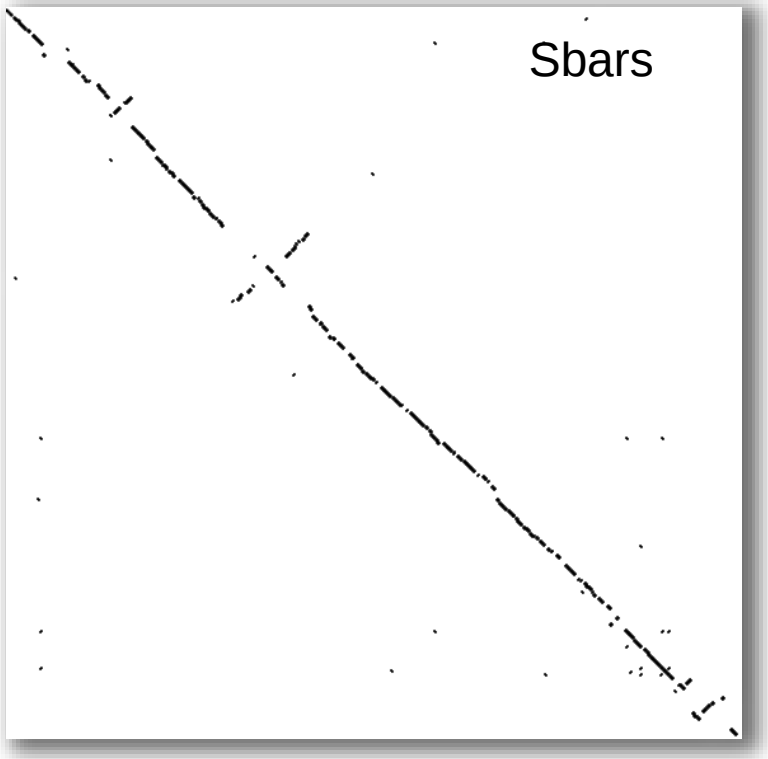
Optimization of parallel algorithm for global alignment of long sequences

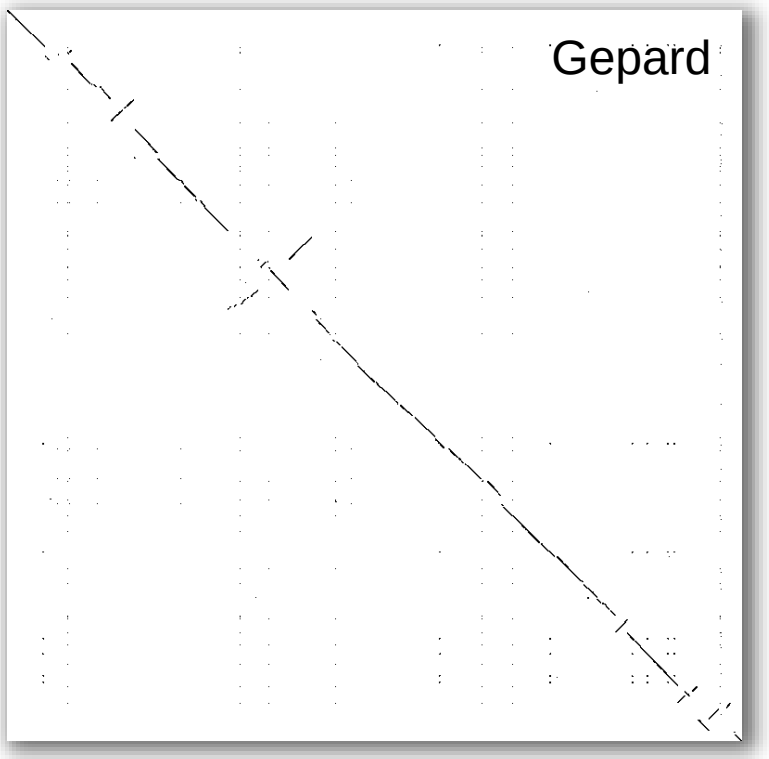
A.N. Pankratov, R.K.Tetuev, M.I.Pyatkov

IMPB RAS – the branch of KIAM RAS, Pushchino

Escherihia coli K12

Shigella flexnery 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.

