Varatharajah[1], M. Aly[1], C. Tan[1], Z. Stephens[1], Z. Kalbarczyk[1], S. Lumetta[1], L. Wang[2], R. Weinshilboum[2], and R. K. Iyer[1]

[1]University of Illinois at Urbana-Champaign, USA.
[2]Mayo Clinic, Rochester, USA.

Introduction

The use of *omics data is transforming the healthcare and life-sciences domains to become more precise, personalized and data-driven. The key challenge in realizing the potential is to effectively merge several different modalities of data (e.g., genomic, metabolomic, epigenetic, medical sensor and patient record data) to produce actionable intelligence that can be used in clinical therapeutic contexts. Further this data fusion and subsequent analytics must be done in a time- and cost-effective manner. This presents interesting analytics, algorithmic and computer-systems challenges dealing with computation and data storage. In this paper, we present an outline of the CompGen machine at the UIUC, which was built in collaboration with the Mayo Clinic and support from IBM. It addresses several problems at the intersection of healthcare data, novel analytical tools and methods, and novel computer system architecture and design.

At it's core, the CompGen machine uses probablistic graphical models called factor graphs, along with supervised, unsupervised learning methods, as well as sequence data processing algorithms to fuse information from several sources to perform inference and prediction. The system uses a hardware-software co-design approach to significantly improve computational performance and energy consumption. Further, using factor graphs to model the performance of various components of the system, we are able to distribute computations effectively between our custom designed accelerators as well as, popularly available accelerators like GPUs and MICs. We demonstrate the efficacy of the approach in solving several important biological and medical problems: 1) Incidence of Diabetes in Populations, 2) Psychiatric Drug Response, and 3) Seizure Prediction and Localization. Figure 1 demonstrates the overarching design of the proposed proposed framework.

Bringing Innovations in Analytics to the Bedside

Mathematical formulations that combine multi-modal data in the form of multi-omic science with longitudinal clinical measures from electronic health records to generate actionable intelligence continues to be an unsolved problem. Actionable intelligence is knowledge inferred from the data that aids in personalizing therapeutics or identifying novel biomarkers as candidates for laboratory experiments. Addressing the limitation of combining multi-modal data is the Analytics and Learning Framework for Omics and Clinical Data (ALMOND). ALMOND currently supports analyses on the following disease and data types:

1. Mixture model-based identification of novel biomarkers for drug mechanisms in triple-negative breast cancer [1].
2. The combination of probabilistic graphs, unsupervised and supervised learning for modelling and predicting drug outcomes in major depressive disorder.
3. Factor graphs based inference to predict surgical re-admissions in diabetic populations using electronic health records.
4. Approximate inference on factor graphs that model EEG data to predict epileptic seizures and localize seizure affected portions of the brain for excision.

Computer Systems for Healthcare Data Analysis

IGen [2]: At the application level, we have analyzed several computational genomics workloads like those for genome assembly, gene prediction, functional similarity analysis between protein sequences, multiple sequence alignment, phylogeny, and germline- and somatic-variant calling to find a small set of computationally intensive "kernels" that are often reused across analyses. These kernels have been performance tuned to modern to CPU, GPU and MIC architectures.

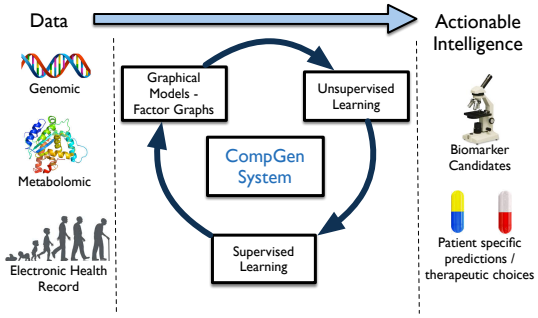

Figure 1: Integrative analysis of multi-modal healthcare data on the CompGen System.

TCGA [3, 4]: At the hardware level, we have designed, architected, prototyped and evaluated the performance of a computational genomics co-processor called TCGA (The Computational Genomics Accelerator) that targets the execution of these kernels. TCGA represents the architecture and programming model of a co-processor that targets the acceleration of computationally intensive kernels in NGS data-analytics applications. TCGA is prototyped on a Xilinx FPGA platform. TCGA uses domain-specific knowledge about the algorithms (kernels) and their input data characteristics to develop various techniques to overcome computational issues in traditional processors.

Symphony: The use of custom co-processors like TCGA, GPUs, and MICs, points to a future where biologists (non-expert users) have to program their workloads to a system of heterogeneous processors. Symphony automates this process by 1) choosing kernel implementations across all accelerators, 2) the placement of kernels in disaggregated clusters of such processors, and 3) movement of data between memories and processors. Symphony builds a cost-performance trade-off model for data-locality, processor affinity, and shared-resource contention between co-located tasks by representing information about system resources in the form a probabilistic graphical model. Using minimal training on representative workloads, Symphony integrates prior knowledge about workloads, performance counter measurements, processor architecture descriptions and interconnect topology to efficiently search the trade-off space to make scheduling decisions.

Using the human variant detection and genotyping workload as a driving example, we will demonstrate a 85× improvement in runtime performance (from 59 hours running on the Blue-Waters supercomputer to 40 minutes, for a human genome at 60× coverage) and a 300× improvement in terms performance-per-unit-energy consumed for the CompGen system.

References

[1] Athreya et al. "Model-based unsupervised learning informs metformin-induced cell-migration inhibition through an AMPK-independent mechanism in breast cancer". *Oncotarget, In Press*, 2017.

[2] Banerjee et al. "Efficient and Scalable Workflows for Genomic Analyses". *I*n Proc. DIDC, 2016.

[3] Banerjee et al. "ASAP: Accelerated Short Read Alignment on Programmable Hardware". *I*n Proc. FPGA, 2017.

[4] Banerjee et al. "On Accelerating PairHMM Computation in Programmable Hardware". *I*n Proc. FPL (to appear), 2017.