GASSST

Global Alignment Short Sequence Search Tool

Jana Applová

Základní údaje

Autoři:

- D. Lavenier ENS-Cachan/IRISA
- G. Rizk IRISA
- D. Fleury IRISA
- (Institute for Research in IT and Random Systems)

- Francie
- verze 1.262
- C++
- open source

Motivace

- potřeba rychlého a přesného zarovnávacího softwaru
- neomezování delecí a inzercí
- efektivita u dlouhých readů

GASSST

- seed-filter-extend
- hashovací tabulka

- PST... pre-computed score table
- TNW... Tiled NW
- FD-vec... frequency distance vectorized filter

$$ECD(F,G) = \sum_{1 \le i \le m} \frac{|u_i - v_i|}{2}$$

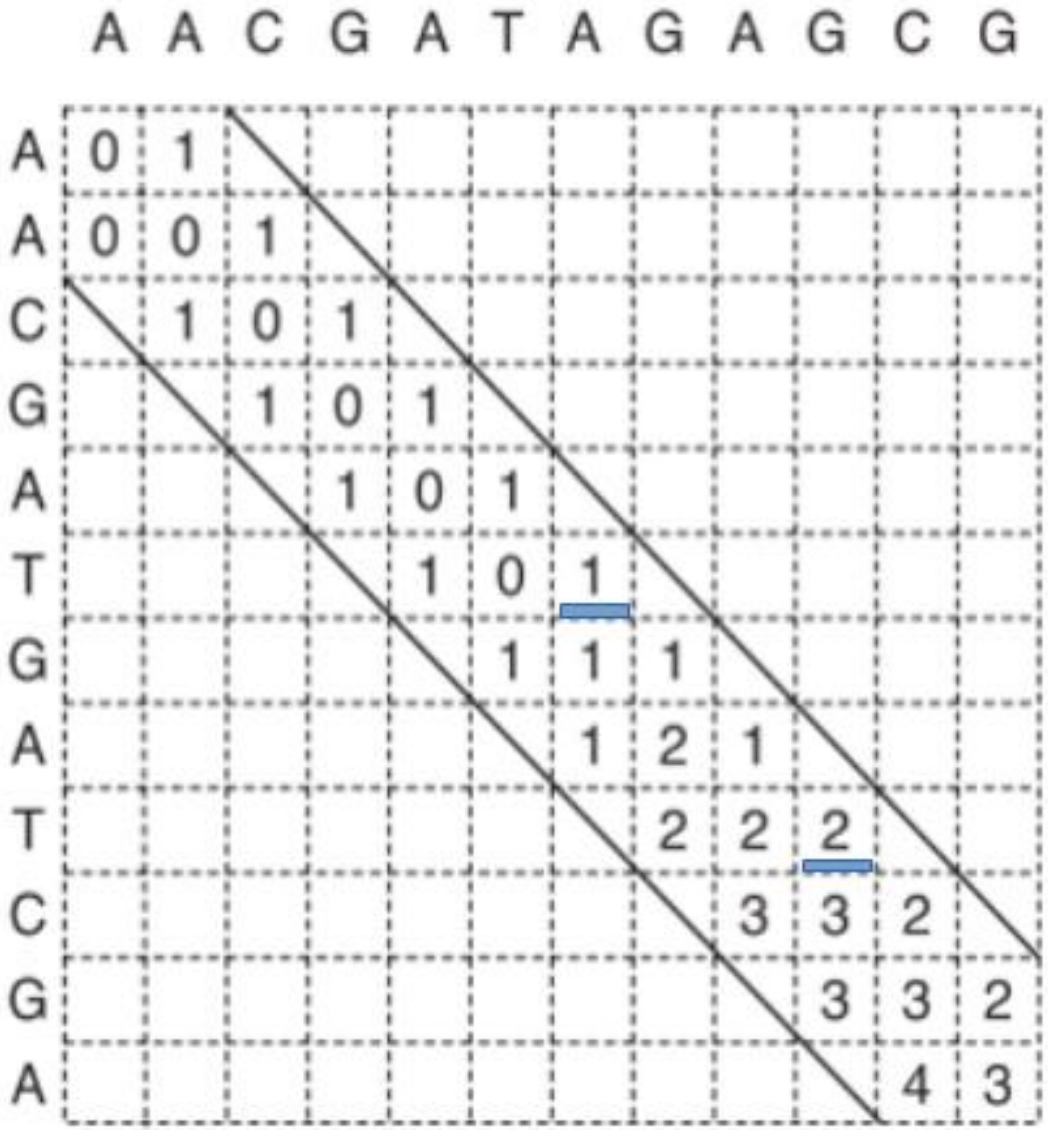
For a sequence $S = s_1 s_2 \ldots_n$ of characters in the alphabet $\Sigma = \{a_1, a_2, \ldots, a_p\}$, the frequency vector $F = \{f_1, f_2, \ldots, f_m\}$ is defined as:

$$\forall i \in [1;m]\, f_i = \sum_{1 \le k \le n} \delta_{s_k, a_i} \tag{4}$$

With

$$\delta_{s_k, a_i} = \begin{cases} 1 & \text{if } s_k = a_i \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

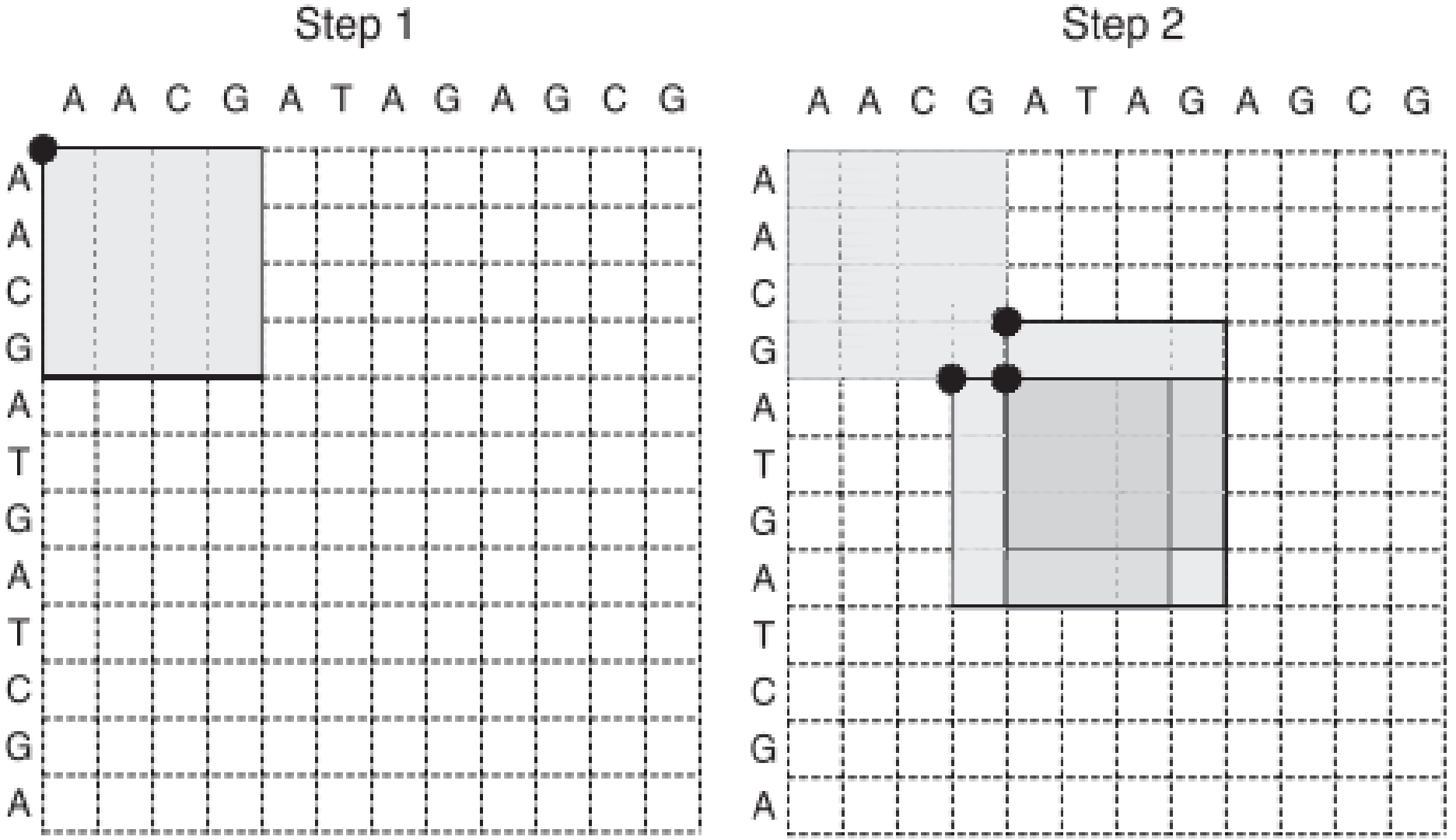
A Traditional banded semi-global alignment



