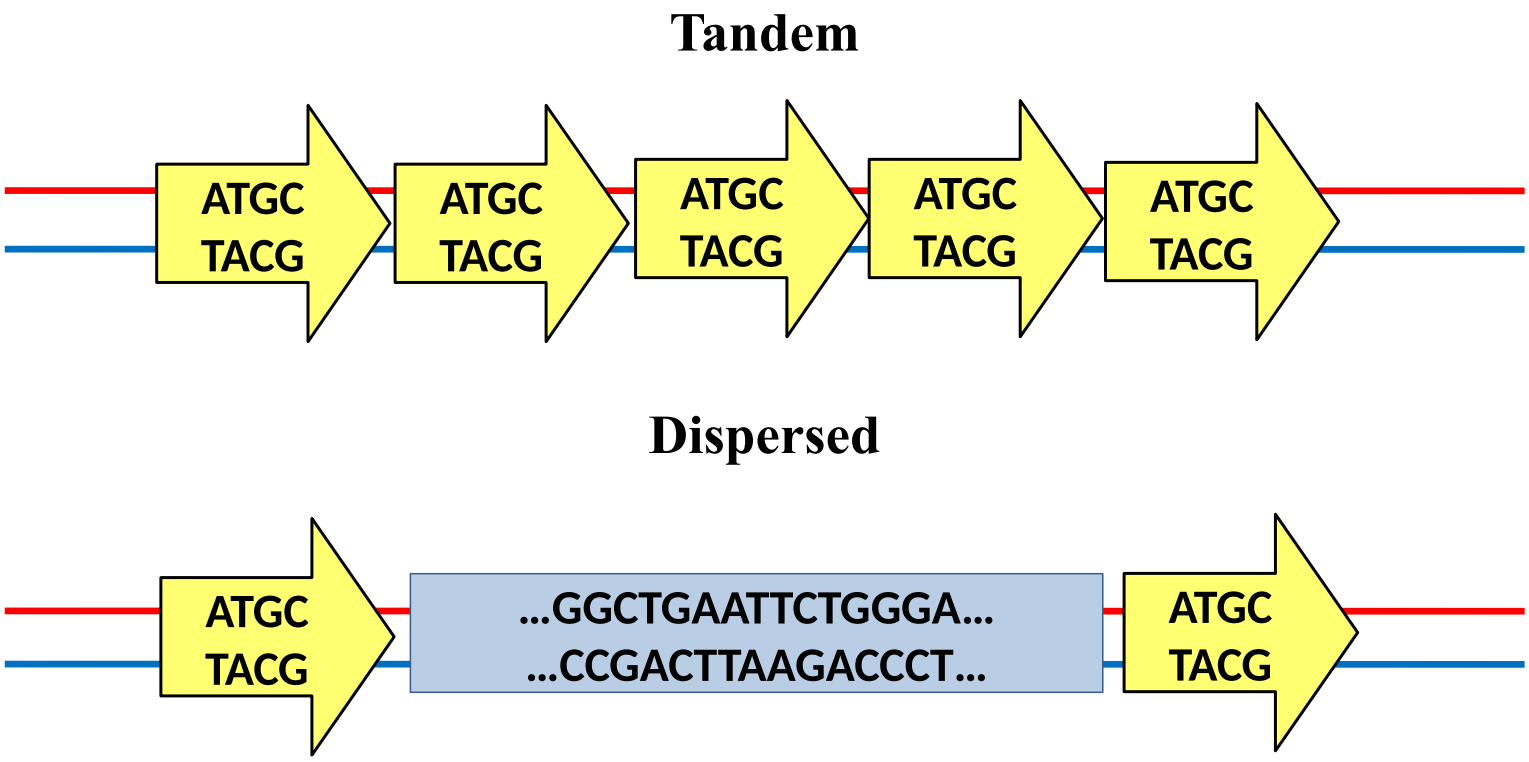
Multiple genome alignment based on a spectral-analytical approach

Pankratov Anton Nikolaevich

IMPB RAS – the branch of KIAM RAS

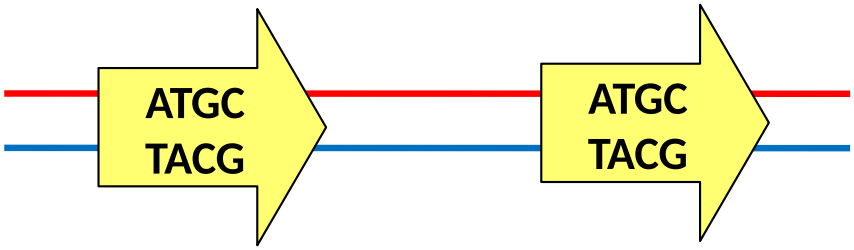
November 26, 2019

Types of repetitive sequences in genomes

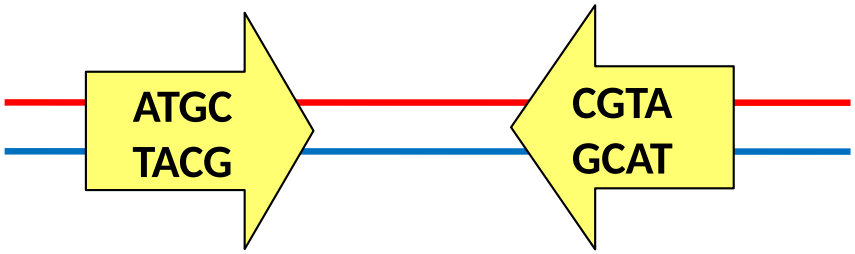


Different orientation of repeats in the double strand of DNA

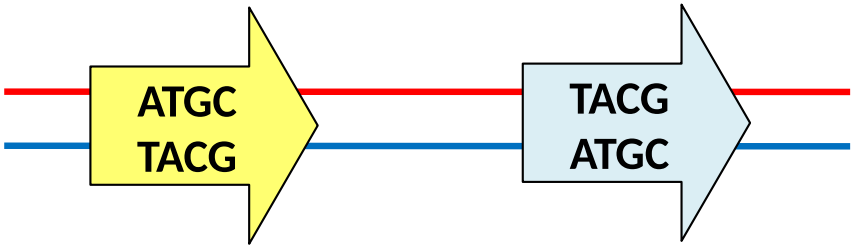
Direct



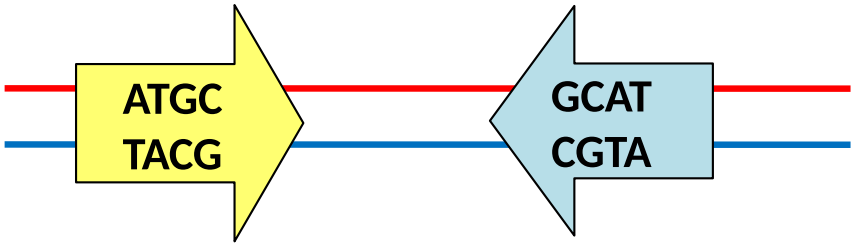
Reverse



Complement

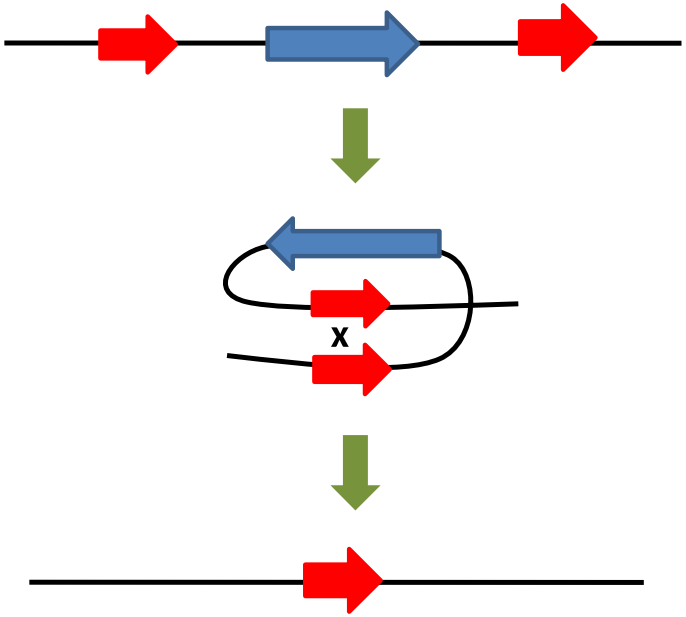


Inverted (reverse-complement)



