Sequence alignment calculation in SIMAP 2.0

Nov 22, 2014

Software version and parameters

- Software version: swipe_swlib1.09
- General parameters: -M BLOSUM50 -G 13 -E 2 -m 88 -s2 -b <nsequences_in_db> -v <nsequences_in_db>

Input files

- Q: query sequences as soft-masked multiple fasta file (low complexity characters in lower case)
- DBI: database sequences as BLAST index from hard-masked multiple fasta file (low complexity characters translated into X)
- DB: database sequences as soft-masked multiple fasta file (low complexity characters in lower case)

Output format

qseqid sseqid score score_symm qstart qend sstart send pident ppos length nident positive mismatch gapopen gaps flags

Alignment phase 1

- Based on swipe, using SSE3
- Fastest available acceleration of Smith-Waterman algorithm, but cannot be combined with composition-based score adjustment
- All pairs having scores lower than -c threshold are discarded
- Scoring based on native substitution matrix (parameters -M, -G, -E)
- Scores only calculated using 7bit routine (sufficient to test against –c cutoff, which is lower than 128)
- Q internally hard-masked
- DBI used as is (hard-masked)

Alignment phase 2

- Based on a combination of swlib (for scores <32k) and swipe (fullsw for scores >=32k; slower than swlib; accelerated by stopping after score is larger than -B threshold)
- Scoring based on composition-based score adjustment of substitution matrix (parameters -M, -G, -E)
- All pairs having scores lower than -B threshold are discarded
- For composition-based score adjustment and score calculation:
    - Q internally hard-masked
    - DBI used as is (hard-masked)

- Differences to BLAST: BLAST only masks the database (introduces asymmetry); BLAST has special rules for very similar sequences (scores not completely continuous)

Alignment phase 3

- Based on a combination of swlib (for scores <32k) and swipe (fullsw; slower than swlib)
- Scoring based on composition-based score adjustment of default BLAST substitution matrix (BLOSUM62/-11/-1)
- All pairs are kept – this step only calculates the final score and alignment attributes using swipe's align function
- For composition-based score adjustment:
  - Q internally hard-masked
  - DBI used as is (hard-masked)
- For score calculation:
  - Q internally unmasked (all characters as upper case)
  - DB internally unmasked (all characters as upper case)
- Differences to BLAST: BLAST only masks the database (introduces asymmetry); BLAST has special rules for very similar sequences (scores not completely continuous)

Performance and symmetry evaluation

Test data and parameters:

- Queries: all sequences from Swissprot from November 2014
- Database: all sequences from Swissprot from November 2014
- all-against-all calculation with varying –c and -B=80

Results:

```
Value of -c: 75
Total runtime: 6774235.0s (78 days, 9:43:54).
Pairs with equal scores: 243528820    Pairs with different scores: 0   Singletons: 0


Value of -c: 65
Total runtime: 8885583.9s (102 days, 20:13:03).
Pairs with equal scores: 275136275    Pairs with different scores: 0   Singletons: 0


Value of -c: 70
Total runtime: 7728945.9s (89 days, 10:55:45).
Pairs with equal scores: 265583329    Pairs with different scores: 0   Singletons: 0


Value of -c: 60
Total runtime: 12178727.8s (140 days, 22:58:47).
Pairs with equal scores: 278379059    Pairs with different scores: 0   Singletons: 0


Value of -c: 55
Total runtime: 19863847.8s (229 days, 21:44:07).
Pairs with equal scores: 279473319    Pairs with different scores: 0   Singletons: 0
```

```
Value of -c: 50
Total runtime: 36158551.8s (418 days, 12:02:31).
Pairs with equal scores: 279850071    Pairs with different scores: 0   Singletons: 0
```

Comparison to BLAST

Test data:

- Queries: 3560 sequences from Swissprot from November 2014
- Database: all sequences from Swissprot from November 2014
- BLAST calculation with ssearch (phase 1) and blastp (phase 2 and phase 3), calculation and alignment parameters are equivalent to those of simap
- SIMAP calculation with varying –c and varying -B

