# High Performance Reconfigurable Computing for Bioinformatics Applications

**Alba M. A. Melo[1], Maria E. T. Walter[1], Carla C. C. Koike[1], Ricardo P. Jacobi[1]**

[1]Department of Computer Science – University of Brasilia (UnB)
Campus UNB – ICC Norte – sub-solo - 70.910-900 – Brasilia – DF – Brazil

{albamm, mia, ckoike, rjacobi}@cic.unb.br

***Abstract.*** *Compute and data intensive applications have been usually executed in high performance computing platforms. Nevertheless, with the exponential growth of the amount of data produced in some research domains, even these parallel systems are unable to produce results in reasonable time. Therefore, in order to further accelerate these applications, parts of them are migrated to hardware and implemented, for instance, in reconfigurable hardware such as FPGAs. Many bioinformatics applications are good candidates to explore the benefits of high performance reconfigurable computing platforms, since they are often very compute intensive, dealing with enormous amounts of data. This paper presents the research already conducted at UnB in the area of High Performance Reconfigurable Computing for Bioinformatics, discusses open problems and presents perspectives for future joint international research.*

## 1. Introduction

Nowadays, biological data is being produced in laboratories all over the world in a rate higher than the speed needed to process them. Public repositories for genomic data, such as the one maintained at NCBI (*National Center for Biotechnology Information*), have attained exponential growth rates. In this scenario, successful biology and medicine research labs are the ones which are able to produce accurate results very rapidly.

Parallel processing can be used to produce results faster, bridging the gap between the generation of genomic data and its analysis. Also, with parallel approaches, exact compute-intensive methods can be executed in reasonable time. However, using parallel processing in Bioinformatics is not straightforward since the problems are often solved by complex methods, with a great amount of data dependency. To further accelerate the production of results, specific hardware can be used in conjunction with parallel processing, generating high performance reconfigurable computing platforms.

The remainder of this paper is organized as follows. In section 2, we briefly describe the research developed by our group in this domain. Section 3 presents ongoing research. In section 4, we discuss the expected scientific impact. Finally, in section 5, we discuss perspectives of new collaborations and conclude the paper.

## 2. Research Accomplishments

From 2002 to 2009, our research group has been working on parallel algorithms, reconfigurable architectures and load balancing strategies for bioinformatics problems.

The first treated problem was the pairwise sequence alignment (Gusfield 1997), since it is a basic building block for solving more complex problems. A pairwise sequence alignment is defined for two sequences, and is obtained by putting one

sequence above the other, such that their similarities can be easily identified. Studying the SW dynamic programming exact algorithm proposed by (Smith and Waterman, 1981) to solve this problem, it was clear that, besides its high computing power needs, memory requirements were also a huge issue. For instance, to compare 23MBP (Mega Base Pairs) sequences, we would need at least 500 TB of memory.

Concerning the dynamic programming solution, we first proposed, in (Melo et al, 2003), a parallel heuristic variant of SW based on Distributed Shared Memory that executes in linear space. A speedup of 4.58 was achieved in an 8-processor cluster. In (Batista et al., 2004), we modified this heuristic by adding a blocking factor obtaining a speedup of 7.28, in the same platform, for the same sequences. Also, a new exact variant is proposed in (Boukerche, Melo, Ayala-Rincon and Walter, 2007) that produced almost linear speedups. Our first work using FPGA to accelerate SW was presented in (Jacobi et al., 2005). Simulation results for Altera devices produced speed-ups of two ordens of magnitude. In (Boukerche, Melo, Sandes and Ayala-Rincon 2007), we proposed a parallel exact variant of the SW algorithm that reduces memory requirements, which was able to compare 1.6MBP sequences.

Finally, we proposed z-align (Batista, Boukerche and Melo, 2008), a parallel exact variant of the SW algorithm that, based on a new concept called divergence, was able to compare 23MBPx24MBP sequences. As far as we know, this is the first work where sequences longer than 3MBP are compared with an exact affine-gap SW variant. For the 3MBP case, we reduced the execution time from 3 days and 8 hours (1 processor) to less than 3 hours (64 processors).