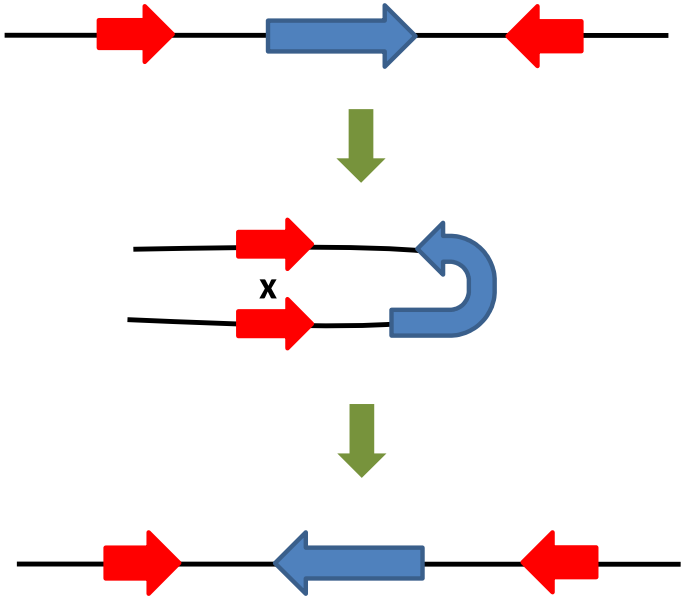
Recombination between repeats leads to block mutations

Deletion



X-linked ichthyosis (STS).
Yen et al. Cell. 1990.

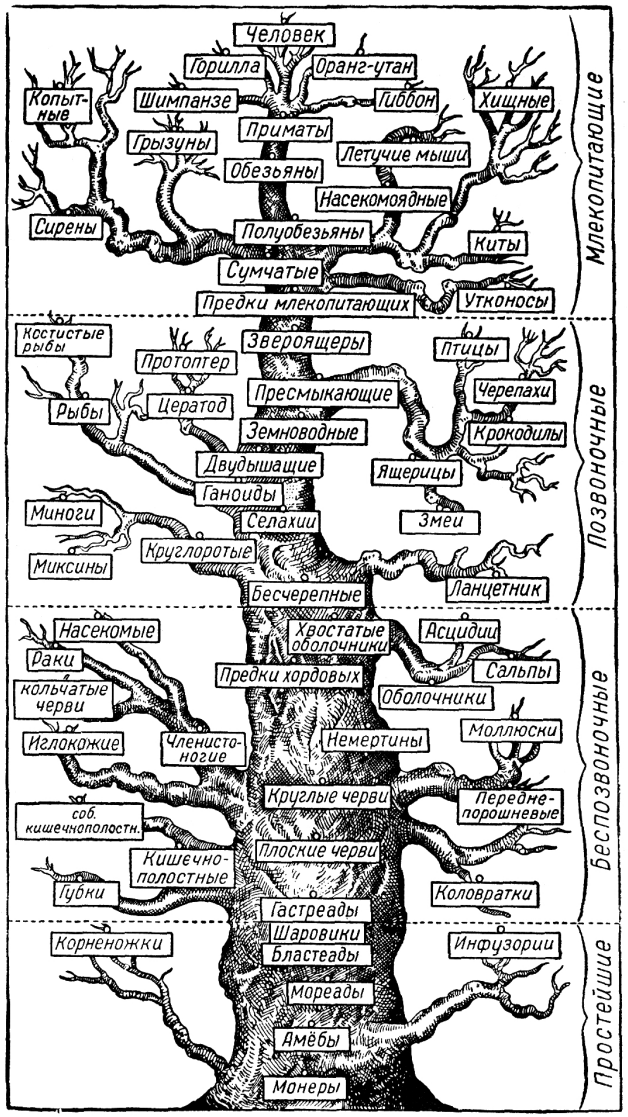
Inversion



Muscular dystrophy (EMD)
Small et al. Nat Genet. 1997.
Hemophilia (F8)
Naylor et al. Hum Mol Genet .1995

The use of repeats in evolutionary research

- Evolutionary research is aimed at identifying the relationship between different taxa. Repetitions are phylogenetic markers.
- The study of the genomes of close organisms. Repeats are extended sites of similarity.

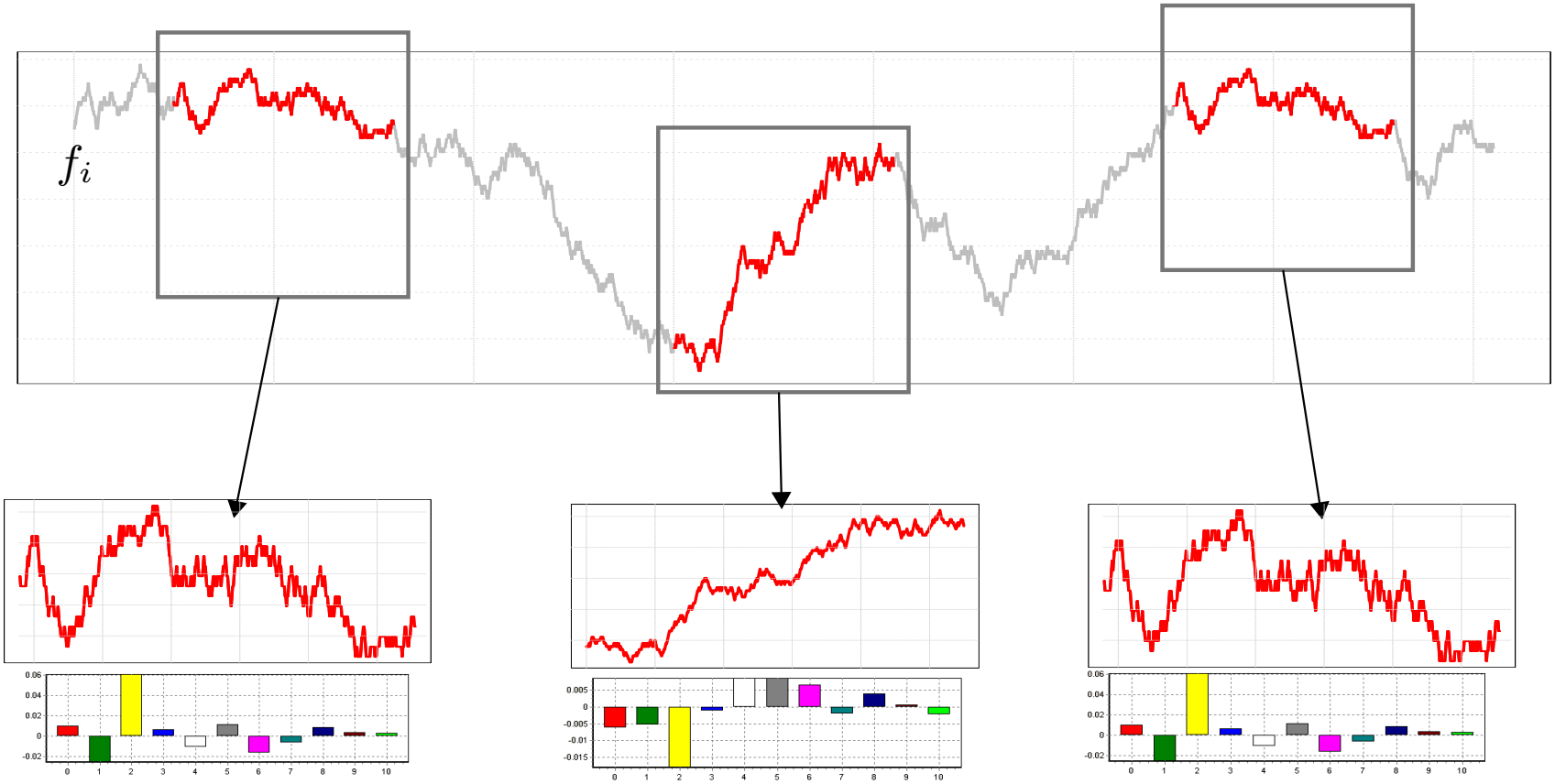


Repeats Search Challenges

- The appearance of mutations - repeats become inaccurate
- The large length of the compared sequences (of the order of 10^9 nucleotides) times the number of analyzed sequences (hundreds to thousands)
- The analysis is quadratic depending on the length of the sequences
- The determination of orthology rather than simple homology
- The determination of reference sequences (pan-genome)

- Research paradigm: to construct a repeat search method based not on letter-by-letter comparison, but on some sort of numerical analysis of nucleotide sequences.

