HUGenomics: a support to personalized medicine research

Lorenzo Di Tucci, Giulia Guidi, Sara Notargiacomo,
Luca Cerina, Alberto Scolari, Marco D. Santambrogio

Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria, Milano, Italy,
lorenzo.ditucci, sara.notargiacomo, luca.cerina, alberto.scolari, marco.santambrogio@polimi.it
giulia.guidi@mail.polimi.it

***Abstract*— In the coming years, human genome research will likely transform medical practices. Genome-wide association studies (GWAS) are an example of the research effort made to allowing scientists to identify genes involved in human disease, reaction to treatments or symptom severity. Indeed, the unique genetic profile of an individual and the knowledge of molecular basis of diseases are leading to the development of personalized medicines and therapies, but the exponential growth of available genomic data requires a computational effort that may limit the progress of personalized medicine. Within this context, we propose the development of a novel hardware and software integrated system, named HUGenomics. The framework aims at becoming an advanced support for personalized medicine research. Thanks to more efficient algorithms and data integration from different biological sources, HUGenomics aims at simplifying the interpretation of biological information and facilitating genomic research process by means of both computational and data visualization tools.**

I. Introduction

The possibility to exploit large *-omics* data, such as genomics, transcriptomics and proteomics, is fostering the research around personalized medicine. Thanks to the availability of these data, important efforts are dedicated to better understand their relationship to individual health, diseases origin and personal responsiveness to medical treatments.
Crucial for healthcare progress, personalized medicine and molecular diagnostics rely on *-omics* data to predict, for instance, the onset of a disease, and current research is very active on data integration, analysis and interpretation [1]. Despite the technological progresses, the computational resources needed for these tasks are still expensive, and the lack of general analysis tools further limits the developments of personalized medicine [2]. Hence, the development of personalized therapies faces two main challenges.
Firstly, we need methods to integrate data from multiple sources that maintain results accuracy and can scale to large, integrated and highly-dimensional datasets. To this aim, it is necessary to develop procedures to reduce the number of variable by means of feature reduction techniques.
The second challenge regards the need of processing large-scale genomic data. From a technological point of view, today's sequencing technologies are replacing genotyping methods based on microarray, which are generally limited to querying only regions of known variation [3].
As of now, several works focused on improving technological efficiency or on solving data integration problem [4]–[9]. However, to allow efficient progress of personalized medicine, it is necessary to address both issues at once.
The framework proposed in this paper, HUGenomics, aims to support research in personalized medicine by unifying data analysis and data integration in a single tool. To this aim, HUGenomics exploits reconfigurable architectures to accelerate data analysis algorithms, allowing users to explore multiple information levels.
To scale out computational capabilities to the size of new datasets and to the complexity of genomic algorithms while maintaining cost-effectiveness, HUGenomics leverages Field Programmable Gate Arrays (FPGAs) for application acceleration. Indeed, FPGA-based accelerators already proved to be more performing and energy-efficient than CPUs with several genomic applications [10]–[12]. HUGenomics also focuses on data integration, using both known and novel algorithms to ease the identification of biological meanings.
To draw the context of HUGenomics and detail its proposals, Section II overviews most relevant literature in the field while Section III provides a brief overview of the general framework adopted and explains the main benefits of the hardware technology used. This section also shows some case studies previously developed. Section IV concludes the paper and discusses future works.

II. Related Works

In the literature, several software and web services are available to extract information from biological and genetic data. These systems are used in molecular dynamics (MD) simulations, next-generation sequencing (NGS), GWAS and in approaches focused on data integration.
Considering MD simulations, they consist in computing atomic trajectories by solving motion numerical equations using empirical force fields; these equations approximate the actual atomic force in biopolymer systems. In this field, NAMD [13] and Gromacs [14] are the most used tools. In particular NAMD is widely used to simulate, for instance, permeation of an ion in a membrane channel and elastic vibrations of proteins. These processes have a key role, for instance, in the development of new engineered enzymes or in understanding how different proteins interact with each other. However, it is still difficult to simulate a whole process of a protein folding using the conventional MD method.

Genome-wide analyses instead search correlations between genomic variants, like Single Nucleotide Polymorphisms (SNPs), and disease phenotypes in large population. This approach present high computational costs that could require days with regular CPUs, but it is visible in [15] and [16] how hardware accelerators like GPUs and FPGAs are capable of reducing the time necessary to hours or minutes. Regarding hardware accelerations in the NGS field, a recent effort is Dragen, the first Bio-IT processor, released by Edico Genome [17]. Dragen radically reduces the computational cost and increases execution speed while maintaining results accuracy. Dragen is integrated on a PCIe card and is exposed via a Platform-as-a-Service interface that external applications can leverage. However, Dragen-based service does not allow for data integration from different sources, which is, in our vision, a key factor for improving biomedical research output and outcomes.

