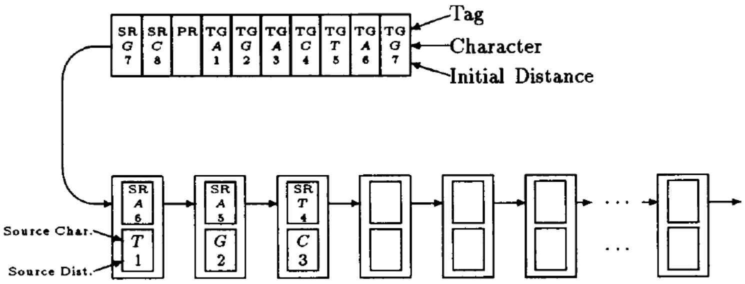
Searching Genetic Databases on Splash 2

Dzung T. Hoang

Year of publication: *1993*
Area: *Applications*



There are parallel problems. There are obscenely parallel problems. And then there is the bioinformatics edit distance problem: given a newly discovered DNA or protein sequence of length n and a massive database of size m, find all sequences in the database that relate to the newly discovered sequence.

The accepted algorithm, Smith-Waterman, is O(nm), which, given the size of the database and genetic sequences, requires a large running time. Naturally, there is a desire to discover all the relationships between all the components of the database: O(K^2) executions of an O(nm) algorithm, on databases which today are now terabytes!

Hoang's paper demonstrated the absolutely massive speedup potential present in FPGA acceleration for DNA sequence matching, showing that a single Splash 2 board (consisting of 17 Xilinx 4010 FPGAs) should be *two orders of magnitude* faster than a MasPar MP-1, a SIMD supercomputer uniquely suited to this problem. A full 16-board Splash 2 configuration would be sixteen times faster. Comparisons with a general-purpose computer are even more favorable, showing an expected four orders of magnitude improvement with a 16 board Splash 2 compared with a typical SPARC. This paper helped create a thriving industry producing custom FPGA solutions for bioinformatics, an industry that continues to demonstrate substantial benefits in compute per dollar and massive benefits in compute per Watt.

Two decades later, it is critical to reflect on why Hoang showed such amazing success. Hoang's systolic structure turns an O(nm) time problem into an O(m) problem using O(n) parallel computational resources. This one-dimensional solution allowed his implementation to scale linearly in the number of FPGAs, while a greater interconnect requirement would have prevented the scaling to multiple chips and multiple boards. The problem also scales effectively without limit, since if one string fails to consume the available resources, there is always room to compare multiple strings simultaneously.

There are also two features that make FPGAs uniquely suited to this problem beyond the problem's inherent parallelizability. The narrow bit-width design efficiently utilizes FPGA resources on the datapath, while the limited branching simplifies the control logic. Together this creates a very compact systolic cell, utilizing only 28 CLBs per cell. This compact nature ensures that the decedents of Hoang's system not only remain fast, but that they are often the most efficient of any alternatives.

Nicholas Weaver

DOI: http://dx.doi.org/10.1109/FPGA.1993.279464