Быстрое распознавание повторов в генетических последовательностях на основе спектрально-аналитического метода

Панкратов А.Н., Пятков М.И., Руднев В.Р., Куликова Л.И.

Институт математических проблем биологии РАН

Виды повторов

Тандемные

…CGAGCATGGACTTTTTGAGCAC…TTTGAGCAC…CACGAGCCACGGA…

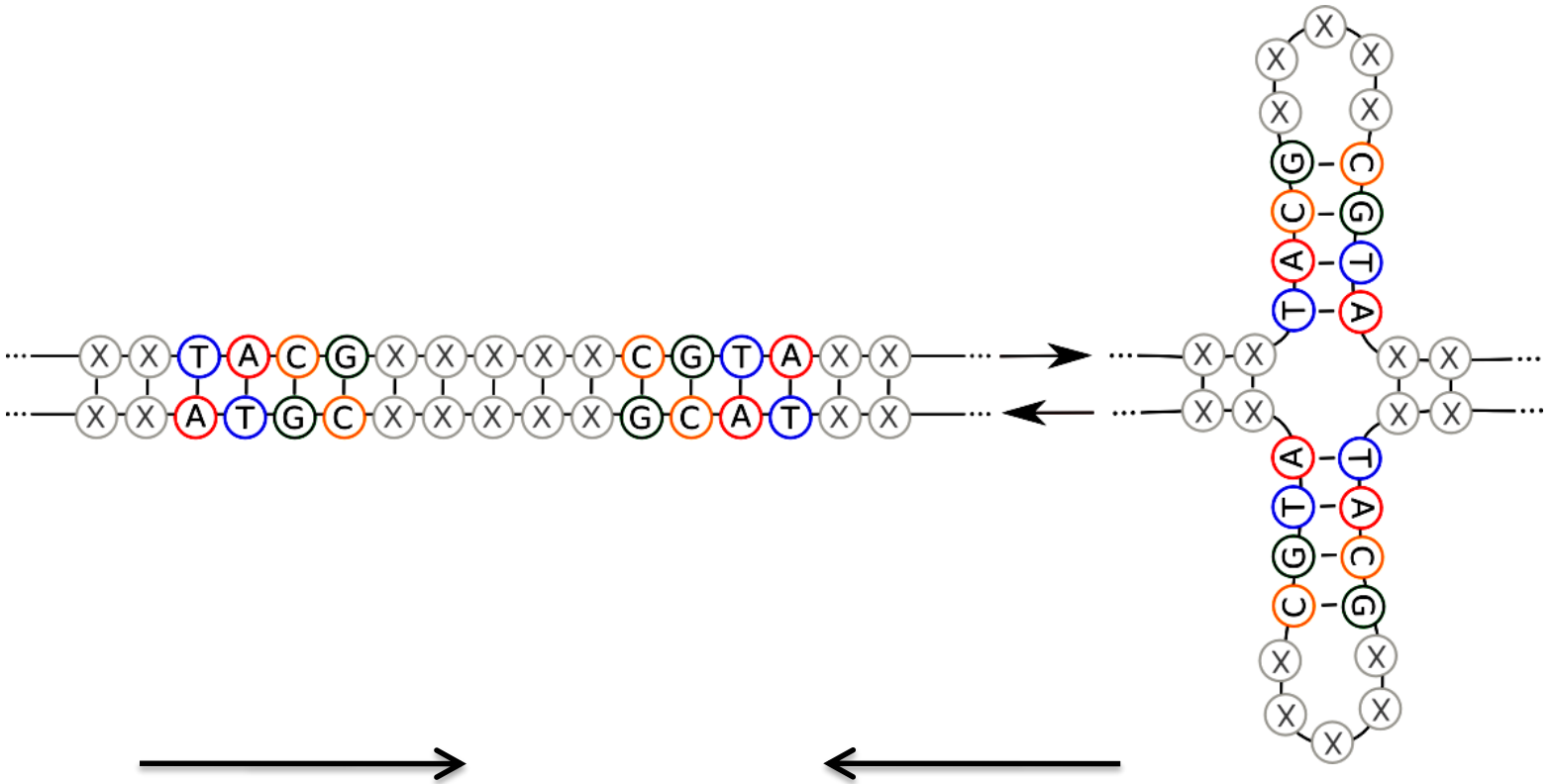
Разнесенные

…CGAGCATTTGAGCAC…TGGACTTTCACGAGCCACTTTGAGCAC…GGA…

Инвертированные

…CGAGCATTTGAGCAC…TGGACTTTCACGAGCCAC…GTGCTCAAAGGA…

Инвертированные повторы



…TACG……CGTA…

комплементарный

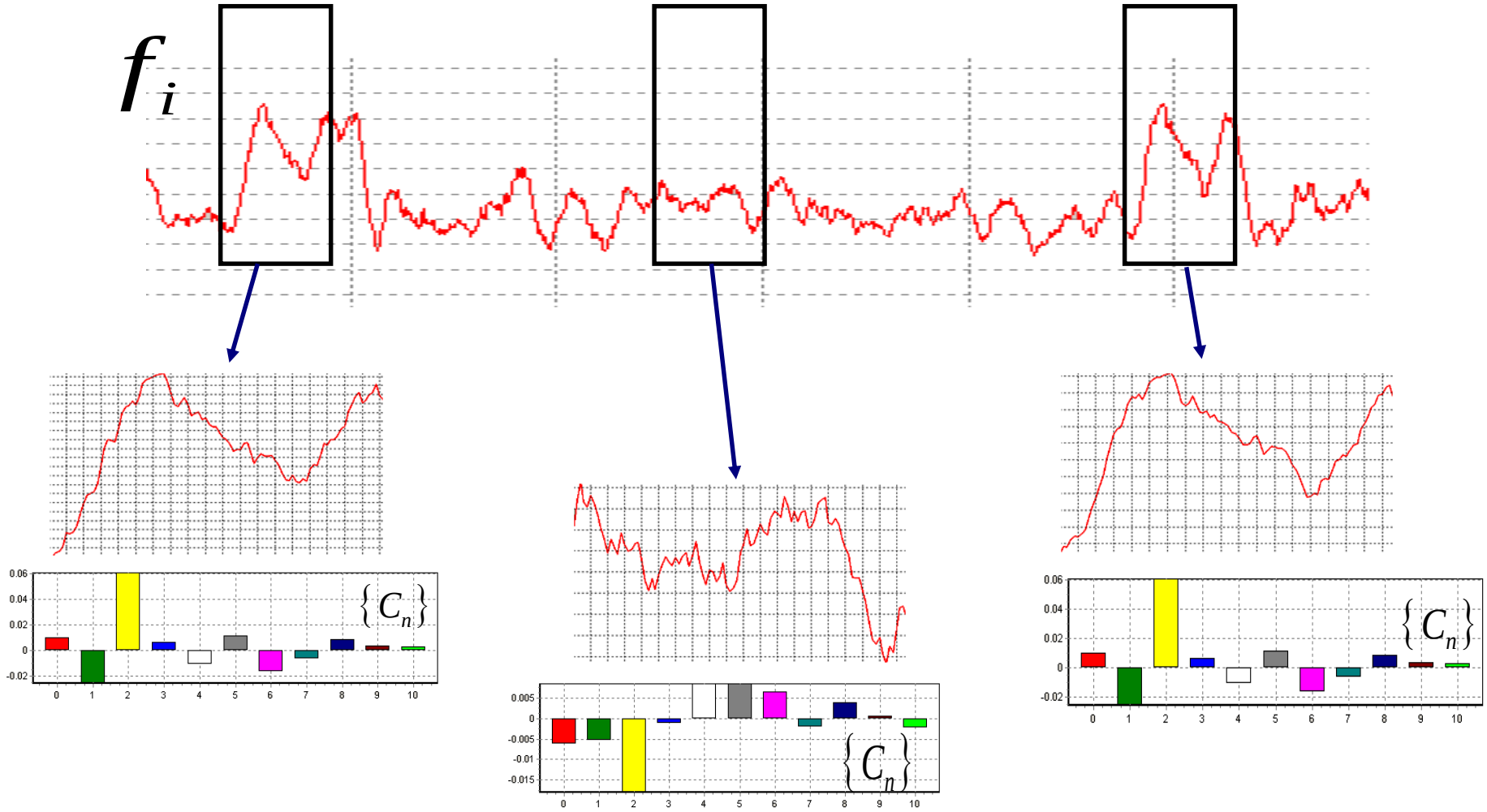
Проблемы поиска повторов

- Точечные мутации
- Протяженные повторы (>= 2000)
- Большая длина исследуемых последовательностей

(10^{6}-10^{9} нуклеотидов)

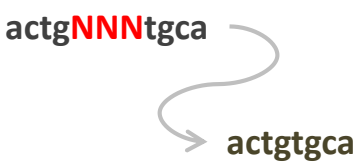