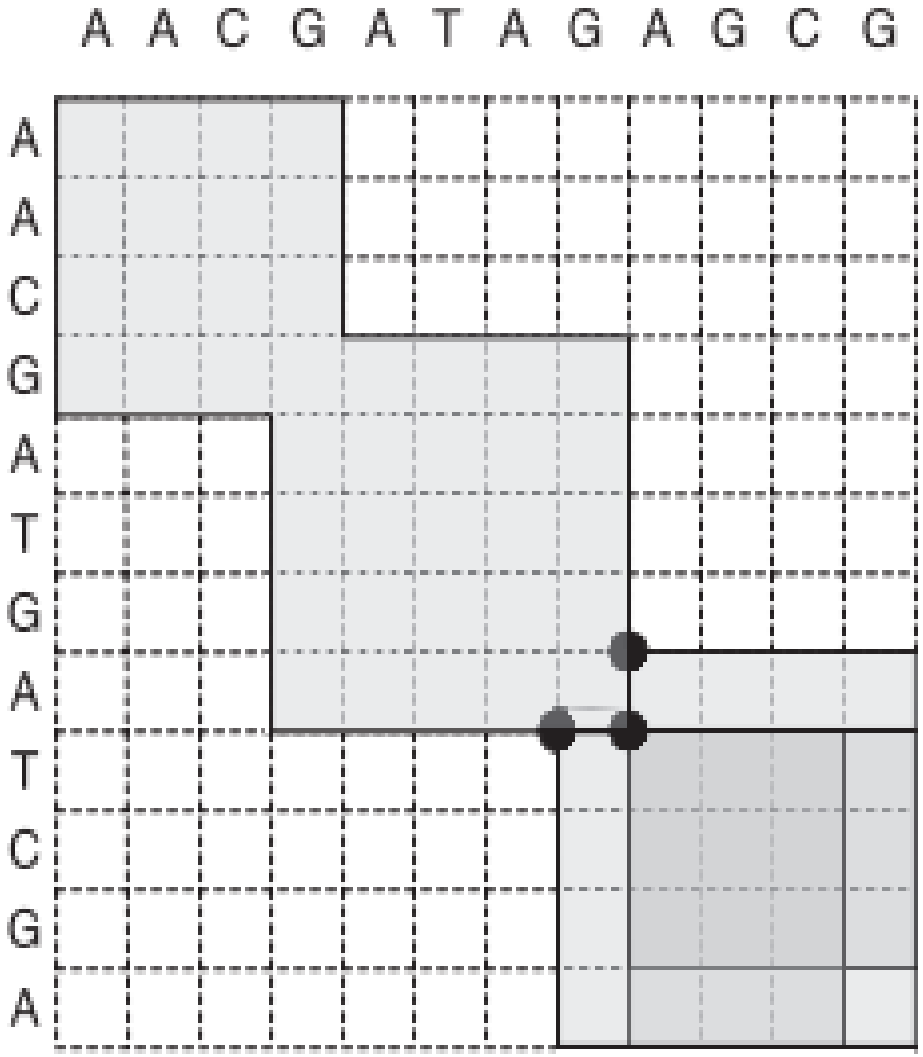
Alignment result : 2 errors

```
A A C G A T A G A G C G
| | | | | |   | |   | |
A A C G A T - G A T C G
```

B New Tiled Algorithm



Step 3



Tiled Algorithm Computations:

Step 1:
PST(AACG,AACG)=0
error = 0

Step 2:
PST(GATG,ATAG)=2
PST(ATGA,ATAG)=1
PST(ATGA,GATA)=2
error = min(1+2, 0+1, 1+2) = 1

Step 3:
PST(AGCG,ATCG)=1
PST(AGCG,TCGA)=2
PST(GAGC,TCGA)=2
error = min(1+1, 1+2, 1+2) = 2