In this context, several works focused on data integration, either starting from materialized datasets or from remote ones. Completely materialized systems, such as EnsMart [4] or BioWarehouse [5], integrate data stored in a warehouse according to a local schema. Instead, systems like TAMBIS [6] or BioMart [7] are mediator-based, in the sense that they are designed to query remotely distributed sources through a virtual mediated schema. Mediator-based approaches provide up-to-date information at the cost of a higher complexity and costs, while materialized approaches are more efficient but cannot be used in online systems.

Moreover, available data integration platforms and workflow systems often cannot provide support for ranking-aware multi-topic searches, since these systems do not usually take into account available partial rankings in the integration process. To address these limitations, [18] proposes Search Computing, a framework that provides the basic tools required to solve complex multi-topic queries over multiple data sources with also ranking information. In order to achieve this goal, the software framework interacts with a collection of cooperating search services, using ranking information to join query results to compose the final output. In this context several works, such as BioSeCo [8] and University of California Santa Cruz (UCSC) Genome Browser Database [9], aim at supporting exploratory integrated bio-search and ranking-aware combination of distributed biomedical-molecular data, in order to answer multi-topic complex biomedical questions. BioSeco presents some limits in terms of efficiency which, as said before, has a crucial role in the development of advanced support in biomedical informatics [19]. Instead, UCSC Genome Browser Database provides genome sequence data integrated with a large amount of related annotations. Moreover, it proposes software optimizations and refinements to support fast interaction with a web-based tool for data graphic visualization.

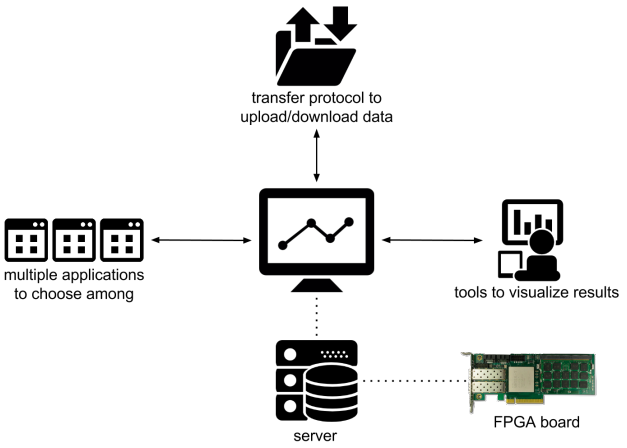

Fig. 1. An Overview of HUGenomics. Thanks to the web interface the user can exploit different services that will be run on a server equipped with an FPGA board.

III. Proposed Solution

This section introduces HUGenomics, which aims to address both the efficiency and data integration problems in a unified manner. By providing hardware-accelerated genomic algorithms, HUGenomics enables novel data integration capabilities, allowing the exploration of several information levels while limiting hardware costs.

In the following, Section III-A overviews the proposed framework, while Section III-B explains the enabling technology behind HUGenomics, reviewing existing computing architectures. Finally, Section III-C concludes with the first two applications that have been integrated in our framework: Smith-Waterman algorithm [20] and Protein Folding [21].

A. Overview of the Framework

HUGenomics is a hardware/software integrated system that aims at becoming the new support tool to the research in the field of personalized medicine. The system exploits the computing capabilities of FPGAs to process the huge amounts of genomic data available today while decreasing energy consumption. Moreover, it provides tools to visualize data in an efficient way, facilitating the identification of biological information.

An overview of the framework used for HUGenomics is visible in Figure 1. The system is accessible via a web-based interface, allowing remote access to HUGenomics and cloud-fashioned deployment and provisioning; similarly, a cloud-like communication interface allows users upload their datasets for processing, as well as downloading intermediate and final results. This model allows final users to focus only on their research tasks, without the need for specialized technical skills. Once the initial data are uploaded, HUGenomics web interface allows users to create custom workflows by choosing different algorithms for data processing: users can use the output of an algorithm as the input to another one according to their own needs, investigating arbitrarily complex pipelines for data processing. For example, users can test algorithm like Principal Components Analysis [22]

or Features Selection [23] for the data preprocessing phase. Each algorithm implemented in the framework benefits from the FPGA-based implementation that runs on the HUGenomics server. For each step of computation, we are also investigating techniques to efficiently and clearly visualize intermediate results, assisting users in the comprehension of huge data amounts.