Гипотеза: методы из области обработки сигналов могут подойти лучше для поиска неточных протяженных последовательностей

Основная идея алгоритма

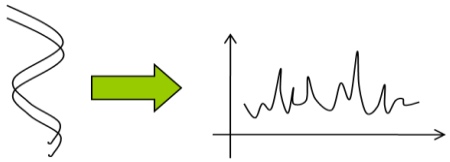


Ф.Ф.Дедус, Л.И.Куликова, С.А.Махортых, Н.Н.Назипова, А.Н.Панкратов, Р.К.Тетуев. Аналитические методы распознавания повторяющихся структур в геномах. ДАН, 2006, т. 411, №5, 1-46.

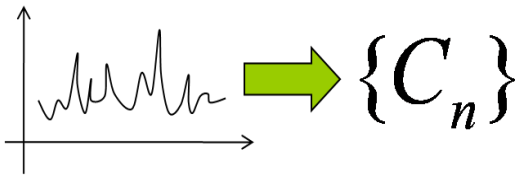
Общая схема работы алгоритма



Предварительная обработка ДНК последовательностей: удаление N регионов, 4-х кратное сжатие



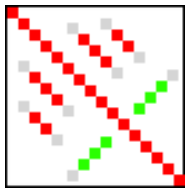
Получение эквивалентного функционального представления: GC%, GA% состав



Спектральное разложение (как минимум 10 –и кратное сжатие)

$$\theta(\{C_n\},\{C'_n\}) < \varepsilon$$

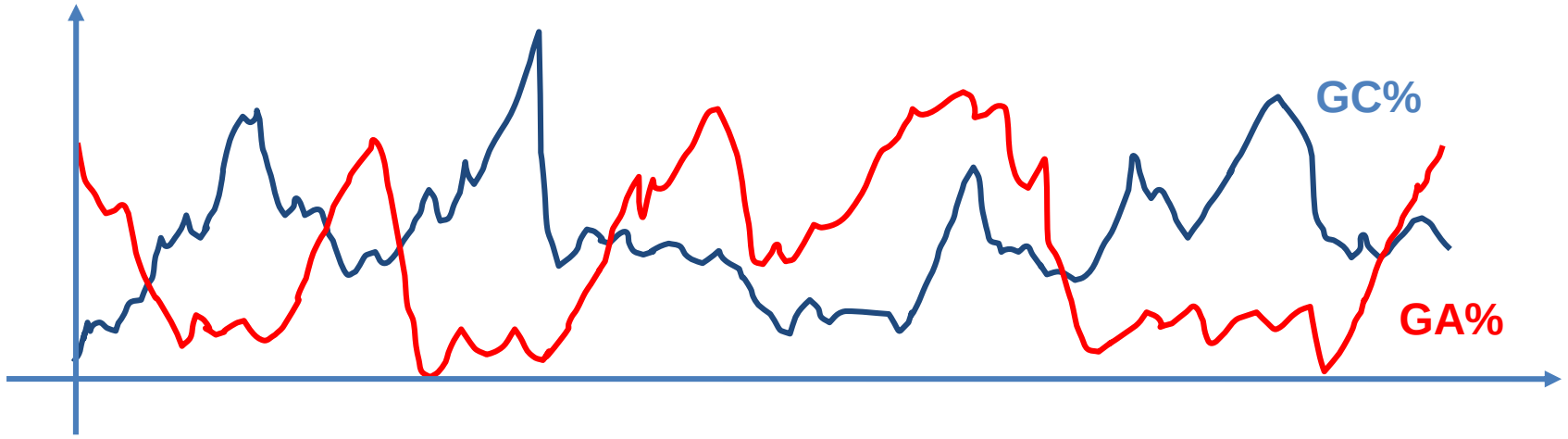
Сравнение спектров фрагментов ДНК-профилей



