

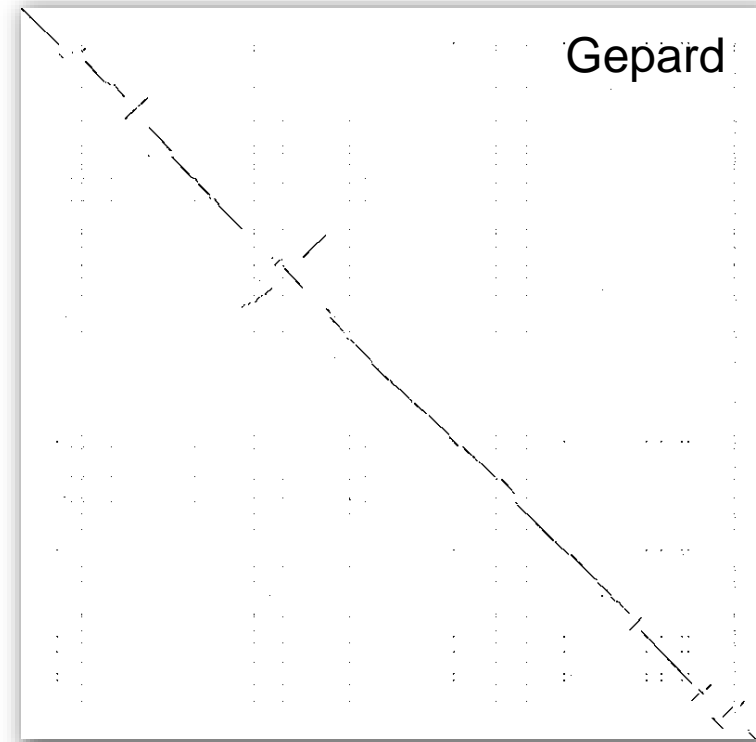
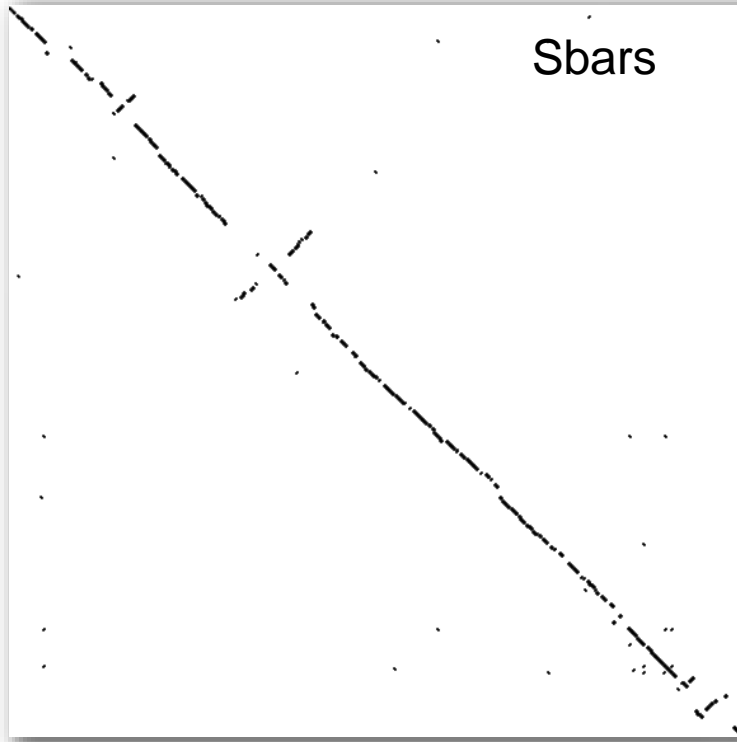
# Optimization of parallel algorithm for global alignment of long sequences

A.N. Pankratov, R.K.Tetuev, M.I.Pyatkov

IMPB RAS – the branch of KIAM RAS, Pushchino

# Escherihia coli K12

Shigella flexneri 2a



M.I. Pyatkov, A.N. Pankratov. SBARS: fast creation of dotplots for DNA sequences on different scales using GA, GC-content// **Bioinformatics**, 2014, volume 30 (12), pages 1765-1766.

Krumsiek J., Arnold R., Rattei T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale // **Bioinformatics**. 2007. Apr. Vol. 23, 8. P. 1026-1028.

Length	Gepard	SBARS
100000 н.п.	< 1 сек	< 1 сек
1000000 н.п.	< 5 сек	< 3 сек
5000000 н.п.	45 сек	< 14 сек
Y chr (2900000 н.п.)	5 мин	27 сек

# Goals of work

- Verification of long repeats detected by spectral method
- Universal web-service for recognition and verification of inaccurate repeats based on distributed computing in browser virtual machines.

# Global alignment of long sequences

ACTATTGAAAGATT-----  
 .....|.||.....  
 -----CG-TTGGGGGTATCA

## Computational complexity

$\sim O(n^2)$

## Space complexity

(Needleman&Wunsh, 1970)

$\sim O(n^2)$

(Miller&Myers, 1988,

Dryga, 2006)

$\sim O(n)$

(Tetuev, Pyatkov, Pankratov,  
 2017)

$\sim O(n^{\frac{4}{3}})$

3	1 1 1 1 1	1 1 1	1 1 1 1 1	1
0	0 0 0 0 0	0 0	0 0 0 0 0	0
2	:	:	:	:
2	:	:	:	:
2	:	:	:	:
0	0 0 1 0	-2 -3 -3	-3 -1 -1 1 0	1
2	2 3 2 3 3	:	:	:
2	3 2 3 3 1	:	:	:
2	3 3 3 3 3	:	:	:
2	3 3 1 2 3	:	:	:
2	3 2 3 2 2	:	:	:
2	3 2 3 3 2	:	:	:
2	3 3 2 3 2	:	:	:
2	3 3 1 1 3	:	:	:
2	3 2 3 3 2	:	:	:
2	3 2 3 3 1	:	:	:
0	-1 -2 0 0	-2 -2 -2	-2 0 -1 1 0	1
2	1 1 1 3 1	1 1 1	1 1 1 1 1	1
0	0 0 0 1	1 1	1 1 1 1	1

# Universal algorithm

LS-CGAT

sbars.impb.ru/aligner/needle.htm#

## Long Sequences Customizable Global Alignment Tool

Copy parameters from...