Algorithm 1 Main computation loop

```
1.  Input:
2.  reference genome, short query sequences.
3.  parameter: n (maximum errors allowed).
4.  Pre-calculation:
5.  Compute the reference genome index
6.  for each query q do
7.      for each overlapping seed s in q do
8.          for each occurrence o of s in reference genome do
9.              if TNW_4(4){q,s,o} < n then
10.                 if FD-vec{q,s,o} < n then
11.                     if TNW_4(16){q,s,o} < n then
12.                         if TNW_7(full){q,s,o} < n then
13.                             if NW{q,s,o} < n then
14.                                 Print Alignment {q,s,o}
```

DEFINITION 1. *We call $TNW_l(n)$ the score of a region of n nucleotides which are adjacent to the right of the seed and which are computed with a PST_l. TNW stands for* Tiled NW. *$TNW_l(full)$ operates on the maximum length available, i.e. from the right of the seed to the end of the query sequence.*

DEFINITION 2. *If S1 and S2 are the two sequences directly adjacent to the right of the seed, we call $PST_l(i, j)$ the pre-computed NW score for the two l nt long words $S1(i, i+l-1)$ and $S2(j, j+l-1)$.*

If G_m is the maximum number of allowed gaps, $TNW_l(n)$ is computed with the following recursion:

If $n \leq l$ then

$$TNW_l(n) = PST_l(1,1) \tag{1}$$

Else

$$TNW_l(n) = \min(A,\ B,\ C) \tag{2}$$

With

$$A = TNW_l(n-l) + PST_l(n-l+1,\ n-l+1)$$

$$B = \min_{1 \leq s \leq G_m} (\max(s,\ TNW_l(n-l)) + PST_l(n-l+1,\ n-l+1-s))$$

$$C = \min_{1 \leq s \leq G_m} (\max(s,\ TNW_l(n-l)) + PST_l(n-l+1-s,\ n-l+1))$$

In the general case, the number of PST_l accesses $N_{acc}[TNW(n)]$ needed for the computation is in $\mathcal{O}(n)$. If G_m is the maximum number of allowed gaps, the exact formula is:

$$N_{acc}[TNW(n)] = \left(\left\lceil \frac{n}{l} \right\rceil - 1\right) \cdot (1 + 2 \cdot G_m) + 1 \tag{3}$$

Data Set	Filter Step	Step filter (%)	Total filter (%)
36 bp	$TNW_4(4)$	64.9	64.9
	$FD-vec$	72.8	90.5
	$TNW_4(16)$	96.6	99.7
	$TNW_7(full)$	45.9	99.8
	NW	63.8	99.9
50 bp	$TNW_4(4)$	41.7	41.7
	$FD-vec$	80.4	88.6
	$TNW_4(16)$	97.5	99.7
	$TNW_7(full)$	88.5	99.97
	NW	67.5	99.99
76 bp	$TNW_4(4)$	0.1	0.1
	$FD-vec$	49.8	49.9
	$TNW_4(16)$	93.7	96.8
	$TNW_7(full)$	94.8	99.8
	NW	70.2	99.95

Délka výpočtu jednotlivých filtrů

Step	Execution time for one hit (ns)	Total percent time spent in this step
General Overhead	13.9	33
$TNW_4(4)$	5.8	14
$FD-vec$	22.4	32
$TNW_4(16)$	78.5	15
$TNW_7(full)$	356.7	3
NW	4486.5	3

Read length	Software	Index (s)	Align (s)	Q20%	Q20 Error rate (%)
36 bp	GASSST	1712	3211	34.5	0.14
	BFAST	520 800	17 520	41.8	0.12
	BWA	5158	3739	35.4	0.17
	PASS	2312	5072	41.4[a]	–
50 bp	GASSST	1719	4090	73.7	0.04
	BFAST	520 800	22 799	80.4	0.10
	BWA	5158	3043	74.5	0.17
	PASS	2144	5384	79.3[a]	–
76 bp	GASSST	1701	8483	81.3	0.04
	BFAST	520 800	161 220	85.4	0.28
	BWA	5158	3101	86.4	0.53
	PASS	1951	118 541	87.7[a]	–