The main idea of the spectral-analytical approach

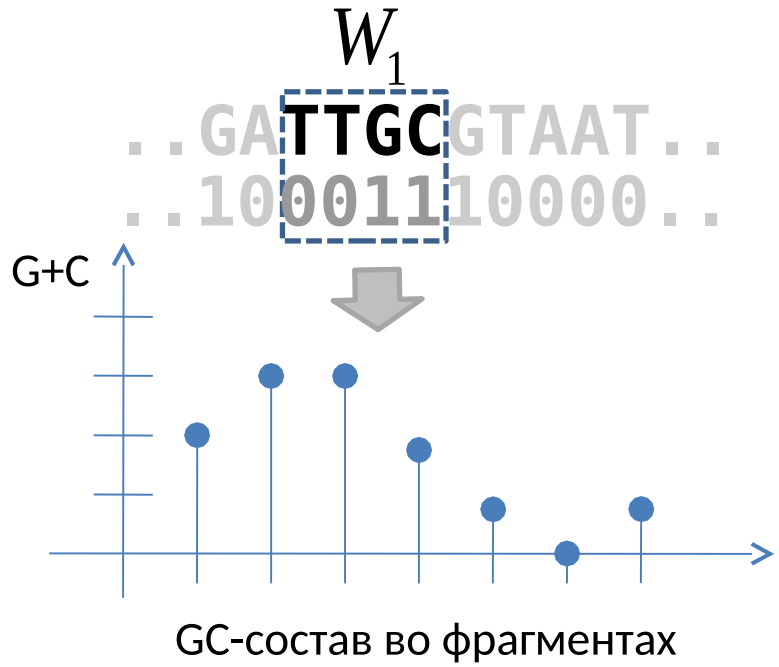


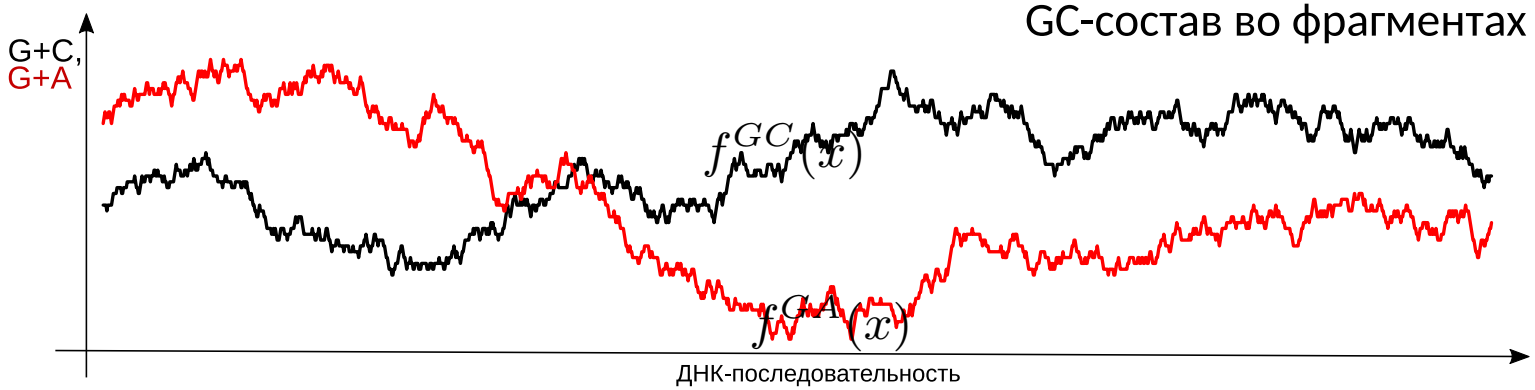
Dedus F.F., Kulikova L.I., Makhortykh S.A., Nazipova N.N., Pankratov A.N. and Tetuev R.K. Analytical Recognition Methods for Repeated Structures in Genomes. Doklady Mathematics, 2006, Vol. 74, №3, pp. 926-929

Conversion of the nucleotide sequence into an analog function

Calculation of the proportion of guanine and cytosine in DNA fragments (GC%):

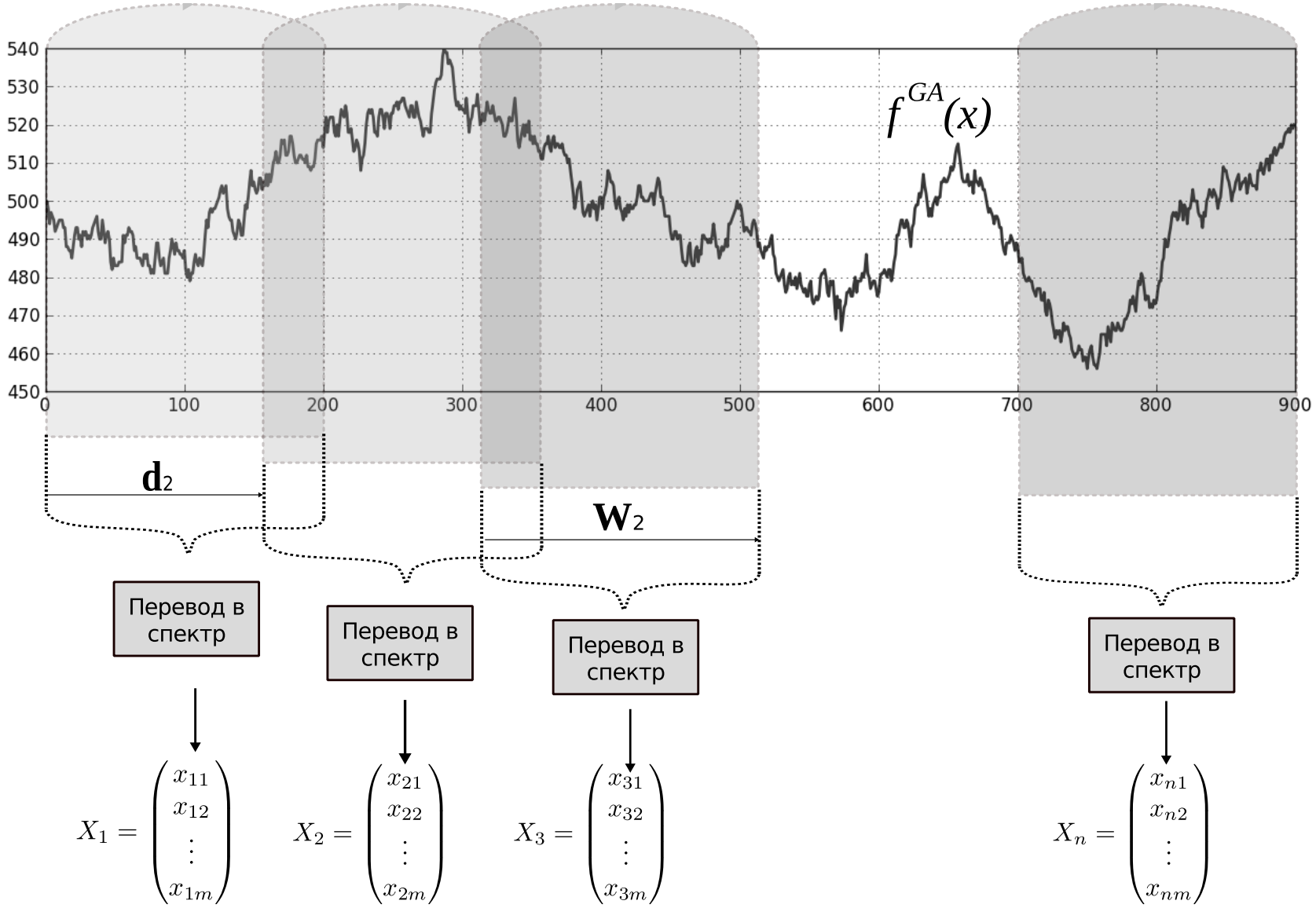
- continuous function
- depends on scale parameter W_1





Theorem on inverse conversion: unambiguous recovery of sequence requires two linearly independent functions (GC%, GA%)