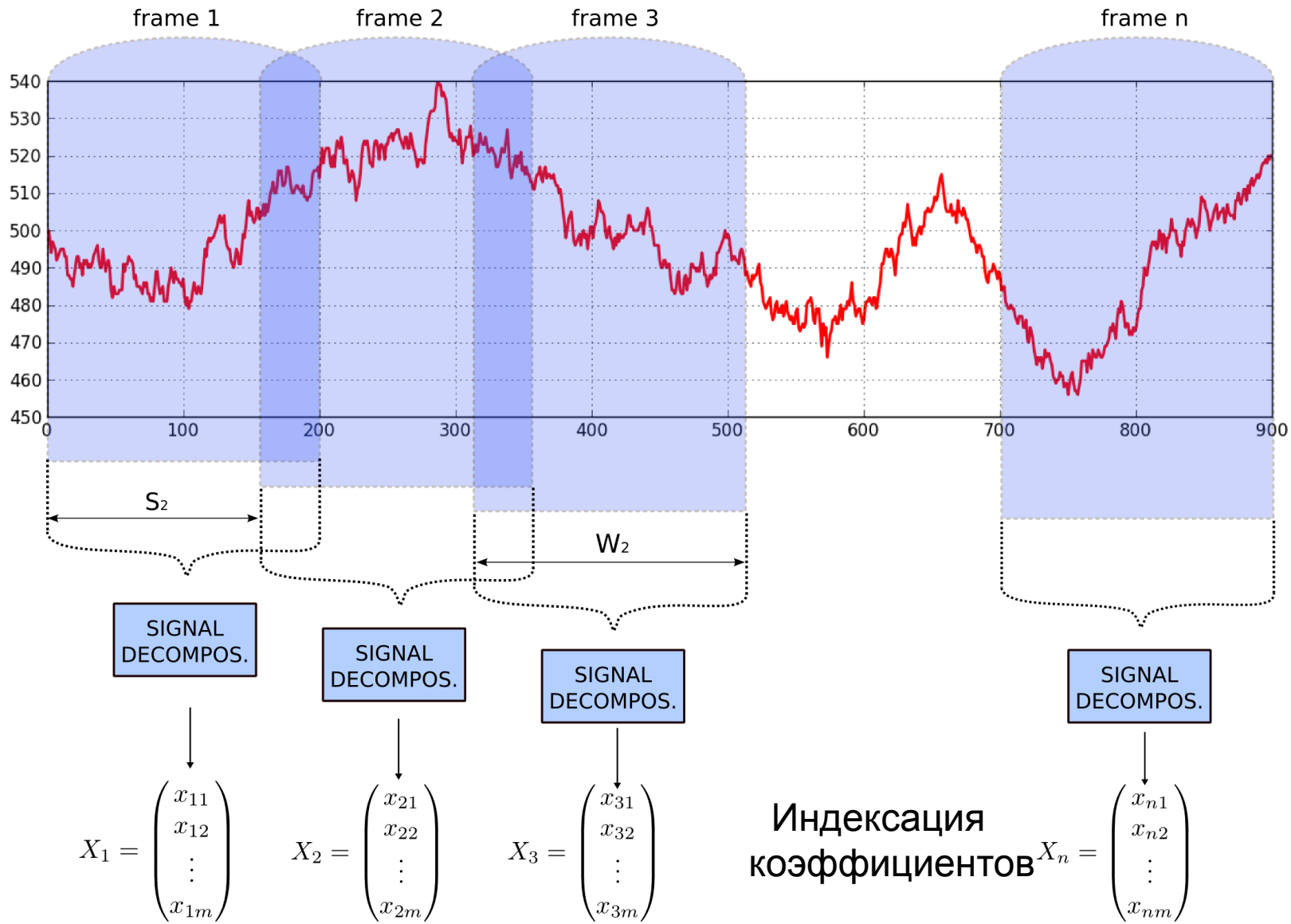
Отображение результатов на точечной матрице и их анализ

Преобразование нуклеотидной последовательности в функциональное представление

Теорема: для однозначного восстановления символьной последовательности достаточно $\lceil \log_2 N \rceil$ функций-аналогов, где N – мощность алфавита



Преобразование функций-аналогов в спектры коэффициентов разложения



Сравнение спектров

$$\theta(\{A_n\},\{B_n\}) < \varepsilon$$

$$\rho^2(f,g) = (f-g, f-g) = \sum_{k=0}^{L-1} (A_k - B_k)^2 (\varphi_k, \varphi_k)$$

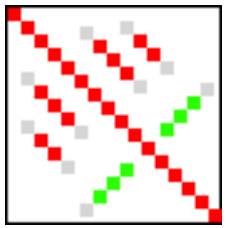
$$(f-g, f-g) = \int_{-\pi}^{\pi} (f-g)^2 dt \le 2\pi W_1^2$$

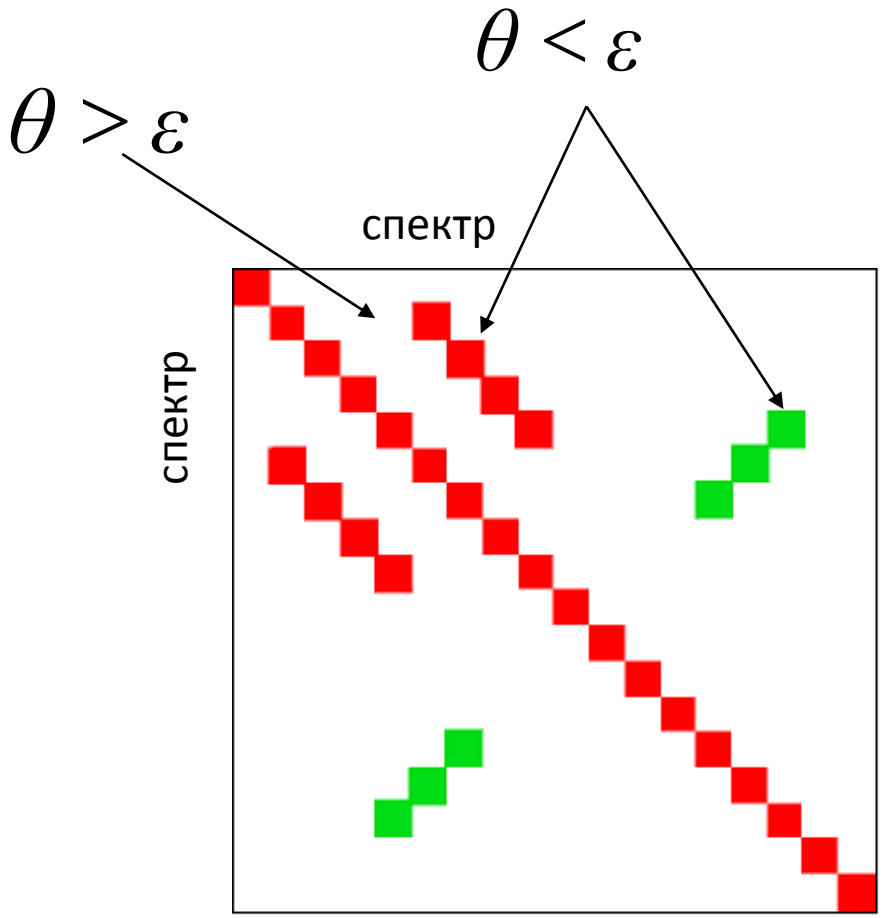
$$\theta(f,g) = \frac{1}{2W_1^2} \sum_{k=0}^{L-1} (A_k - B_k)^2 \le \varepsilon \le 1$$

Метрика:

1) монотонна по числу коэффициентов,
2) нормирована
3) инвариантна к масштабу

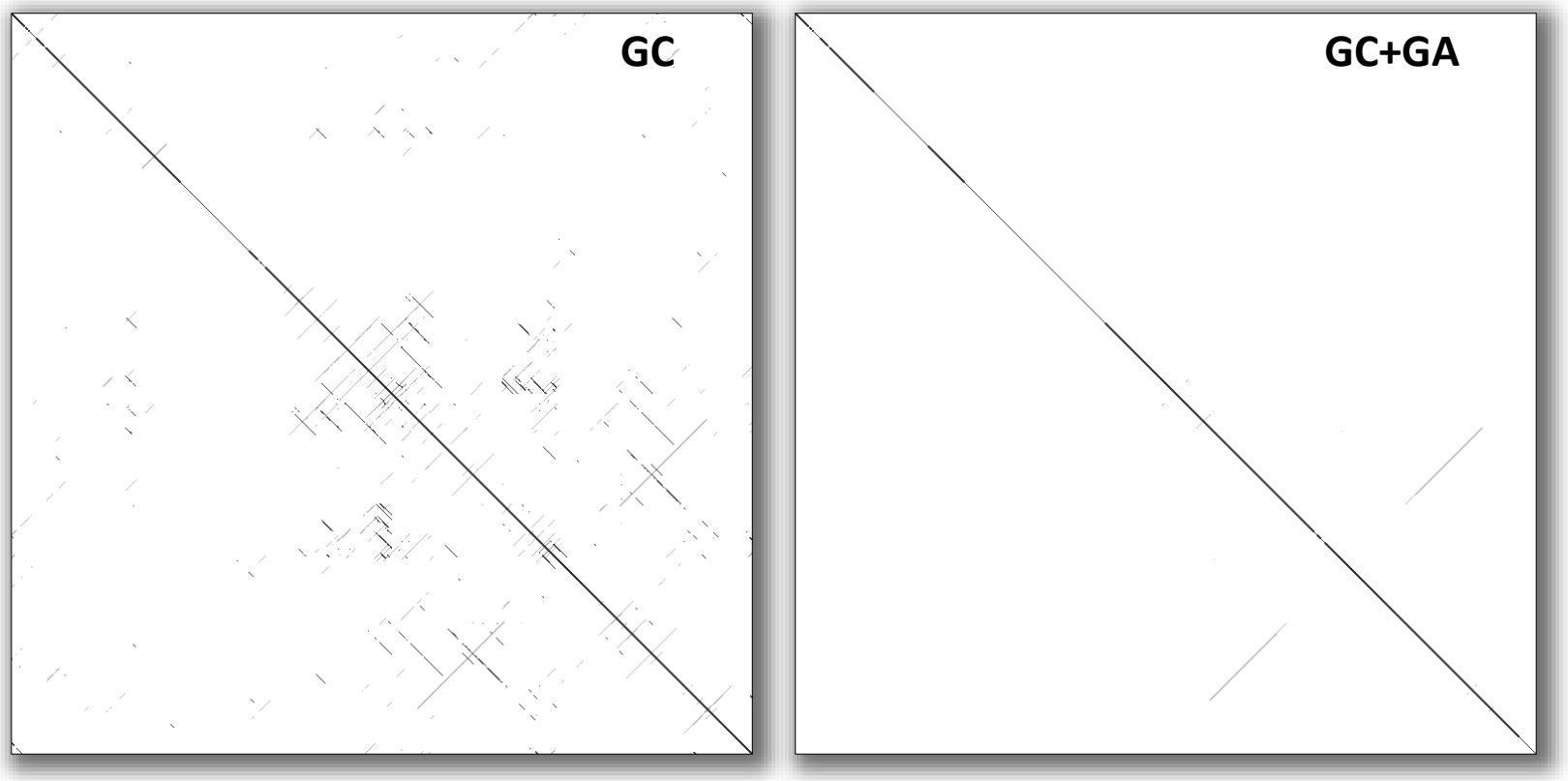
Отображение результатов на матрице спектральной схожести





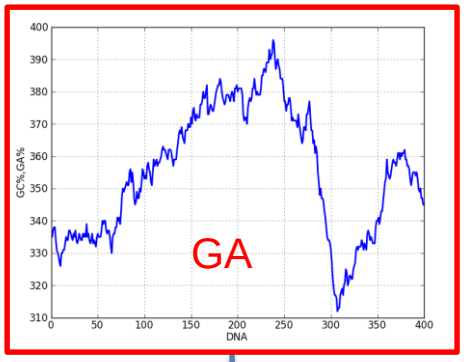
Автоматический поиск заданных образцов

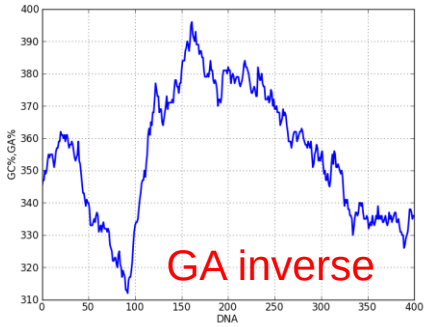
Фильтрация



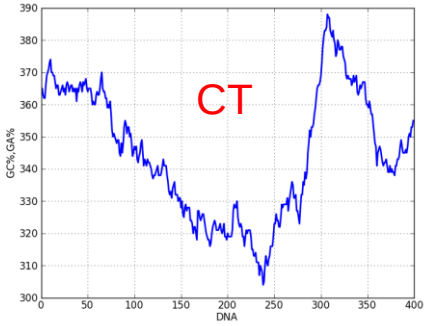
$$(\theta(f^{GC}, g^{GC}) \leq \varepsilon) \wedge (\theta(f^{GA}, g^{GA}) \leq \varepsilon)$$

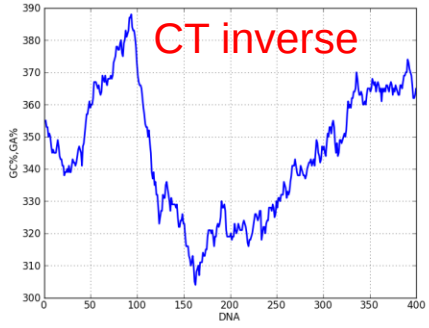
$$\begin{cases} C_{2k}^{GA} = C_{2k}^{GArev} \\ C_{2k+1}^{GA} = -C_{2k+1}^{GArev} \end{cases}$$





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





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

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

Сложность алгоритма:

Матрица фиксированного размера строится за <u>линейное</u> время относительно длины последовательности ***n***