Program	Mode	Metrics	50 bp			100 bp			200 bp			500 bp		
			2%	5%	10%	2%	5%	10%	2%	5%	10%	2%	5%	10%
GASSST	fast	Align sec	584	781	1720	794	981	1160	2030	2314	3051	6573	8453	11 859
		Sensitivity%	45.8	42.8	36.1	54.2	53.3	44.9	58.6	56.3	53.8	61.0	59.8	58.8
		Accuracy%	99.2	98.5	93.8	90.5	89.4	86.9	91.9	91.5	89.7	93.0	92.8	91.7
GASSST	accurate	Align sec	1709	2741	4706	1290	1887	3262	3452	6173	8744	12 864	18 737	34 222
		Sensitivity%	46.7	44.6	37.5	51.0	50.3	43.5	53.4	52.8	51.7	57.5	55.7	55.5
		Accuracy%	99.8	99.3	93.5	99.7	99.4	97.5	99.9	99.8	99.3	99.9	99.9	99.7
BFAST		Align sec	2279	2044	1756	15 263	15 787	11 452	–	–	–	–	–	–
		Sensitivity%	46.5	43.0	32.1	52.7	51.0	48.5	–	–	–	–	–	–
		Accuracy%	98.8	96.1	85.2	99.0	98.7	95.8	–	–	–	–	–	–
BWA		Align sec	792	1392	1572	1862	4941	3364	4660	2145	185	–	–	–
		Sensitivity%	48.2	38.6	16.8	54.8	41.0	7.3	53.3	11.9	0.1	–	–	–
		Accuracy%	99.2	97.4	93.5	99.7	99.0	97.9	99.8	99.6	96.7	–	–	–
BWA-SW		Align sec	–	–	–	–	–	–	4699	3546	2365	13 027	9646	7835
		Sensitivity%	–	–	–	–	–	–	54.9	50.3	25.2	57.3	56.1	45.4
		Accuracy%	–	–	–	–	–	–	99.4	96.9	85.7	99.2	96.8	85.2
SSAHA2		Align sec	–	–	–	27 740	41 978	45 295	22 285	27 504	65 420	179 095	415 252	275 622
		Sensitivity%	–	–	–	45.3	43.5	38.6	53.2	51.4	48.4	59.5	58.8	55.8
		Accuracy%	–	–	–	99.8	99.1	95.3	99.8	99.1	96.0	99.8	99.2	95.3
PASS		Align sec	2012	2281	5085	14 387	26 033	30 022	103 338	139 436	180 943	–	–	–
		Sensitivity%	50.0	43.8	31.3	51.6	37.5	16.4	49.3	16.6	2.8	–	–	–
		Accuracy%	96.6	93.2	82.4	98.5	94.0	86.8	97.2	93.6	92.4	–	–	–

Povinné parametry

- -d <bank_file>
  : (string) Name of a DNA sequence bank in FASTA format

- -i <query_file>
  : (string) Name of a DNA sequence bank in FASTA format - Short sequences

- -o <output_file>
  : (string) Name of the output file to store the results

- -p <identity_percentage>
  : (integer) Minimum percentage of identity. Must be in the interval [0 100]

Doplňkové volby

- -w <size_seed>
- -m <output_format>
- -n <thread number>
- -l <complexity_filter>
- -t <size_partition>
- -r <reverse_complement>
- -g <gaps_number>
- -h <max alignment per read>
- -s <sensitivity level>
- -b <output best alignments>

Minimum percentage of identity [0..100] 100

GASSST options

Output format Standard Gassst format

Seeds size to use ?

Number of gaps allowed ?

Maximum number of alignment per query (0 is no limit) 100

Sensitivity level, in [0..5] ? 2

Output best alignments ? Yes

Enable low complexity filter Yes

Enable reverse complement search Yes

Odkazy

- http://bioinformatics.oxfordjournals.org/content/26/20/2534.full.pdf+html (zpracovávaný text)
- http://www.irisa.fr/symbiose/projects/gassst/
  (zde dostupný zdrojový kód)
- https://www.biocatalogue.org/
  - http://gassst.genouest.org
    - http://mobyle1.genouest.org/cgi-bin/Mobyle/portal.py?#forms::gassst
      (zde možno vyzkoušet GASSST online, potřeba e-mailu)
- http://www.genomequest.com
  (další možnost, kde si vyzkoušet GASSST online, potřeba založení účtu)

INRIA and GenomeQuest entered a collaboration on licensing and development early 2010 (see http://wiki.genomequest.com/index.php/LSPMUL_GASSST for more information).

Děkuji za pozornost!