Conversion of analog functions into spectra of Fourier coefficients



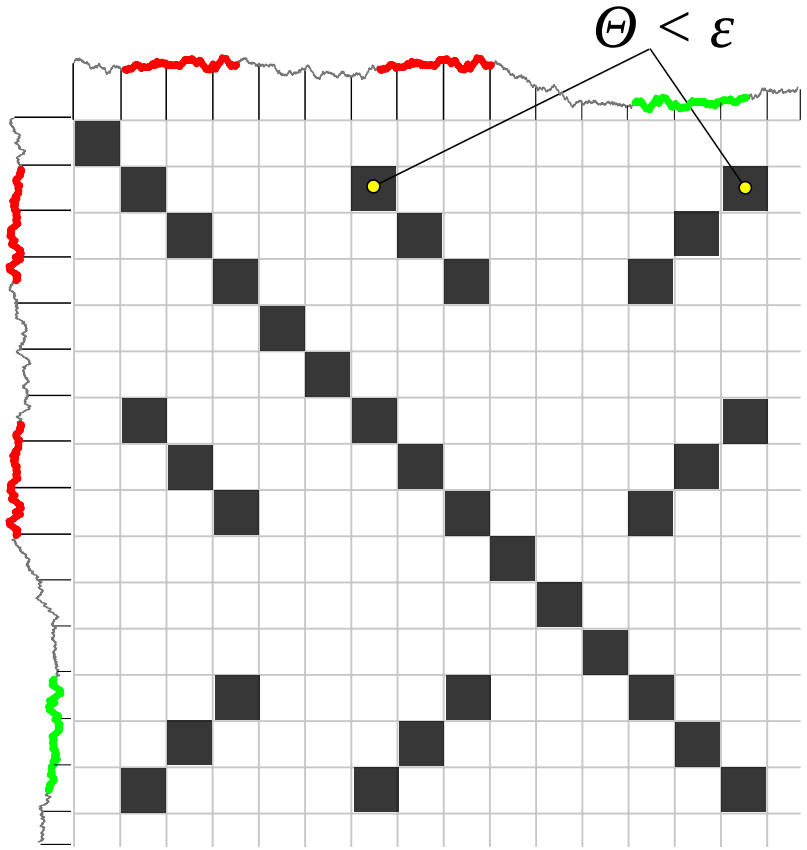
Comparison of coefficients and construction of a dot matrix

$$\sum_{i=0}^{L-1}(A_i - B_i)^2\pi \leq \int_{-\pi}^{\pi}(f-g)^2 \leq 2\pi W_1^2$$

$$\Theta(f,g) = \frac{1}{2W_1^2}\sum_{i=0}^{L-1}(A_i - B_i)^2 \leq \varepsilon \leq 1$$

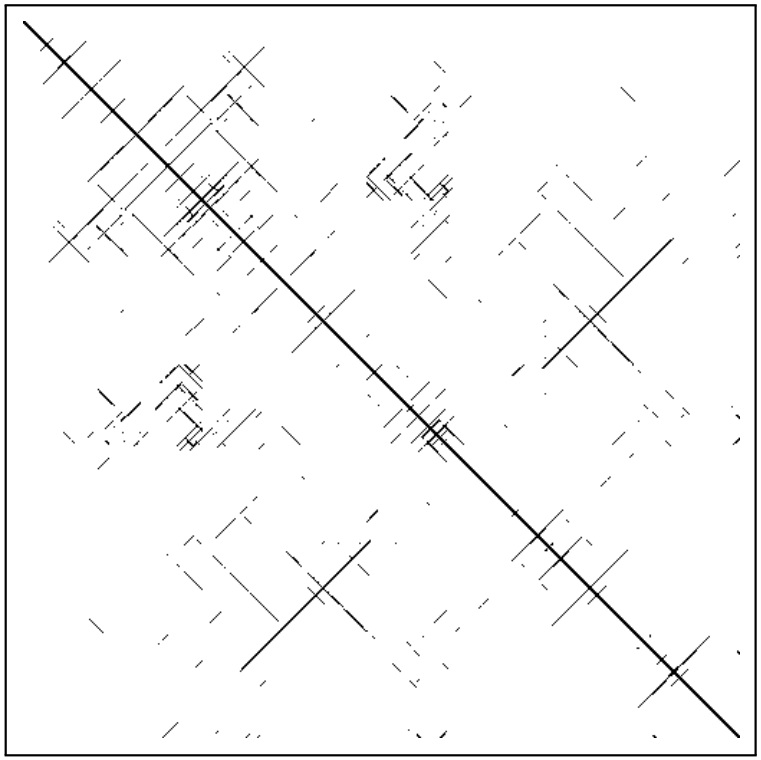
Metric

- bounded
- scale invariant
- monotonic in the number of coefficients

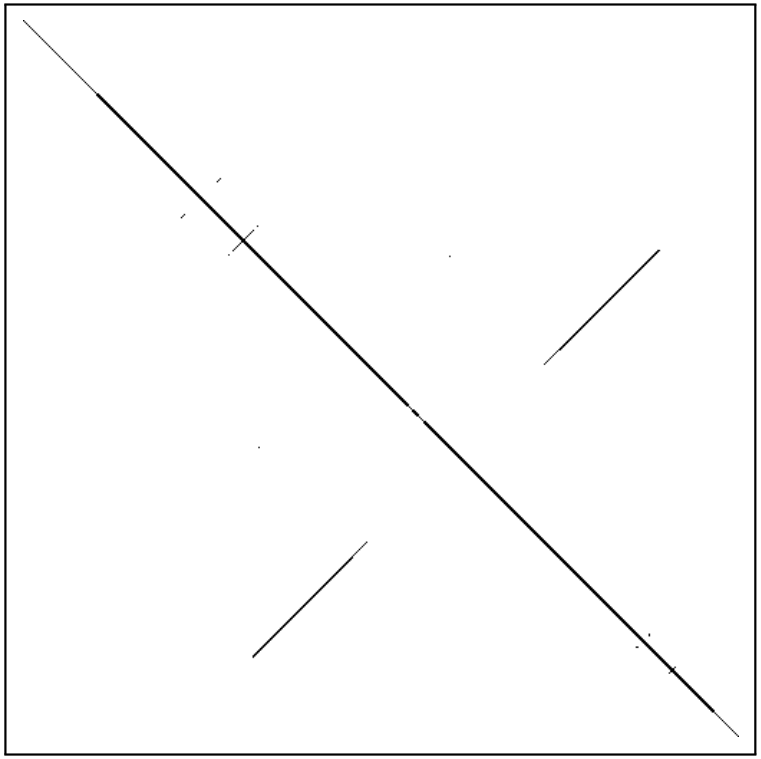


Each matrix point is the result of pairwise comparison of coefficient vectors

Decision rule: conjunction of threshold decision rules for each analog function



GC%

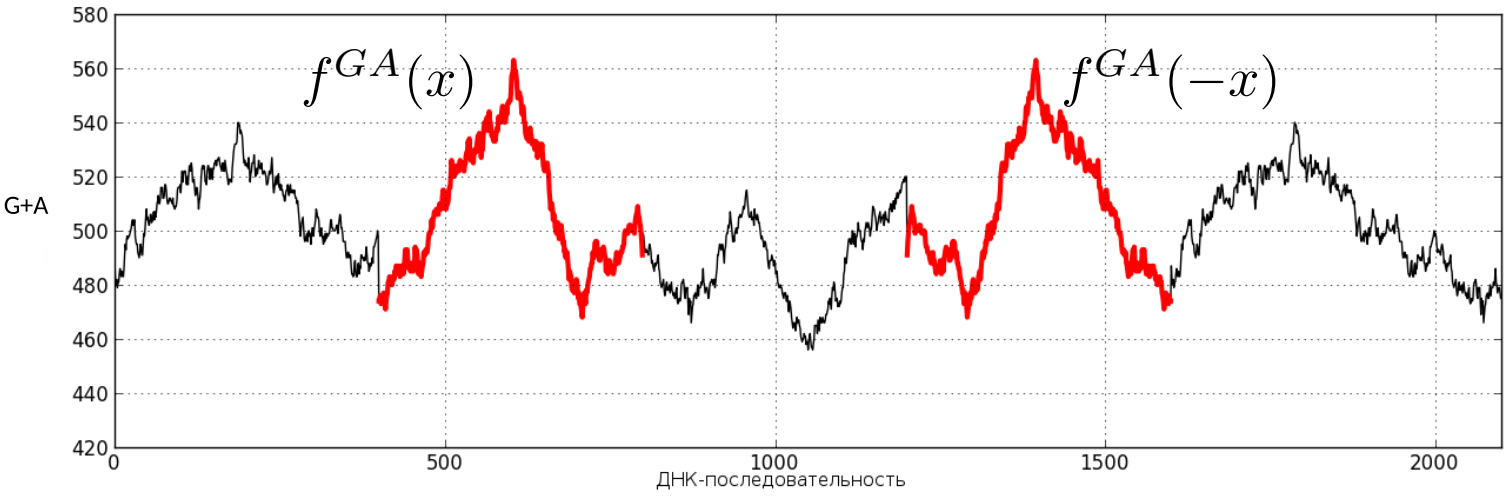


(GC%) & (GA%)