$An + Bn + Cn^2$ без пропорционального масштабирования параметров W_2 и S_2

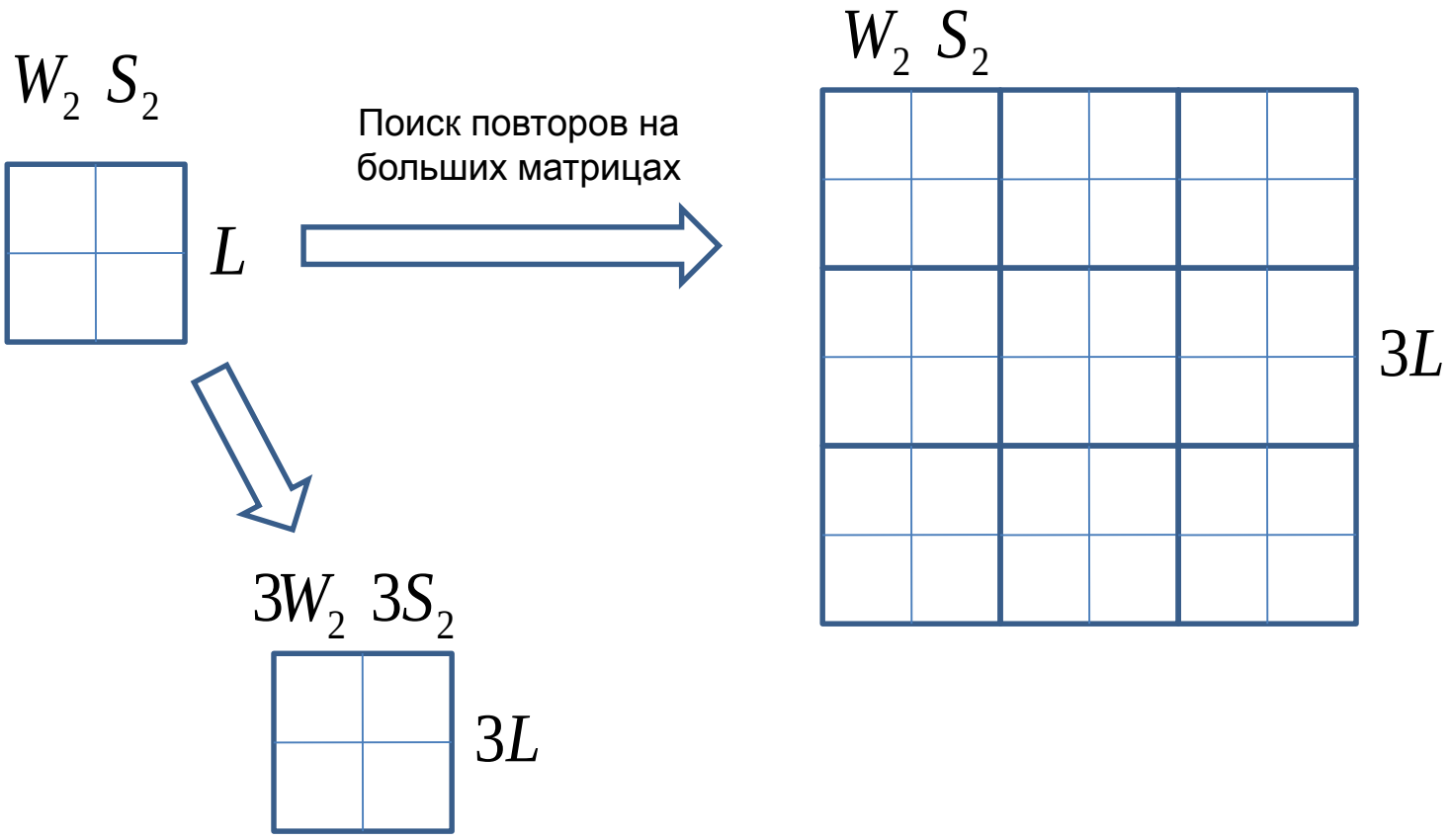
$An + Bn + c$ с масштабированием

An - вычисление кривых GC,GA содержания

Bn - вычисление векторов коэффициентов разложения

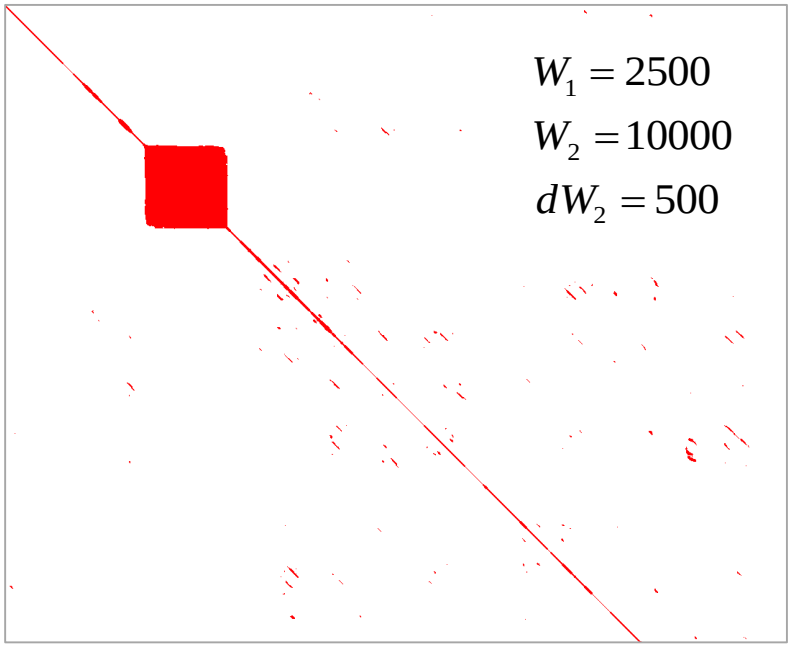
Cn - сравнение векторов коэффициентов

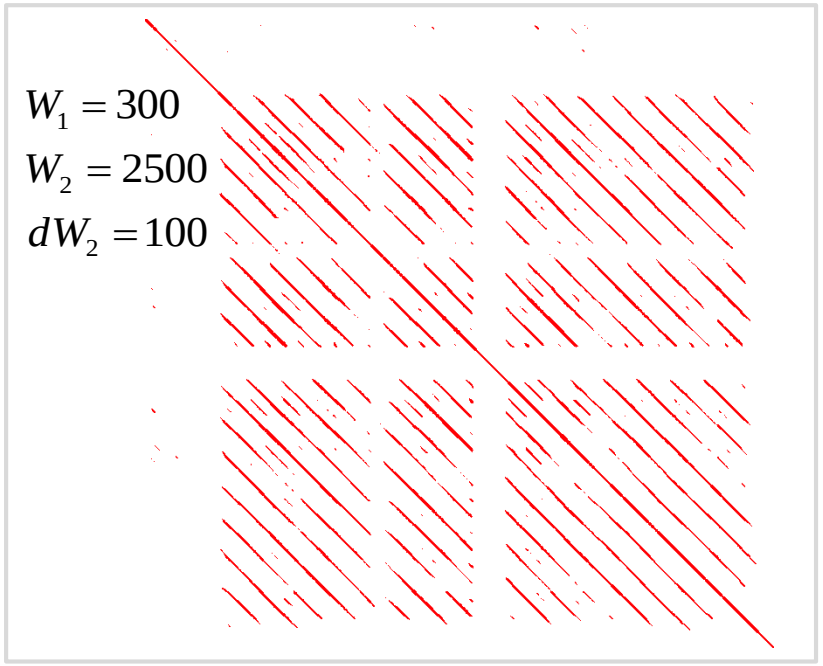
Оптимизация спектрального метода

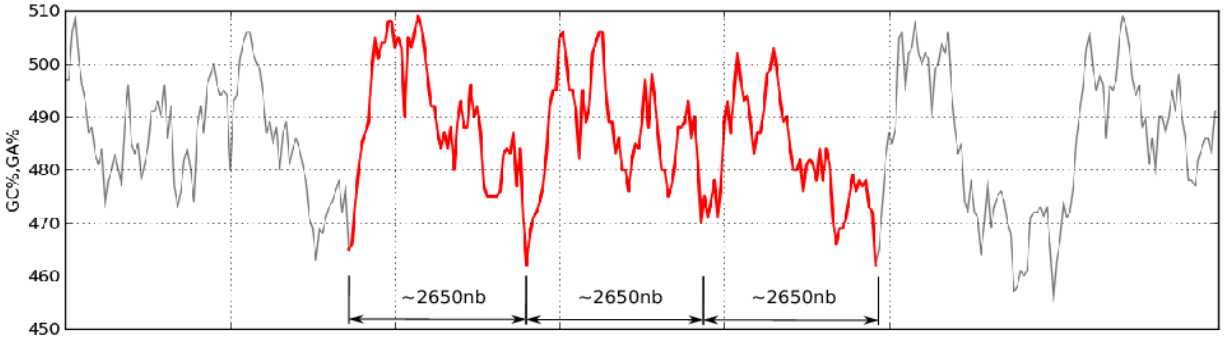


Integrated Performance Primitives (Intel IPP) – векторизация
Open Multi-Processing (OpenMP) – многопоточность