B. Hardware Acceleration Technology

The development of biomedical sciences led to a huge increase in the amount of collected data, which required biologists to work in collaboration with computer scientists to optimize their compute-intensive applications and keep analysis times reasonable. However, the exponential growth of data sources needs new ways of management and analysis [24]. In recent years, hardware acceleration technologies like Application Specific Integrated Circuits (ASICs), Graphics Processing Units (GPUs), Field Programmable Gate Arrays (FPGAs) and co-processors (as the Intel Xeon PHI) showed to be often more energy-efficient and performing with *-omics* algorithms than Central Processing Units (CPUs), thanks to their ability to harness the intrinsic parallelism of genomic algorithms. Therefore, the HUGenomics framework relies on hardware acceleration, offloading part of the computation to the accelerator: in HUGenomics, a commodity host machine controls the execution and provides the user with the final results, while an accelerator (connected to the host) runs the heaviest part of the computation.
We considered different hardware accelerators for the HUGenomics framework, and chose FPGAs as they offer the best trade-off between performance, energy consumption and cost effectiveness. Indeed, while ASICs are best for performance and energy saving, their development cost is justified only for massive production. On the other hand, GPUs are available off-the-shelf, have high performance (especially for floating-point operations) and good flexibility, but have very high power consumption due to their high frequency. Instead, FPGAs are the best fit for many genomic applications (especially for those with no floating-point operations), as they are flexible, available off-the-shelf and have good performance and lower energy consumption than GPUs and CPUs [25], and can also handle I/O operations and provide fault-tolerance [26]. Thus, using multiple FPGAs in a single HUGenomics computing facility enables performance scalability for large datasets and keeps the total energy cost reasonable, overcoming one of the most severe obstacles to scalability [27]. Furthermore, FPGAs can be re-configured to implement different functions without performance degradation: so that HUGenomics can use the same accelerators to perform different exploration phases; this limits hardware provisioning costs and allows adding new accelerated algorithms at no additional overhead.

C. Case Studies

Two applications have been already implemented and accelerated via FPGA for the HUGenomics framework. The first application we implemented is based on the *Smith-Waterman* algorithm. It is a dynamic programming algorithm that performs pairwise local sequence alignment for DNA sequences. The Smith-Waterman algorithm is guaranteed to find the optimal local alignment with respect to the scoring system that is used for the computation. The algorithm takes as input two strings representing the database and query sequences and some parameters representing the scoring system. The output provided shows the user how the query string aligns to the database sequence. This algorithm is widely used in biology as it provides the user with information about how much a newly discovered DNA sequence is similar to another one that has been already studied. For example, it is a key algorithm in the genome analysis pipeline for personalized cancer research. Our implementation performs the main steps of computation on the FPGA board in a highly parallel and efficient way. Table I compares the execution times of software and hardware implementations, our implementation achieves performance in the order of 42 GCUPS (giga cell update per seconds) on a Kintex Ultrascale FPGA and shows the best ratio of performance over power consumption of the entire state of the art, as described in [28]. The second application implements an accelerated *protein folding* algorithm. Protein folding is the physical process by which a sequence of amino acids folds into its 3D structure, that identifies the final function of the protein. The knowledge of the tertiary structure is crucial, for instance, to create personalized drugs and treatments. For this application, the user provides an initial amino acids sequence and a starting set of dihedral angles using the web-based interface, and then just starts the computation. HUGenomics splits the computation between the CPU and the FPGA device and then returns to the user a graphical visualization of the folded protein. The protein folding application embedded in HUGenomics outperforms the existing implementation on CPU by a factor of 1.61x, as summarized in Table I, and still leaves room for more optimization, as shown in [10], [11].

TABLE I
COMPARISON BETWEEN SOFTWARE AND HARDWARE IMPLEMENTATIONS.

$Algorithm$	$Time_{CPU}[ms]$	$Time_{FPGA}[ms]$	$Speedup$
Smith-Waterman	3440.98	6.3	546.19
Protein Folding	0.4425	0.275	1.61

In addition to the Smith-Waterman and the protein folding algorithms, other applications are currently under investigation, such as a Copy Number Variation detection algorithm from whole exome-sequencing data [29] and Burrows-Wheeler's algorithm, a compression algorithm well suited to genomic data [30]. Finally, we are investigating the BLAST algorithm [31], a sequence-matching algorithm at the base of many common applications: unlike Smith-Waterman, BLAST returns results that are sub-optimal with respect to the similarity metrics, employs a similarity treshold [32] to

limit those results and is used where sub-optimal results are also acceptable. For example, it is used in the initial screening of sequence data within another application for protein structure prediction [33], also under investigation.