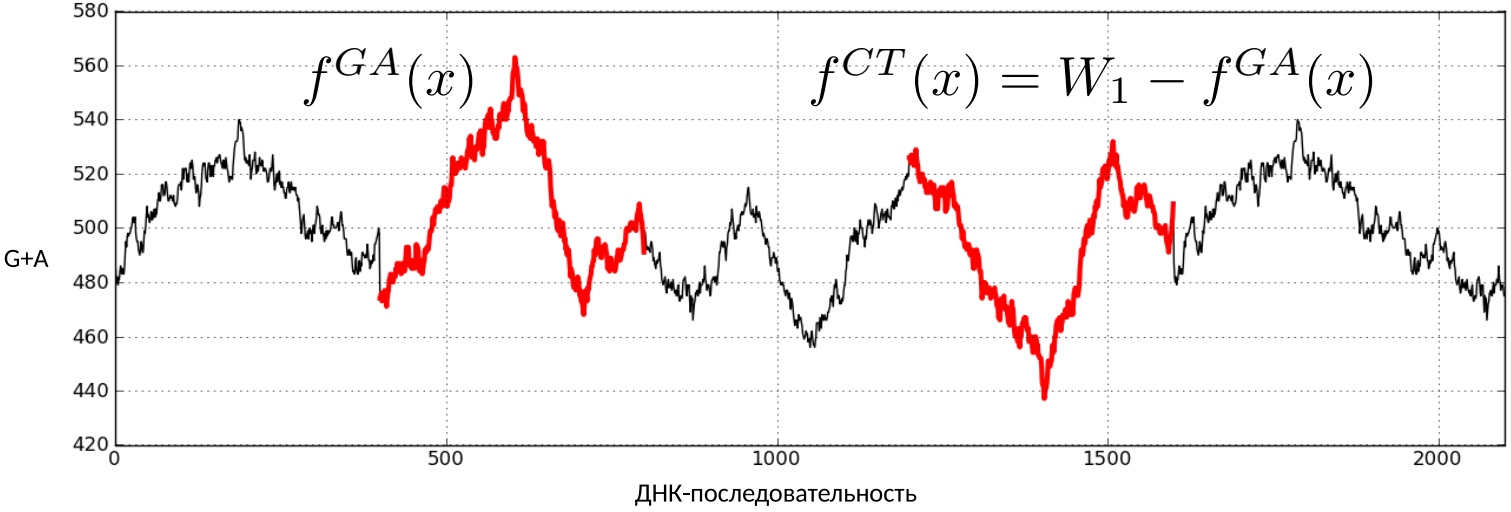
$$(\Theta(f^{GC}, g^{GC}) \leq \varepsilon)\&(\Theta(f^{GA}, g^{GA}) \leq \varepsilon)$$

Reverse and complement transform of corresponding analog functions

Reverse transform

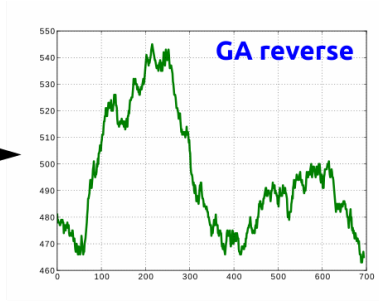


Complement transform

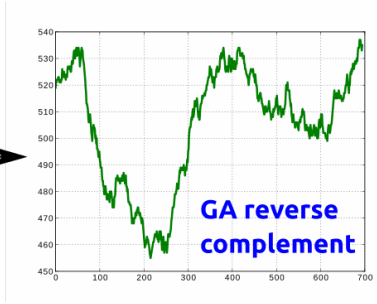


Transformations over analog functions lead to corresponding transformations in the space of Fourier coefficients

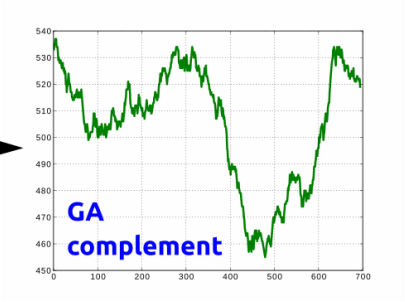
$$\begin{cases} C_{2k}^{GArev} = C_{2k}^{GA} \\ C_{2k+1}^{GArev} = -C_{2k+1}^{GA} \end{cases}$$



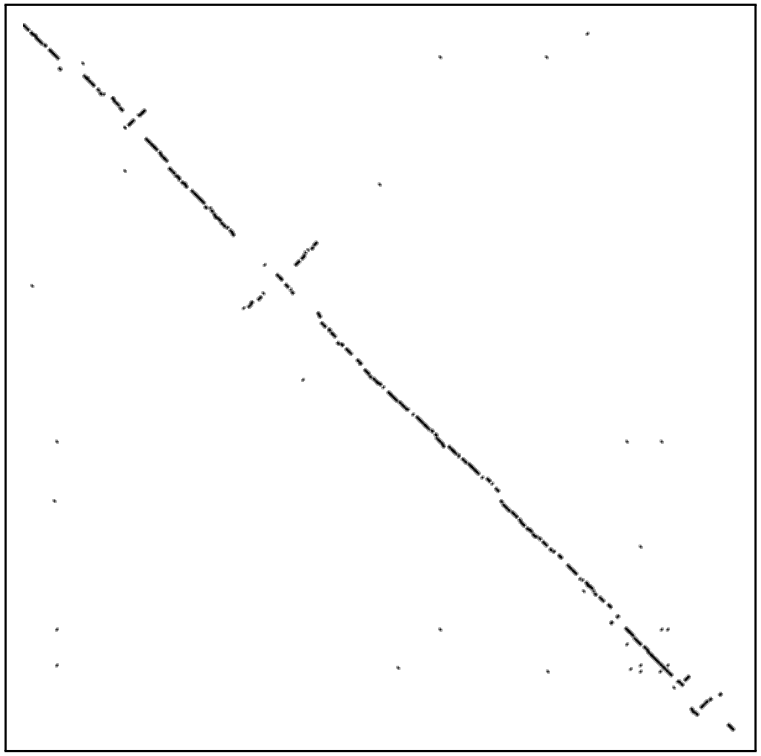
$$C_0^{CT} = \sqrt{2}W_1 - C_0^{GA}$$
$$\begin{cases} C_{2k}^{CT} = -C_{2k}^{GA} \\ C_{2k+1}^{CT} = C_{2k+1}^{GA} \end{cases}$$



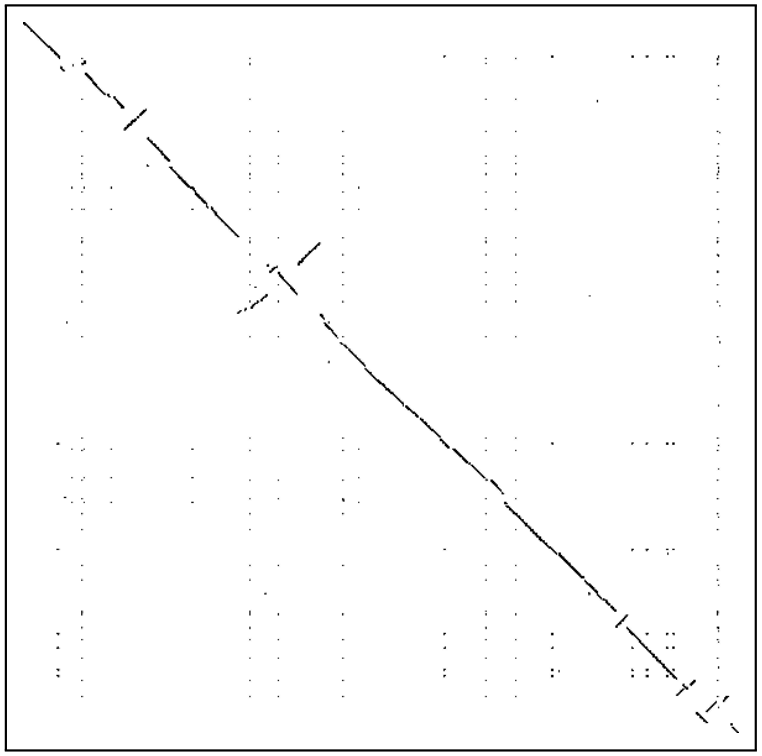
$$C_0^{CT} = \sqrt{2}W_1 - C_0^{GA}$$
$$C_k^{CT} = -C_k^{GA}$$



Dotplot of the bacterial genome by different methods



SBARS (Pyatkov M., Pankratov A. Bioinformatics. 2007.)