Тандемный повтор (MSU1)

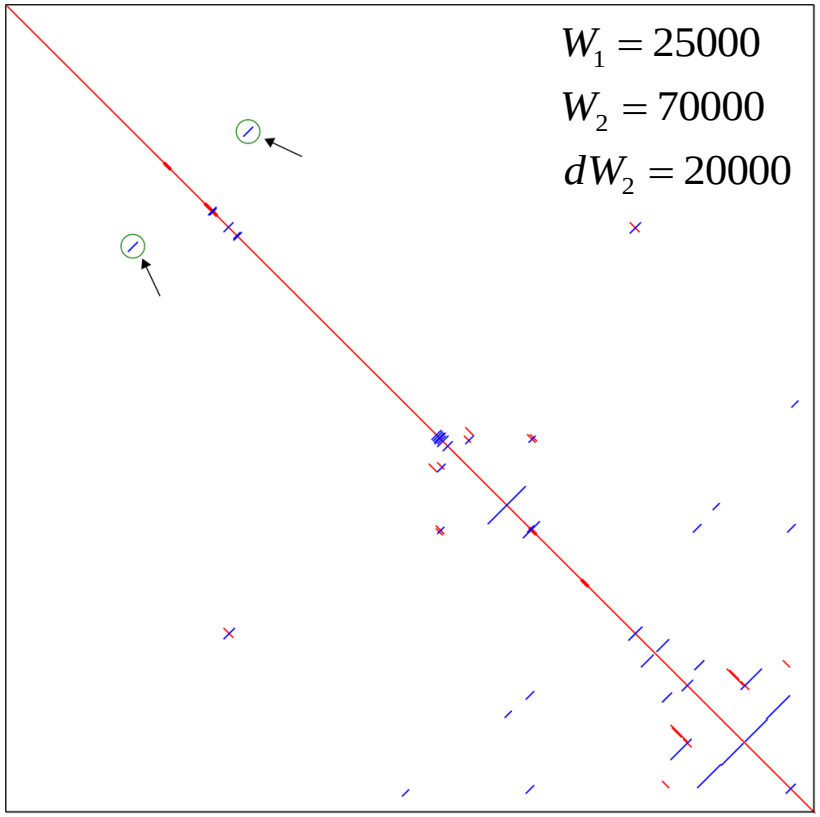


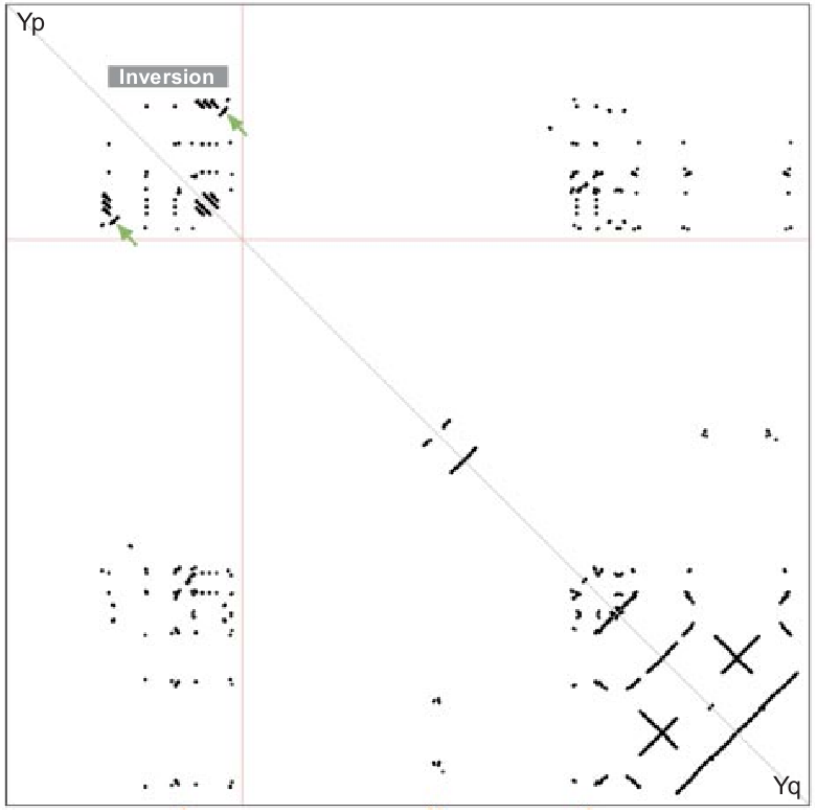




Pyatkov M.I., Filippov V.V., Pankratov A.N. Consensus of repeated region of rabbit chromosome 17 containing over 15 huge approximate tandem repeats. Repbase Reports. 2012. Vol.12, No.3.