IV. Conclusions and Future Works

This paper presented HUGenomics, a solution to analyze and integrate data from different biological sources, like genome sequencing and proteomics. HUGenomics is an hardware/software solution based on FPGA technology that aims to help current research by providing data integration functionalities, necessary for personalized medicine.

The first two implementations produced under HUGenomics framework, namely the Smith-Waterman algorithm and the protein folding, have shown favorable results, outperforming software solutions and competing with FPGA's state-of-art in both compute speed and power consumption.

Considering the design of HUGenomics in Section III, its future developments run along three major axes. The first axis is the extension of the algorithmic library and the development of a wider set of case studies, starting from the four applications currently under investigation of Section III-C to cover a growing number of processing pipelines while providing clear benefits in scalability and power consumption. The second axis concerns the investigation of machine learning algorithms within the context of HUGenomics, for example to predict results quality. Indeed, current research [34] already applied artificial intelligence and machine learning techniques to predict drug efficacy, opening the way to new research directions between personalized medicine and machine learning, to which HUGenomics can contribute. As a third direction, it is necessary to explore the data types HUGenomics can potentially receive in input: we believe that the ability to manipulate flexible and heterogeneous data structures, like proteomics and trascriptomics data, while exploiting hardware acceleration and reconfiguration is fundamental for precision medicine. Finally, HUGenomics should keep focused on usability via its graphic interface, allowing users with any level of competence to process data and inspect results productively.

In conclusion, we believe that HUGenomics can contribute to data analysis for personalized medicine and innovative drug design, accelerating current protocols and reducing development costs associated with these research activities.

References

[1] E. R. Mardis, "The $1,000\ genome, the$ 100,000 analysis?" *Genome medicine*, vol. 2, no. 11, p. 1, 2010.

[2] B. Yngvadottir, D. G. MacArthur, H. Jin, and C. Tyler-Smith, "The promise and reality of personal genomics," *Genome biology*, vol. 10, no. 9, p. 1, 2009.

[3] G. H. Fernald, E. Capriotti, R. Daneshjou, K. J. Karczewski, and R. B. Altman, "Bioinformatics challenges for personalized medicine," *Bioinformatics*, vol. 27, no. 13, pp. 1741–1748, 2011.

[4] A. Kasprzyk, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox, and E. Birney, "Ensmart: a generic system for fast and flexible access to biological data," *Genome research*, vol. 14, no. 1, pp. 160–169, 2004.

[5] T. J. Lee, Y. Pouliot, V. Wagner, P. Gupta, D. W. Stringer-Calvert, J. D. Tenenbaum, and P. D. Karp, "Biowarehouse: a bioinformatics database warehouse toolkit," *BMC bioinformatics*, vol. 7, no. 1, p. 1, 2006.

[6] R. Stevens, P. Baker, S. Bechhofer, G. Ng, A. Jacoby, N. W. Paton, C. A. Goble, and A. Brass, "Tambis: transparent access to multiple bioinformatics information sources," *Bioinformatics*, vol. 16, no. 2, pp. 184–186, 2000.

[7] D. Smedley, S. Haider, B. Ballester, R. Holland, D. London, G. Thorisson, and A. Kasprzyk, "Biomart–biological queries made easy," *BMC genomics*, vol. 10, no. 1, p. 1, 2009.

[8] M. Masseroli, M. Picozzi, G. Ghisalberti, and S. Ceri, "Explorative search of distributed bio-data to answer complex biomedical questions," *BMC bioinformatics*, vol. 15, no. 1, p. 1, 2014.

[9] D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas *et al.*, "The ucsc genome browser database," *Nucleic acids research*, vol. 31, no. 1, pp. 51–54, 2003.

[10] G. Guidi, E. Reggiani, L. Di Tucci, G. Durelli, M. Blott, and M. D. Santambrogio, "On how to improve fpga-based systems design productivity via sdaccel," in *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE, 2016, pp. 247–252.

[11] G. Guidi, L. Di Tucci, and M. D. Santambrogio, "Profax: A hardware acceleration of a protein folding algorithm," in *Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI), 2016 IEEE 2nd International Forum on*. IEEE, 2016, pp. 1–6.

[12] M. Paolieri, I. Bonesana, and M. D. Santambrogio, "Recpu: A parallel and pipelined architecture for regular expression matching," in *Vlsi-Soc: Advanced Topics on Systems on a Chip*. Springer, 2009, pp. 1–20.