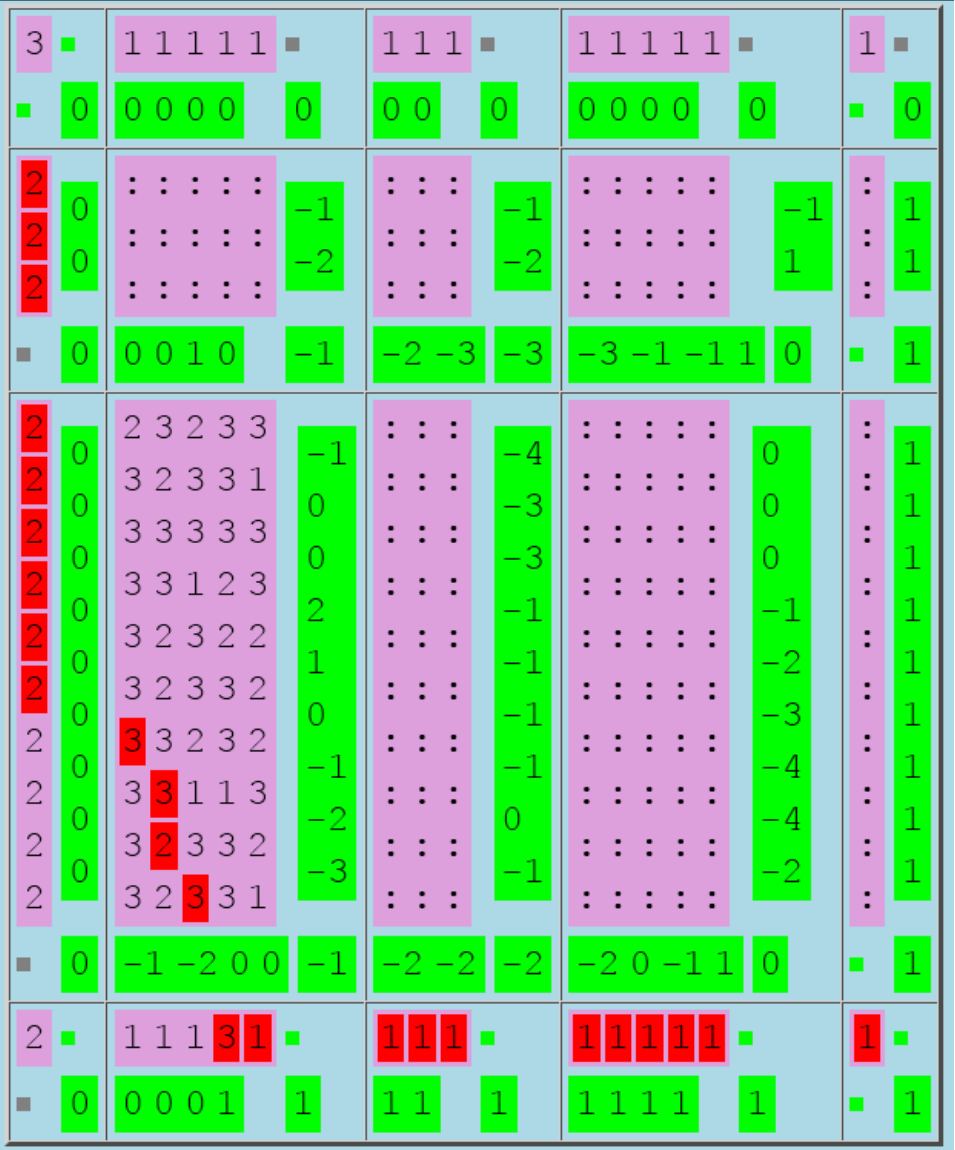
Global alignment of long sequences

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

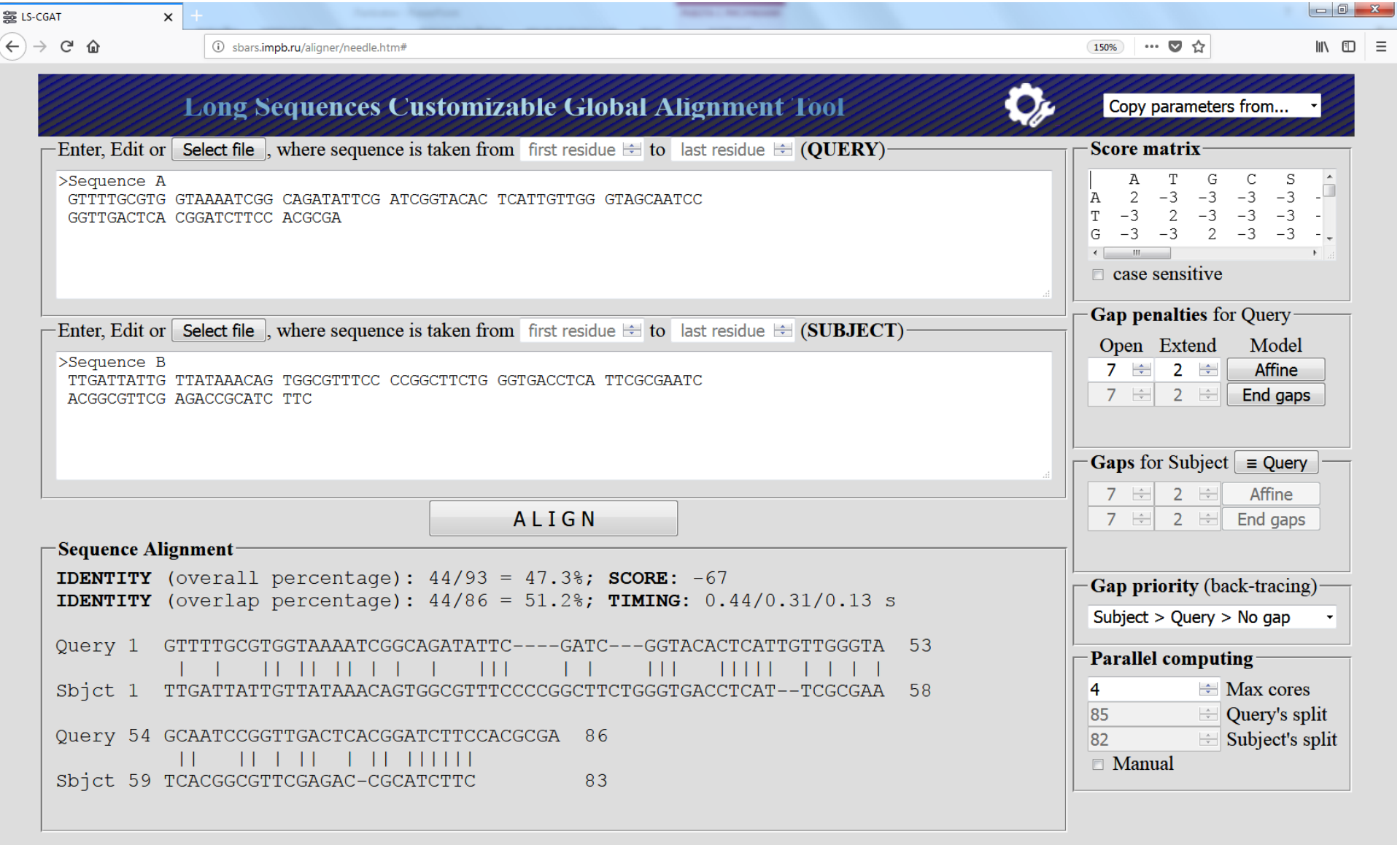
Computational complexity
~ O(n^2)
Space complexity
(Needleman&Wunsh, 1970)