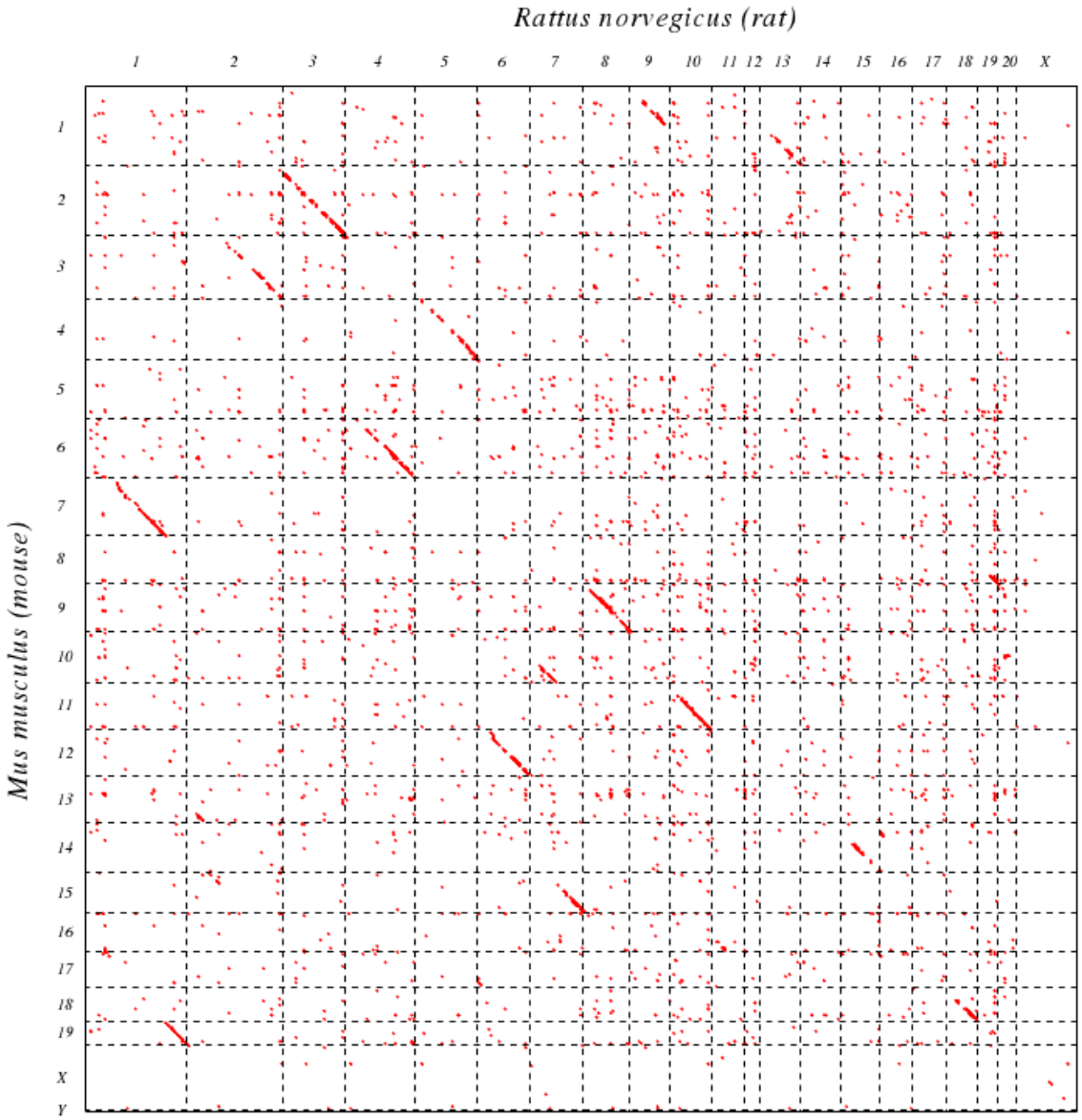
Y хромосома человека



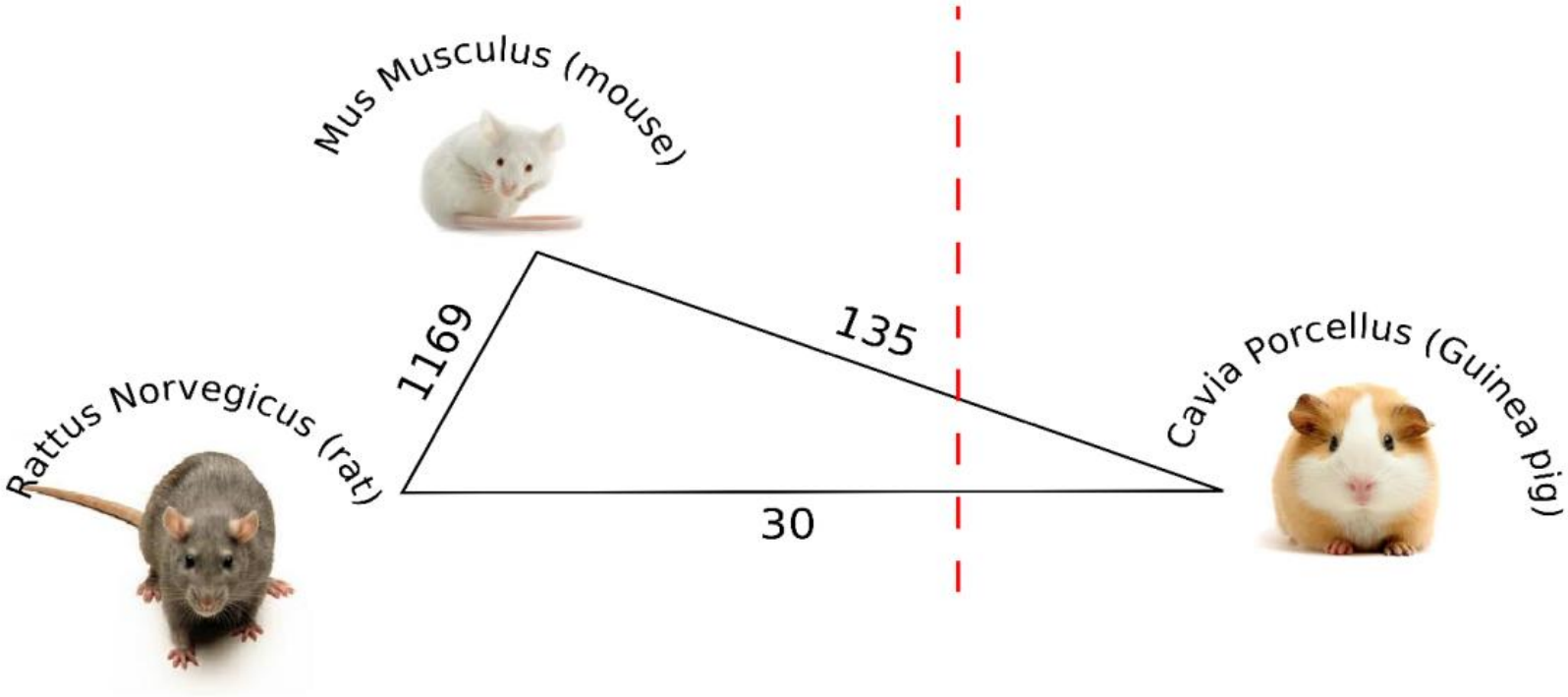


Tilford CA et al, A physical map of the human Y chromosome. Nature. 2001. No. 409, P.943--945