We also proposed in (Boukerche, Correa, Melo, Jacobi, Rocha 2007a) a hardware accelerator to execute an SW variant. In this case, our simulated FPGA prototype achieved a speedup of 246.9 over the software implementation, in a 100MBP x 100BP sequence comparison.

In order to execute BLAST in grid environments, we proposed PackageBLAST (Sousa and Melo, 2006), which distributes BLAST queries among the grid nodes using multiple task allocation strategies. In this case, a speedup of 11.28 over the best machine was achieved in a 16-machine heterogeneous platform.

In (Boukerche, Correa, Melo, Jacobi, Rocha 2007b) we propose the execution of the DIALIGN (Morgenstern et al. 1998) sequence comparison algorithm in an FPGA-based architecture. A speedup of 383.41 in a 160KBP x 190KBP comparison was obtained, reducing the execution time from more than 3 hours to 28.83s.

Besides, we developed a Peer-to-Peer (P2P) architecture for bioinformatics applications that executed in 13 heterogeneous machines distributed by three geographically distinct research institutions of Brasilia. As a case study we used BLAST (Ribeiro, Walter, Togawa, Costa, Pappas 2008). In the experiments, we obtained a gain performance of 80%, by comparing the time of running BLAST with input files containing at most 800 biological sequences against a big database (NCBI nr), both in a stand alone machine (the best machine) and in our P2P architecture.

## 3. Ongoing research

We are currently interested in the Multiple Sequence Alignment Problem (MSA), already shown to be NP-Complete. A MSA of $k > 2$ sequences $S = \{S_1, S_2, ..., S_k\}$ is obtained in such a way that chosen spaces (*gaps*) are inserted into each of the $k$ sequences so that the resulting sequences have the same length $l$. Then, the sequences are arranged in $k$ rows of $l$ columns each, so that each character or space of each sequence is in a unique column (Gusfield 1998). We are currently working on a Parallel Island Injection Genetic Algorithm to solve this problem in a reasonable time, using no

information about the classes of the sequences being compared. Now, we are designing a parallel version of the linear space variant of DIALIGN, that will be used for MSA.

We are also interested in extending our P2P architecture to other research institutions of the MidWest Region of Brazil, and to implement other applications in this framework such as efficient computational methods based on probabilistic theories to identify non-coding RNAs.

Finally, we are implementing a Sequence-Profile alignment solution that uses Hidden Markov Models (HMMs) in FPGA. In this case, a biological sequence is compared to a profile that characterizes a family of sequences.

## 4. Expected Scientific Impact

Pairwise and Multiple Sequence Alignment are basic operations in Bioinformatics and many sophisticated methods use them as a fundamental building block. Therefore, improving the accuracy of these alignments will certainly improve the accuracy of the methods that rely on them.

Even though the SW algorithm for pairwise sequence alignment has quadratic time and space complexity, the heuristic method BLAST is much more used in practice, since it can produce good average results very quick. However, as long as the sizes of highly similar sequences being compared increase, the results produced by BLAST are not so good. In (Boukerche, Batista and Melo 2009), we presented the case of the Anthrax comparison, where our z-align strategy was able to create an alignment of 5220960 bases between two different *Bacillus anthracis* strains (Ames and Sterne), resulting in a similarity of 98.1%. The same comparison using the BlastN program generated an alignment of size 36159, with a similarity of 0.69%, which does not correspond to the reality. Thus, we claim that the use of exact methods for pairwise sequence comparison can lead to a significant higher accuracy. The benefits of z-align can be better exploited if some parts of it are implemented in hardware, thus reducing drastically its execution time.

Although Multiple Sequence Alignment problem is known to be NP-Complete, the use of high performance reconfigurable computing to solve this problem can accommodate more elaborated methods that potentially could lead to better accuracy.

## 5. Perspectives of New Collaborations

The research conducted by our group has a multi-disciplinary characteristic, involving 4 research domains: parallel algorithms, high performance computing infrastructure, reconfigurable architectures and bioinformatics.

By now, we have software-only parallel solutions and hardware-only solutions for some bioinformatics problems. Nevertheless, the ideal situation would be a complete hardware/software platform where many bioinformatics problems could be solved in restricted time. In order to attain this complete solution, international collaborations are highly desirable. Research cooperation with groups working on load balancing, reconfigurable architecture design, hardware/software codesign, complex systems modeling, and large data sets could be very interesting to achieve an efficient hardware/software solution.