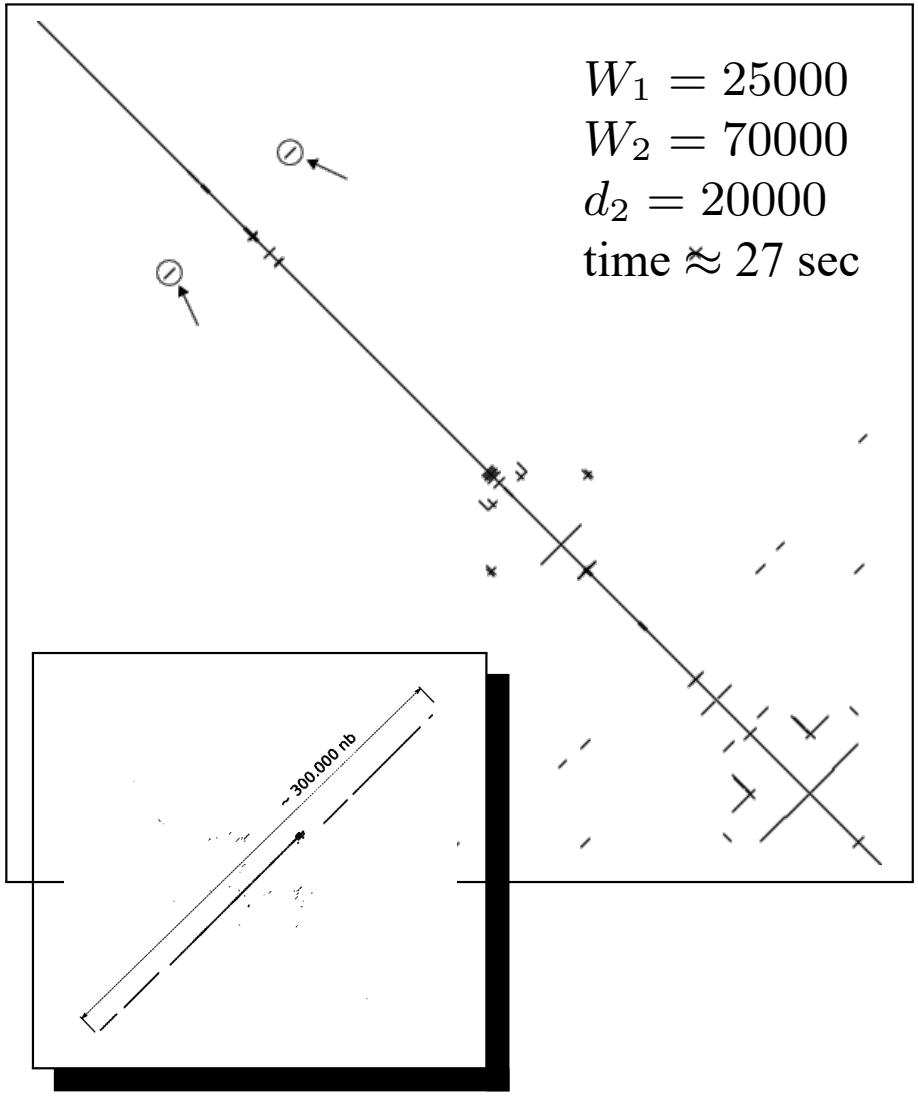


Gepard (Krumsiek J. et al. Bioinformatics. 2007.)

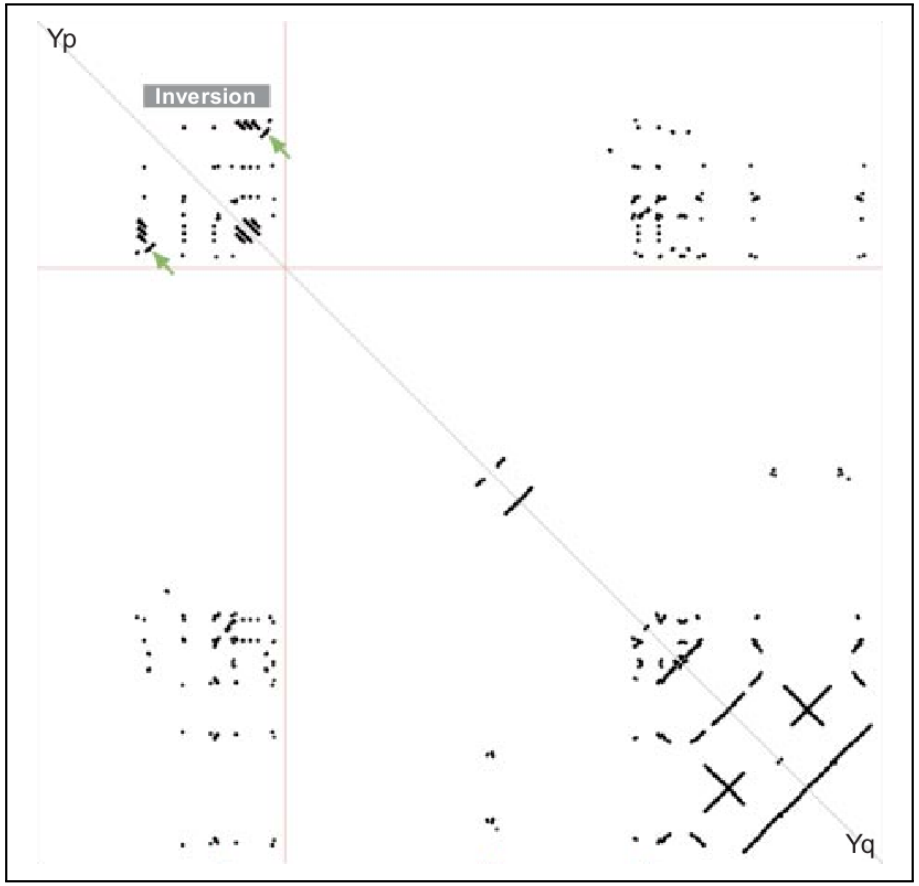
Sequence length	Gepard	SBARS
100000 b.p.	≈ 1 сек	≈ 1 сек
1000000 b.p.	5 sec	5 seq
5000000 b.p.	45 sec	14 seq
Y chr (27000000 b.p.)	5 min	27 seq

Y human chromosome

in silico(by SBARS)

