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:
  o Q internally hard-masked
  o 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:
  o Q internally hard-masked
  o DBI used as is (hard-masked)
• For score calculation:
  o Q internally unmasked (all characters as upper case)
  o 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, 0:13:03).
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

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

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