[13] M. T. Nelson, W. Humphrey, A. Gursoy, A. Dalke, L. V. Kalé, R. D. Skeel, and K. Schulten, "Namd: a parallel, object-oriented molecular dynamics program," *International Journal of High Performance Computing Applications*, vol. 10, no. 4, pp. 251–268, 1996.

[14] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen, "Gromacs: fast, flexible, and free," *Journal of computational chemistry*, vol. 26, no. 16, pp. 1701–1718, 2005.

[15] J. Gonzalez-Dominguez, L. Wienbrandt, J. C. Kassens, D. Ellinghaus, M. Schimmler, and B. Schmidt, "Parallelizing epistasis detection in gwas on fpga and gpu-accelerated computing systems," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 12, no. 5, pp. 982–994, 2015.

[16] L. Wienbrandt, J. C. Kassens, M. Hübenthal, and D. Ellinghaus, "Fast genome-wide third-order snp interaction tests with information gain on a low-cost heterogeneous parallel fpga-gpu computing architecture," *Procedia Computer Science*, vol. 108, pp. 596–605, 2017.

[17] P. van Rooyen, "Bridging tech and biotech," *Nature biotechnology*, vol. 33, no. 6, pp. 581–583, 2015.

[18] S. Ceri, A. Abid, M. A. Helou, D. Barbieri, A. Bozzon, D. Braga, M. Brambilla, A. Campi, F. Corcoglioniti, E. Della Valle *et al.*, "Search computing: Managing complex search queries," *IEEE Internet Computing*, vol. 14, no. 6, pp. 14–22, 2010.

[19] C. A. Kulikowski, E. H. Shortliffe, L. M. Currie, P. L. Elkin, L. E. Hunter, T. R. Johnson, I. J. Kalet, L. A. Lenert, M. A. Musen, J. G. Ozbolt *et al.*, "Amia board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline," *Journal of the American Medical Informatics Association*, vol. 19, no. 6, pp. 931–938, 2012.

[20] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of molecular biology*, vol. 147, no. 1, pp. 195–197, 1981.

[21] I. Lotan, F. Schwarzer, and J.-C. Latombe, "Efficient energy computation for monte carlo simulation of proteins," in *International Workshop on Algorithms in Bioinformatics*. Springer, 2003, pp. 354–373.

[22] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[23] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[24] N. R. C. (US) and I. of Medicine (US) Committee on the Organizational Structure of the National Institutes of Health., "Enhancing the vitality of the national institutes of health: Organizational change to meet new challenges." 2003.

[25] S. Kestur, J. D. Davis, and O. Williams, "Blas comparison on fpga, cpu and gpu," in *2010 IEEE computer society annual symposium on VLSI*. IEEE, 2010, pp. 288–293.

[26] D. Singh and C. K. Reddy, "A survey on platforms for big data analytics," *Journal of Big Data*, vol. 2, no. 1, p. 8, 2014. [Online]. Available: http://dx.doi.org/10.1186/s40537-014-0008-6

[27] P. Kogge, K. Bergman, S. Borkar, D. Campbell, W. Carson, W. Dally, M. Denneau, P. Franzon, W. Harrod, K. Hill *et al.*, "Exascale computing study: Technology challenges in achieving exascale systems," 2008.

[28] L. Di Tucci, K. O'Brien, M. Blott, and M. D. Santambrogio, "Architectural optimizations for high performance and energy efficient smith-waterman implementation on fpgas using opencl," in *2017 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2017, pp. 716–721.

[29] D. Backenroth, J. Homsy, L. R. Murillo, J. Glessner, E. Lin, M. Brueckner, R. Lifton, E. Goldmuntz, W. K. Chung, and Y. Shen, "Canoes: detecting rare copy number variants from whole exome sequencing data," *Nucleic acids research*, vol. 42, no. 12, pp. e97–e97, 2014.

[30] H. Li and R. Durbin, "Fast and accurate long-read alignment with burrows–wheeler transform," *Bioinformatics*, vol. 26, no. 5, pp. 589–595, 2010.

[31] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.

[32] G. Wieds, "Bioinformatics explained: Blast versus smith-waterman," *CLCBio. http://www. clcbio. com/index. php*, 2007.

[33] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker, "Protein structure prediction using rosetta," *Methods in enzymology*, vol. 383, pp. 66–93, 2004.

[34] M. K. Leung, A. Delong, B. Alipanahi, and B. J. Frey, "Machine learning in genomic medicine: A review of computational problems and data sets," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 176–197, 2016.