Enter, Edit or  , where sequence is taken from first residue  to last residue  (QUERY)

```
>Sequence A
GTTTTGCGTG GTAAAATCGG CAGATATTCG ATCGGTACAC TCATTGTTGG GTAGCAATCC
GGTTGACTCA CGGATCTTCC ACGCGA
```

Enter, Edit or  , where sequence is taken from first residue  to last residue  (SUBJECT)

```
>Sequence B
TTGATTATTG TTATAAACAG TGGCGTTTCC CCGGCTTCTG GGTGACCTCA TTCGCGAATC
ACGGCGTTCG AGACCGCATC TTC
```

### Score matrix

	A	T	G	C	S
A	2	-3	-3	-3	-3
T	-3	2	-3	-3	-3
G	-3	-3	2	-3	-3

case sensitive

### Gap penalties for Query

Open	Extend	Model
<input type="text" value="7"/>	<input type="text" value="2"/>	<input type="button" value="Affine"/>
<input type="text" value="7"/>	<input type="text" value="2"/>	<input type="button" value="End gaps"/>

### Gaps for Subject

<input type="text" value="7"/>	<input type="text" value="2"/>	<input type="button" value="Affine"/>
<input type="text" value="7"/>	<input type="text" value="2"/>	<input type="button" value="End gaps"/>

### Gap priority (back-tracing)

Subject > Query > No gap

### Parallel computing

Max cores  
 Query's split  
 Subject's split  
 Manual

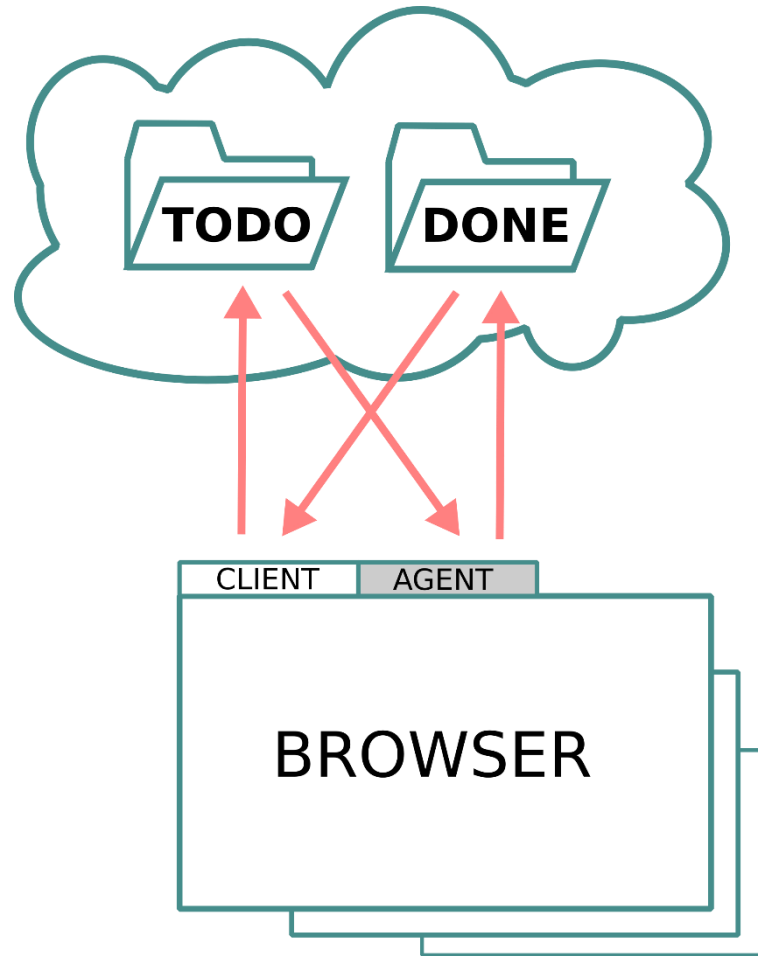
### Sequence Alignment

**IDENTITY** (overall percentage): 44/93 = 47.3%; **SCORE**: -67  
**IDENTITY** (overlap percentage): 44/86 = 51.2%; **TIMING**: 0.44/0.31/0.13 s

```
Query 1  GTTTTGCGTGTTAAAATCGGCAGATATTC----GATC---GGTACACTCATTGTTGGGTA  53
          | |  ||  ||  ||  ||  |  |||  |  ||  |||  ||||  |  |  |  |
Sbjct 1  TTGATTATTGTTATAAACAGTGGCGTTTCCCCGGCTTCTGGGTGACCTCAT--TCGCGAA  58

Query 54 GCAATCCGGTTGACTCACGGATCTTCCACGCGA  86
          ||  ||  |  ||  |  ||  ||||  |
Sbjct 59 TCACGGCGTTCGAGAC-CGCATCTTC  83
```

# Distributed cloud computing browser system for web application



# Results

Random DNA sequences have a 25% similarity in the case of Hamming distance and 50% similarity when adding inserts / deletions

Editing the replacement matrix allows one to solve biotechnological issues, for example, searching for a complex motive ATGCNNNNATGC