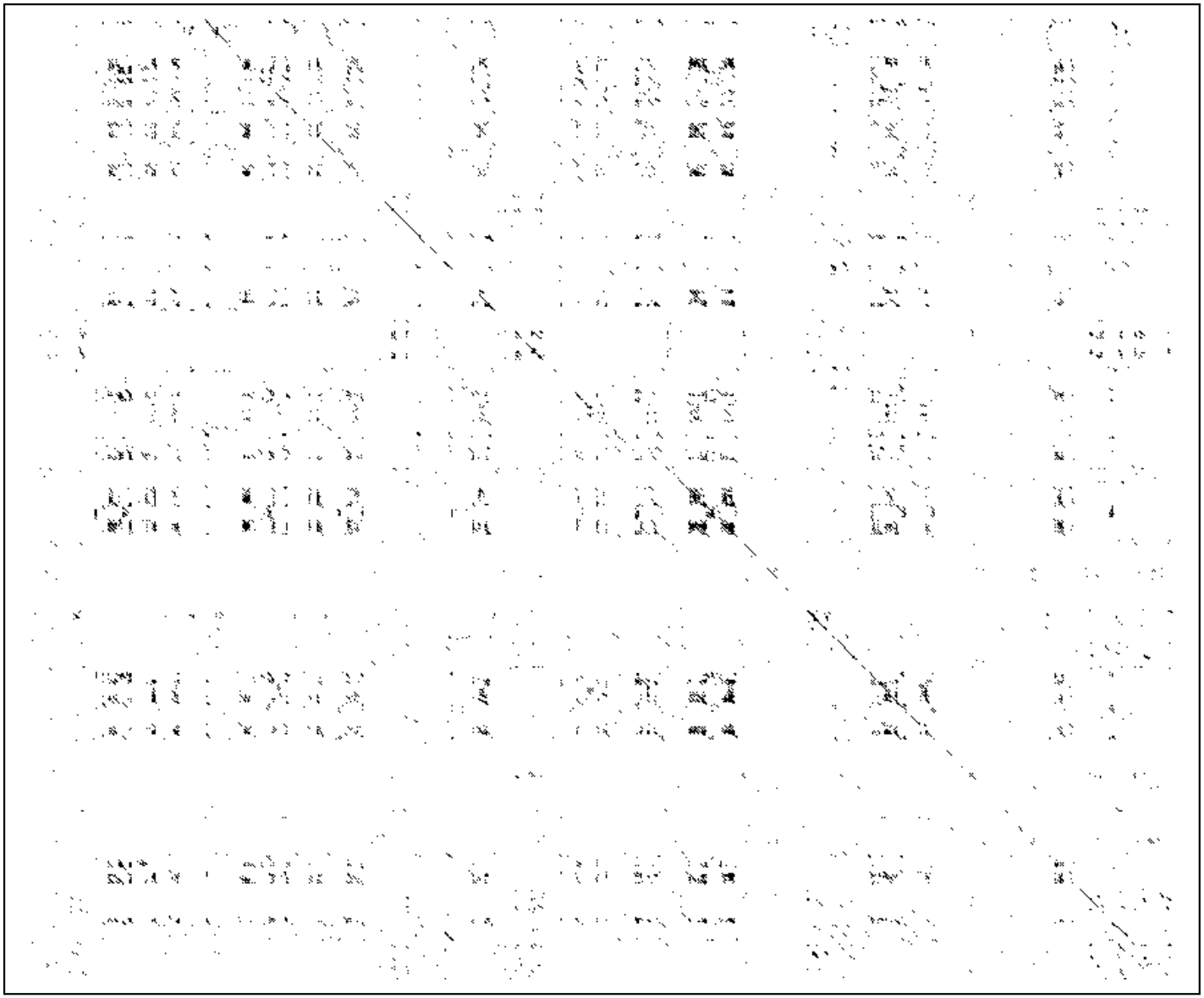


in vetro(by DNA hybrid)



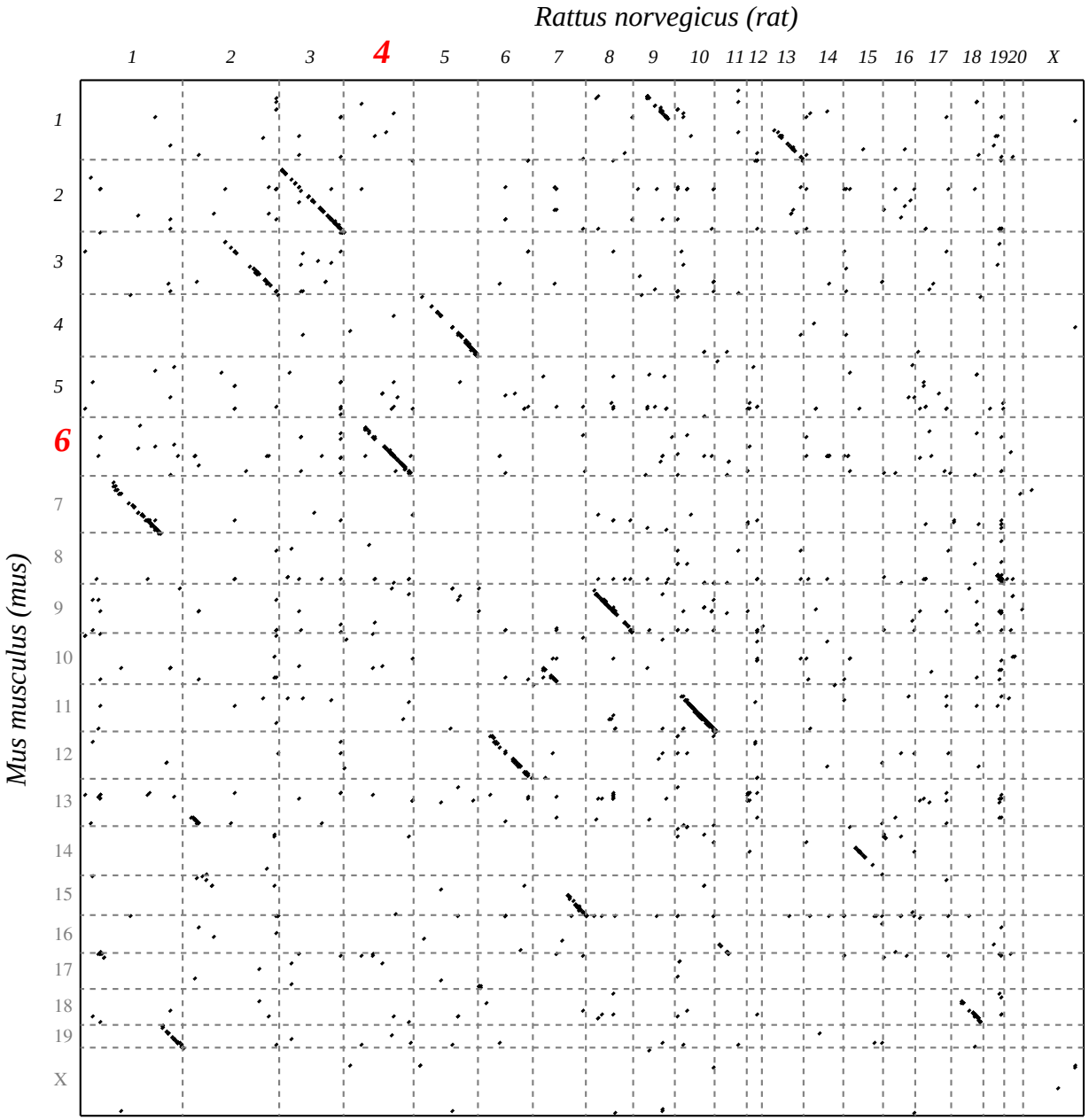
Tilford CA et al. Nature. 2001

6 mouse chromosome (Mus musculus) vs 4 rat chromosome (Rattus norvegicus)



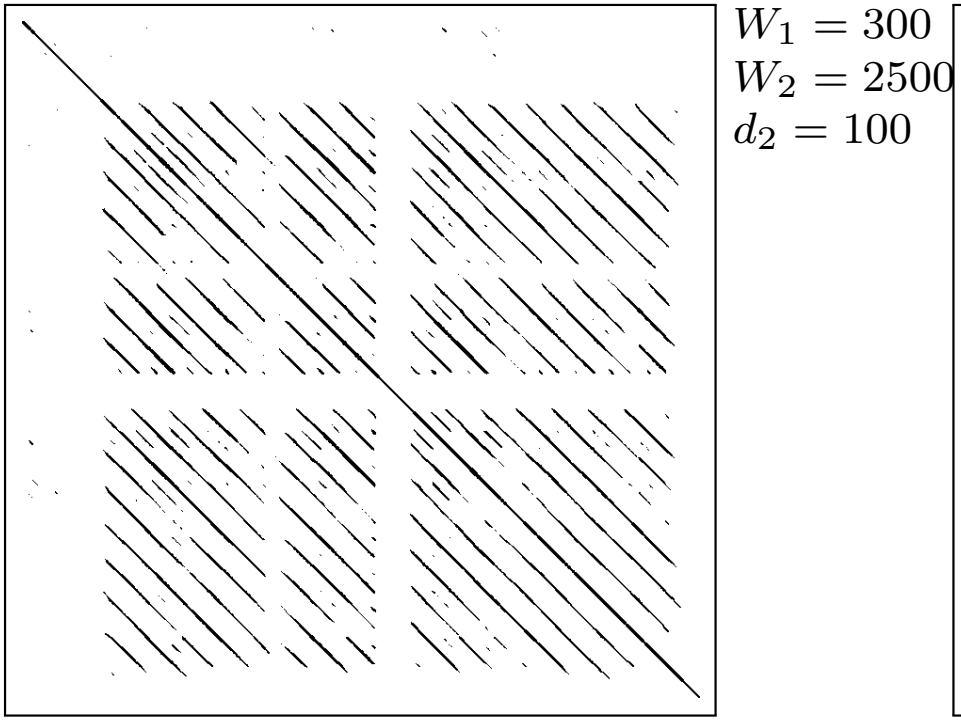
$W_1 = 10^5$
$W_2 = 7 * 10^5$
$d_2 = 2 * 10^5$
time $\approx$ 1 min
length $\approx 10^8$

