Local sequence alignment for an associative model of parallel computation

Shannon I. Steinfadt
Department of Computer Science, Kent State University, Kent, Ohio 44242

Introduction

The goal of sequence alignment is to find similarity between the different strings of genetic information.

similar characters → *similar structure*
→ *similar function*

In addition to functional similarity (finding what a gene does by comparing it to genes with known function), alignment can create models for ancestral relationships, or aid in drug discovery.

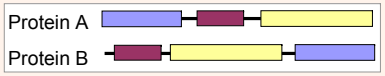

Fig. 1. Proteins with three non-intersecting local alignments.

Local Sequence Alignment

Sequence alignment is an optimization problem, maximizing scores for local, or sub-sequence, alignments using given weights. The well-known Smith-Waterman algorithm[1] always finds the highest scoring alignment, but is slow because it uses the dynamic programming (DP) method. This has led to approximation (heuristic) algorithms like BLAST, whose alignments are fast, but they may not be the "best."

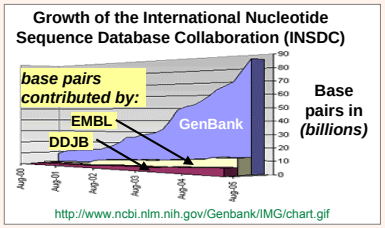

Fig. 2. Exponential growth of public sequence data means more to align with; the faster an alignment, the better.

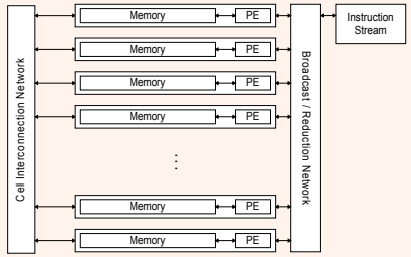

Fig. 3. A high-level view of the ASC model. The MASC model includes more than one instruction stream.

Parallel Models

The goals were to create a fast local sequence alignment algorithm that produces more information:

- Returning the top *k*-ranked, non-intersecting local alignments from a single run
- Finds all *k*-ranked local alignments in the same time it takes to find one

The ASsociative (ASC) and Multiple ASsociative Computing (MASC) models[2] are SIMDs with an associative property and some additional hardware features.

ASC and MASC features:

- Fast processing
- Constant time search and respond operations
- Constant time global reduction operations, i.e. retrieving the minimum / maximum value among processing elements (PEs)
- A base of existing algorithms[3] that can be adapted
- An existing programming language and emulator

Algorithm Adaptation

The examples use the following weights:
Gap Insert: 3 Gap Extend: 0
$d(S1_i, S2_j) = 10$ when $S1_i = S2_j$
$d(S1_i, S2_j) = 10$ when $S1_i \neq S2_j$

PE j[$] index / S2

PE i[$] index / S1		***0*** Δ	***1*** C	***2*** U	***3*** G	***4*** G
0	Δ	0	0	0	0	0
1	C	0	10	7	7	7
2	A	0	7	7	4	4
3	U	0	7	17	14	14
4	U	0	7	17	14	14
5	G	0	7	14	27	24

Fig. 4. Smith-Waterman dynamic programming table showing the dependency free anti-diagonal that the ASC algorithm executes in parallel.

The following image is a *partial* mapping of the DP table to the ASC model.

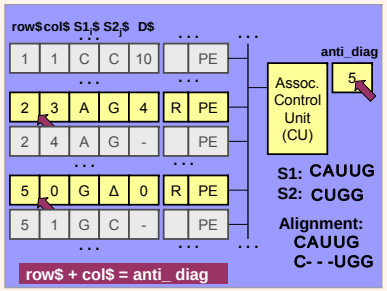

Fig. 5. Mapping the adapted Smith-Waterman algorithm onto the ASC model. This shows select PEs with marked *responders* that have "active" status and will be processed in parallel. "Inactive" PEs will not be processed in the next step.

Future Work – Conclusions

This work is the foundation for additional work that includes:

- Extending algorithm from ASC→MASC
- Pairwise alignment with multiple local (non-intersecting) alignments
- Multiple sequence alignment (aligning three or more sequences)
- Achieve better timings, practical use by migrating to physical architectures
  - The MIMD-SIMD cluster
  - The "SIMD on a board" World*Scape* PCI board

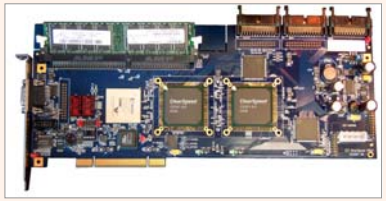

Fig. 6. World*Scape* Dual 64 PCI SIMD Board with 50 GFLOPS performance.

The goal of this work is to provide fast, accurate, more detailed alignments that aid in the bioinformatics field for those using sequence alignment.

References

[1] Gotoh, O. "An Improved Algorithm for Matching Biological Sequences." *J. of Molecular Biology* 162, 705-708, 1982.
[2] Potter, J., J. Baker, A. Bansal, S. Scott, C. Leangsuksun, and C. Asthagiri. "ASC: An Associative Computing Paradigm." *IEEE Computer*, 27(11): 19-25, November, 1994.
[3] Esenwein, M., J. Baker, "VLCD String Matching for Associative Computing and Multiple Broadcast Mesh", *Proc. of 9th IASTED International Conf. on Parallel and Distributed Computing Systems*, 69-74, 1997.

For further information

Please contact *ssteinfa@cs.kent.edu*. More information on this and related projects can be obtained at *www.cs.kent.edu/~parallel.*

KENT STATE UNIVERSITY