~ O(n^2)

(Miller&Myers, 1988, Dryga, 2006)
~ O(n)

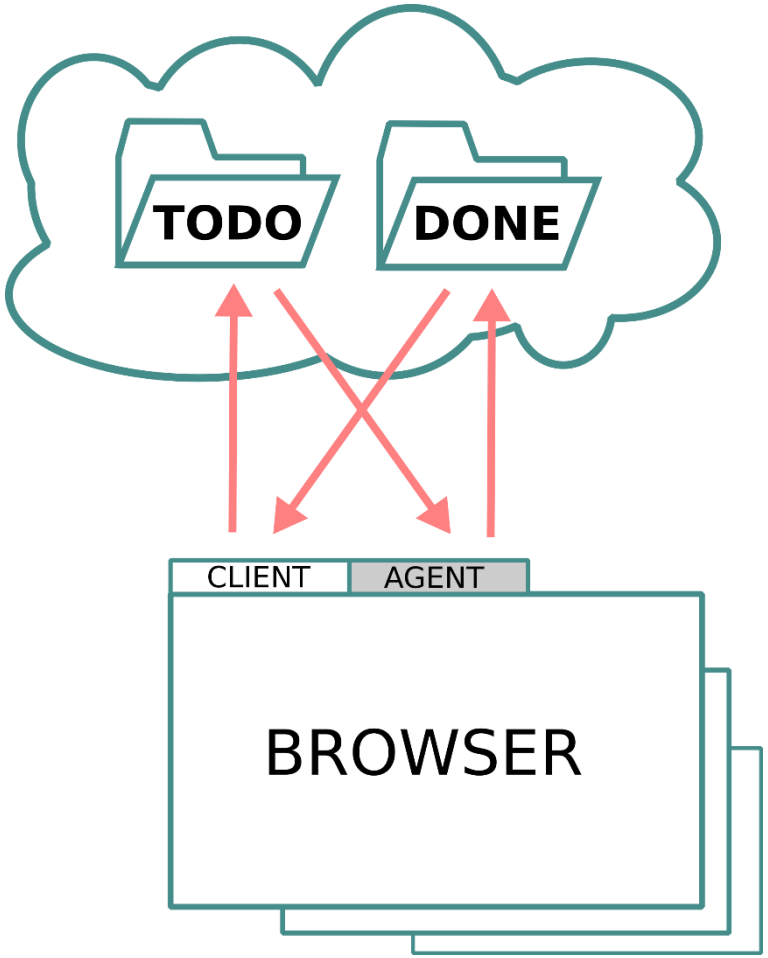
(Tetuev, Pyatkov, Pankratov, 2017)
~ O($n^{\frac{4}{3}}$)



Universal algorithm



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