Whole genome alignment of mouse and rat

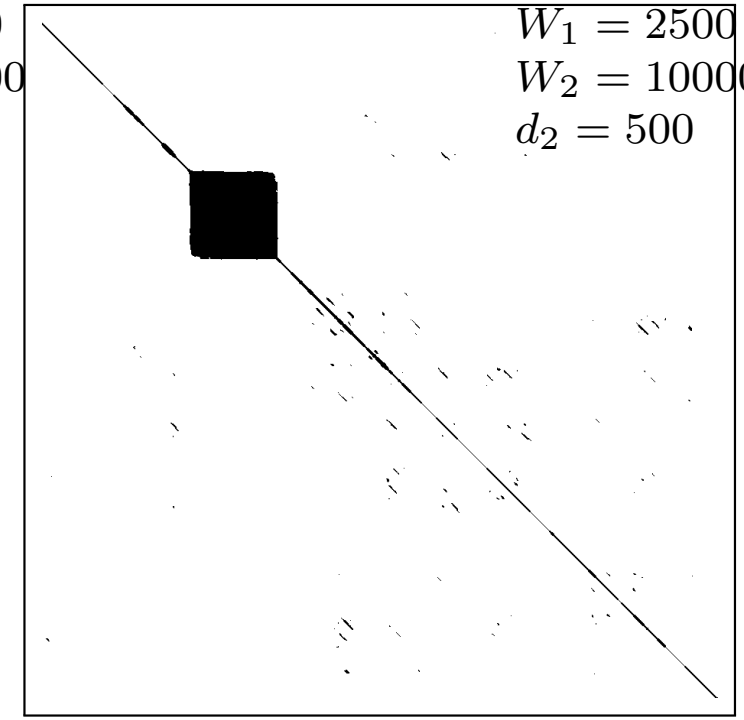


$W_1 = 2.5 * 10^6$
$W_2 = 10^7$
$d_2 = 2.5 * 10^6$
time $\approx$ 40 sec
length $\approx 10^9$

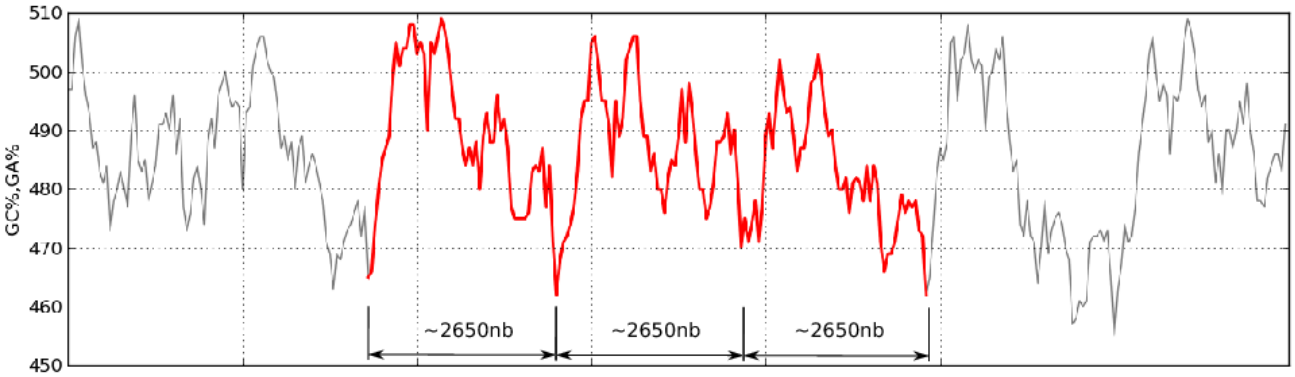
Tandem repeats



$W_1 = 300$
$W_2 = 2500$
$d_2 = 100$



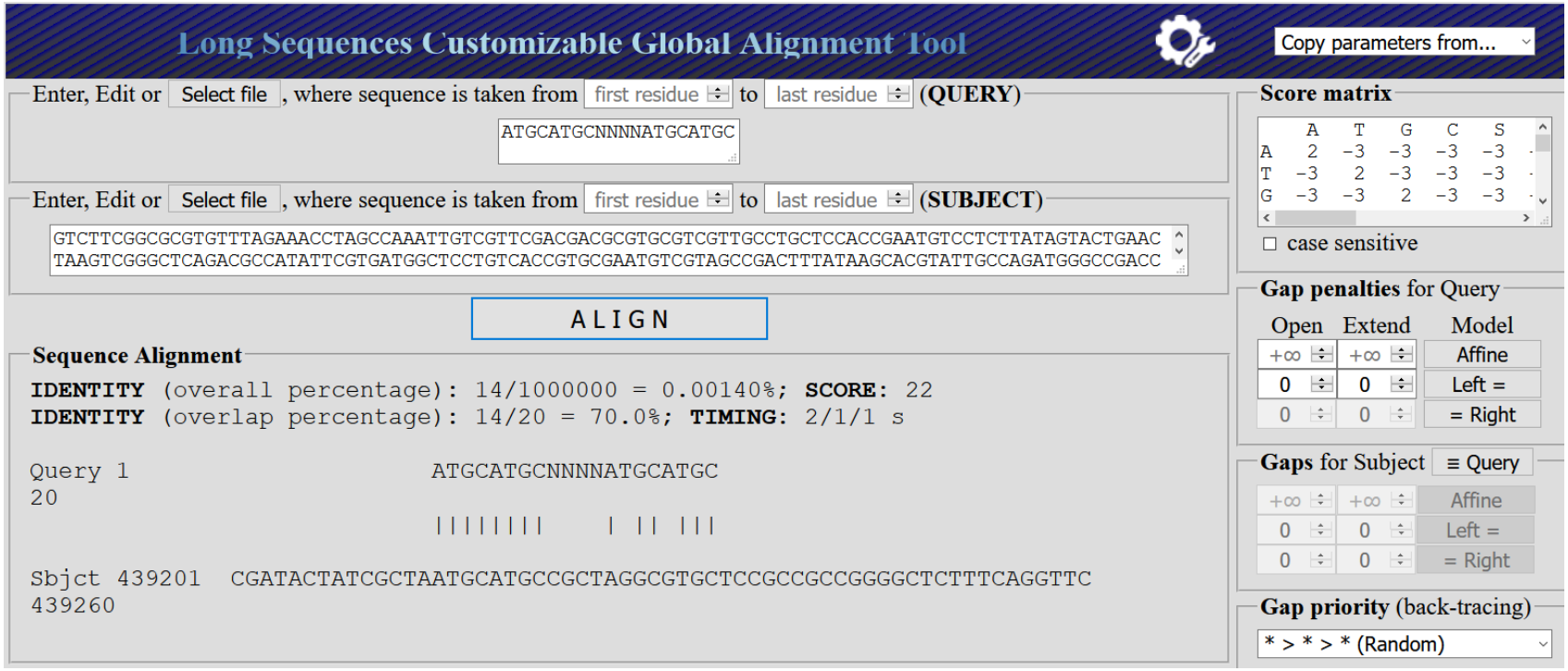
$W_1 = 2500$
$W_2 = 10000$
$d_2 = 500$



Differences
Between Copies:
Min: 4,2%
Max: 22,7%
Consensus: 47,0%

Pyatkov M.I., Filippov V.V., Pankratov A.N. Consensus of repeated region of rabbit chromosome 17 containing over 15 huge approximate tandem repeats. Repbase Reports. 2012. Vol.12, No.3.

Universal alignment tool for long repeats



Space complexity:

- $O(n^2)$ - Needleman & Wunsh (full matrix), 1970
- $O(n)$ - Miller & Myers (recursive), 1988; Dryga (grid & recursive), 2006
- $O(n^{4/3})$ - Tetuev, Pyatkov, Pankratov (grid), 2017