Another research cooperation area that can be envisaged is reconfigurable supercomputing, where FPGAs are used to accelerate algorithms in a variety of applications domains, such as cryptography, video compression and virtual reality, among others. We are also currently working on H.264/AVC video encoding with FPGAs, which has huge processing requirements.

Probabilistic inference, as employed in Bayesian modeling, is another domain where there is exact and approximated calculations due to big time and space complexity. Nowadays, Bayesian methods, including Bayesian Networks, Markov Chains, Hidden Markov Model, among others, are widely used in several applications, including bioinformatics, and we are highly interested in developing hardware/software platforms to perform these calculations in restricted time.

## References


Batista, R.B., Silva, D. N., Melo, A. C. M. A., Weigang, L. (2004) "Using a DSM application to locally align DNA sequences", In: *CCGrid*, Chicago, p.372-378.

Batista, R. B., Boukerche, A., Melo, A. C. M. A. (2008), "A Parallel Strategy for Biological Sequence Alignment in Restricted Memory Space". *Journal of Parallel and Distributed Computing*, v.68, n. 5, p. 548-561.

Batista, R. B., Boukerche, A., Melo, A. C. M. A. (2009) "Exact Pairwise Alignment of Huge Biological Sequences with the Z-Align Parallel Strategy", In*: Workshop NIDISC at IPDPS*, Rome, Italy, May.

Boukerche, A., Melo, A. C. M. A., Ayala-Rincon, M., Walter, M. E. M. T, (2007) "Parallel Strategies for Local Biological Sequence Alignment in a Cluster of Workstations". *Journal of Parallel and Distributed Computing*, v. 67, n. 2, p. 170-185, 2007.

Boukerche, A., Melo, A. C. M. A., Sandes, E., Ayala-Rincon, M. (2007), "An exact parallel algorithm to compare very long biological sequences in clusters of workstations". *Cluster Computing*, v. 10, n. 1, p. 187-202.

Boukerche, A. Correa, J. M., Melo, A. C. M. A, Jacobi, R. P., Rocha, A. F. (2007a) Reconfigurable Architecture for Biological Sequence Comparison in Reduced Memory Space. In: *Workshop NIDISC at IPDPS*, Long Beach, USA, March.

Boukerche, A. Correa, J. M., Melo, A. C. M. A, Jacobi, R. P., Rocha, A. F. (2007b), "An FPGA-based Accelerator for Multiple Biological Sequence Alignment with DIALIGN". In: *Int. Conf. on High Performance Computing*, Goa, India, p. 71-82.

Gusfield, D. (1997), "*Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*", Cambridge University Press.

Jacobi, R. P., Ayala-Rincon, M., Carvalho, L. G., Quintero, C. H. L. ; Hartenstein, Reiner W., "Reconfigurable systems for sequence alignment and for general dynamic programming". In: *Genetics and Molecular Research*, São Paulo, Brasil, v. 4, n. 3, p. 543-552, 2005.

Melo, R. C. F, Walter, M. E. M. T., Melo, A. C. M. A., Batista, R. B., Santana, M. N. P., Martins, T., Fonseca, T., (2003), "Comparing Two Long Biological Sequences Using a DSM System", In 9th *Euro-Par*, Klagenfurt, Austria, p. 517-524.

Ribeiro, E., Walter, M. E. M. T., Togawa, R. C., Costa, M. M., Pappas, G. (2008) "p2pBIOFOCO: Proposing a Peer-to-Peer System for Distributed BLAST Execution", In 10th *HPCC-08*, DaLian, China. p. 594-601.

Smith, T., Waterman, M. (1981), "Identification of common molecular sub-sequences". *Journal of Molecular Biology*, v.147, 1981, p.195-197.

Sousa, M. S., Melo, A. C. M. A.: PackageBLAST: An Adaptive Multi-Policy Grid Service for Biological Sequence Comparison (2006). In: *ACM SAC*, Dijon, France, v. 1. p. 156-160.