Results:

Table 1: Total runtime in seconds for combinations of –c and –B

	-B threshold							
-c	80	75	70	65	60	55	50	45
75	47143.2	48285.0	47446.6	46110.5	48973.0	47093.3	46620.5	45775.5
70	47546.0	46297.2	46162.5	43664.8	45520.2	45474.0	46683.8	50654.7
65	55469.8	55871.4	60520.5	63314.3	66312.2	67738.6	68420.4	65494.8
60	68558.7	72695.8	74521.6	81562.0	90251.1	99821.4	103089.7	103836.9
55	118009.8	115900.5	119758.7	129094.0	148289.7	168955.5	181626.4	192865.0
50	208842.2	204260.0	211069.1	224900.8	241744.8	292710.6	343110.6	364582.0
45	376456.5	373698.2	380278.3	384959.1	408444.1	465917.6	558228.1	650618.8

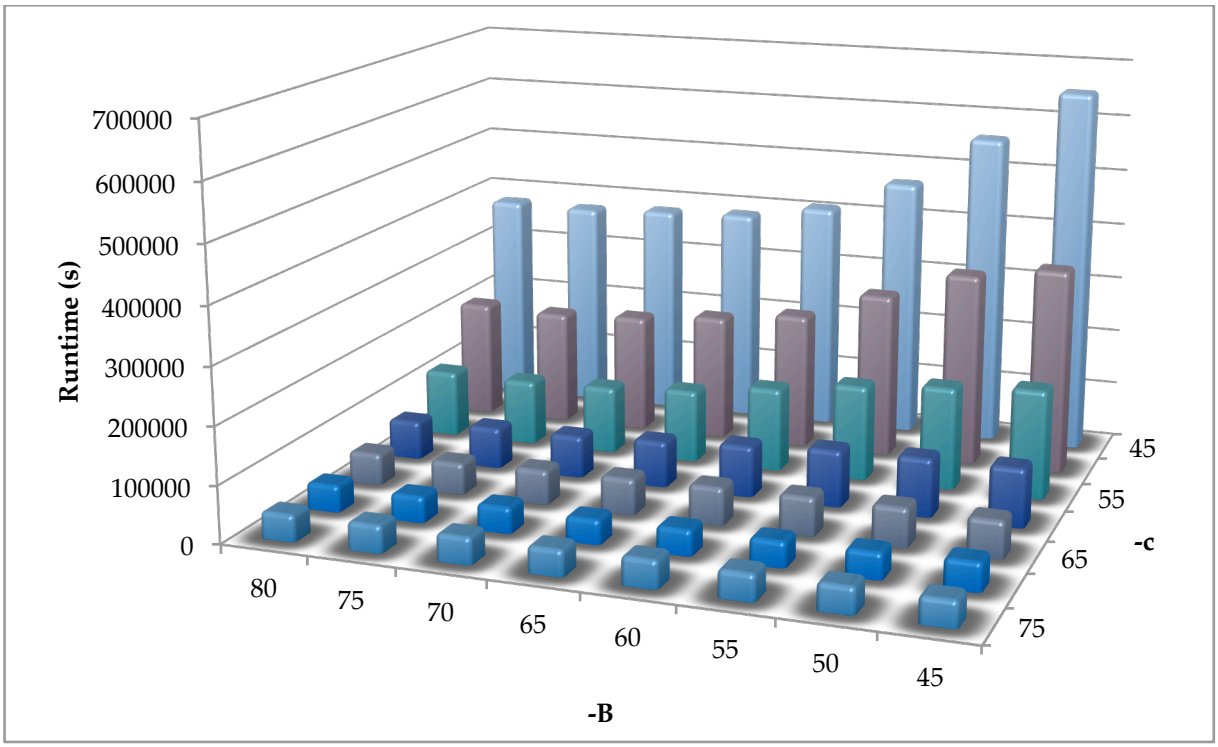

Fig 1: Total runtime for combinations of –c and –B