PETROGLYPHS TO PETAFLOPS
CUG 2005

Implementation of PSI Smith-Waterman Algorithm on Cray X1

Lukasz Bolikowski

ICM, Warsaw University

Project stages

- Original Smith-Waterman on Cray SV1
- Filtering using BMM unit
- Position-Specific Iterated S-W on Cray X1

PETROGLYPHS TO PETAFLOPS
CUG 2005

Sequence alignment

- **Input**: two amino acid sequences
  - MDRKVTPGSTCAVFGLGGVGLSAIMGFIL
  - MKLNPGSSGHGGMGATMTSAVMGDRNN
- **Output**: an alignment
  - KVTPGSTCAVFGLGGVG---LSAIMG

```
K+ PGS+    G GG+G    SA+MG
KLNPGSS----GHGGMGATMTSAVMG
```

Score calculation

- Scoring table

	A	R	N	D
A	4	-1	-2	-2
R	-1	5	0	-2
N	-2	0	6	1
D	-2	-2	1	6

- Gap penalties
  - For opening (~ 10)
  - For extending (~ 1)
- Alignment score is simple to calculate!
  - Add scores of individual pairs
  - Substract gap penalties

Algorithms

- Smith-Waterman (S-W)
  - Exact algorithm (never fails)
  - Dynamic programming, costly
- Basic Local Alignment Search Tool (BLAST)
  - Heuristic algorithm
  - Extremely fast

PSI approach

- Position-Specific Iterated method
  - Input: *a query* and *a database*
  - Choose an initial scoring table
  - Run the query on the database
  - Use the results to construct a new table
  - Iterate (until the tables converge)
- PSI BLAST is popular
- No known PSI S-W implementation

Filtering with BMM

- **Observation**: S-W is costly
- **Action**: use a fast filter that calls S-W only for sequences that pass certain tests
- **Means**: BMM unit and vectorized bit operations

Filtering with BMM

- **Observation**: sign of a score is the most important information
- **Action**: use 0-1 for dynamic programming tables in S-W
- **Means**: Bit matrix multiply

Filtering with BMM

- **Means**: Bit matrix multiply
  - Represent letters as 64-bit vectors (= a word)
  - Prepare a bit matrix Q for the query
  - For each sequence in the database: fill the dynamic programming table by bit matrix multiply with Q

Flitering with BMM

- **Observation**: high-scoring alignments have long ungapped segments rich in +s
- **Action**: identify the regions, call S-W
- **Means**: a series of shifts, ANDs and ORs

Filtering with BMM

- **Means**: a series of shifts, ANDs and ORs

  -
    ```
    00111001101110001000
     00111001101110001000
      00111001101110001000
    ```

  -
    ```
    0011111111111110111000
    011111111111110111000
    11111111111110111000
    ```

  -
    ```
    00111111111110001000
    ```

Filtering with BMM

- **Observation**: high-scoring alignments have islands of +s
- **Action**: identify the islands, call S-W
- **Means**: a series of shifts and ANDs

PETROGLYPHS TO PETAFLOPS
CUG 2005

Filtering with BMM

- **Means**: a series of shifts and ANDs

  ```
  ▪ 00111001101110001000
     00111001101110001000
       00111001101110001000
  ▪ 0000100000001000000000
  ```

Quality of results

- For several families of proteins:
  - Run BLAST, S-W, PSI S-W, PSI S-W w/ filter
  - Use SWISS-PROT as a database
  - Set a weak threshold of reporting sequences
  - Record the number of sequences found

Quality of results

Family	BLAST	S-W	PSI S-W	PSI S-W w/ filter
Serine protease inhib.	155	157	161	121
Ras	500	568	1407	200
Globin	57	147	786	48
Hemagglutinin	141	142	170	108
Cytochrome P450	602	662	716	312
Alcohol dehydrogenase	221	232	287	129

Execution time

- Database of 10,000 random sequences
- Execution times for 1 SSP
  - PSI S-W: **80** seconds
  - PSI S-W + conservative filter: **31** seconds
  - PSI S-W + aggressive filter: **7** seconds

Conclusions

- Slower than (PSI) BLAST, but finds better alignments
- Cray X1 capabilities were required
  - Vectorized Smith-Waterman
  - BMM and vector bit operations for the filter
  - Computational power for PSI S-W

PETROGLYPHS TO PETAFLOPS
CUG 2005

Authors

- Rafal Maszkowski
- Lukasz Bolikowski
- Maciej Cytowski
- Maciej Dobrzynski
- Maria Fronczak
- Witold Rudnicki

Thank you!

Lukasz Bolikowski

bolo@icm.edu.pl