Карта сравнения геномов мыши и крысы

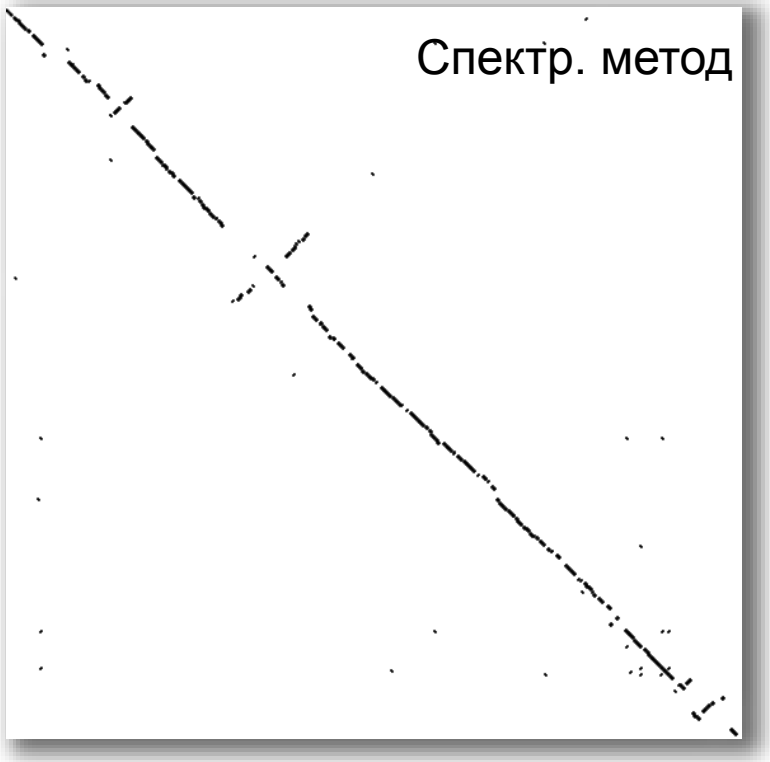


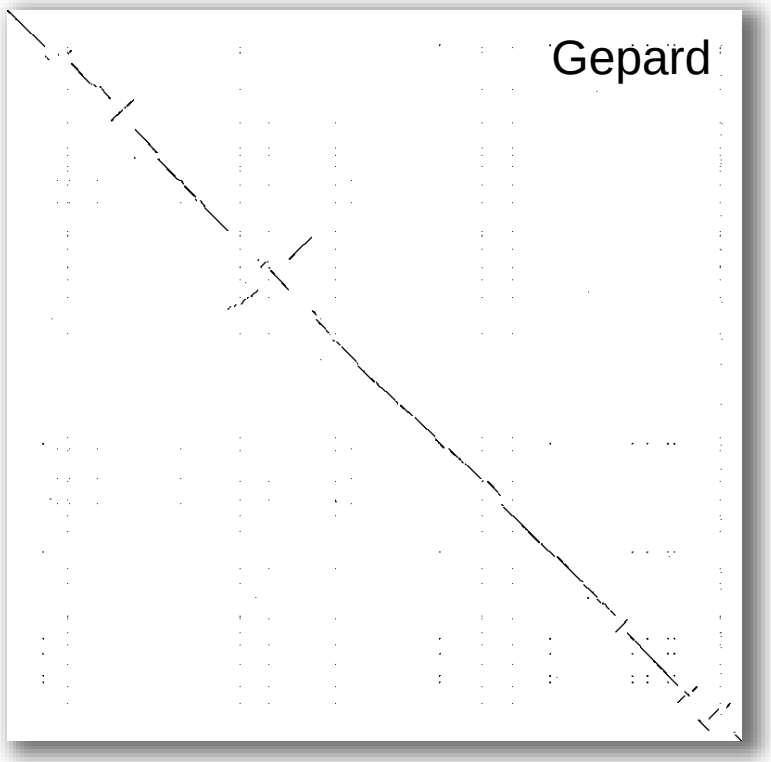
Сравнение геномов



Escherihia coli K12

Shigella flexnery 2a



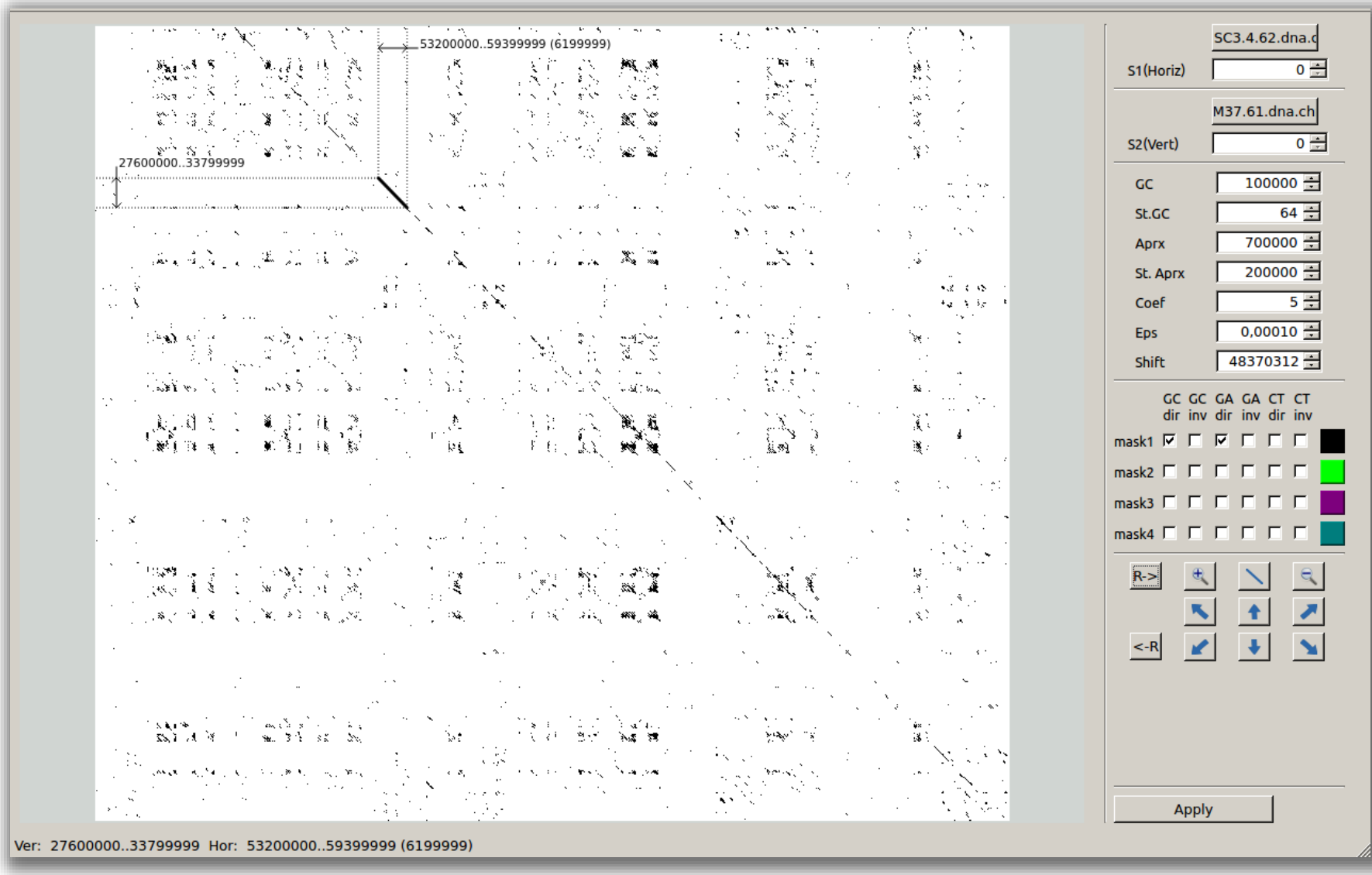


Длина ~ 5.000.000 н.п.

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.

Длина последовательности	Gepard	Спектр. метод
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

S.B.A.R.S (Spectral Based Approach for Repeats Search)

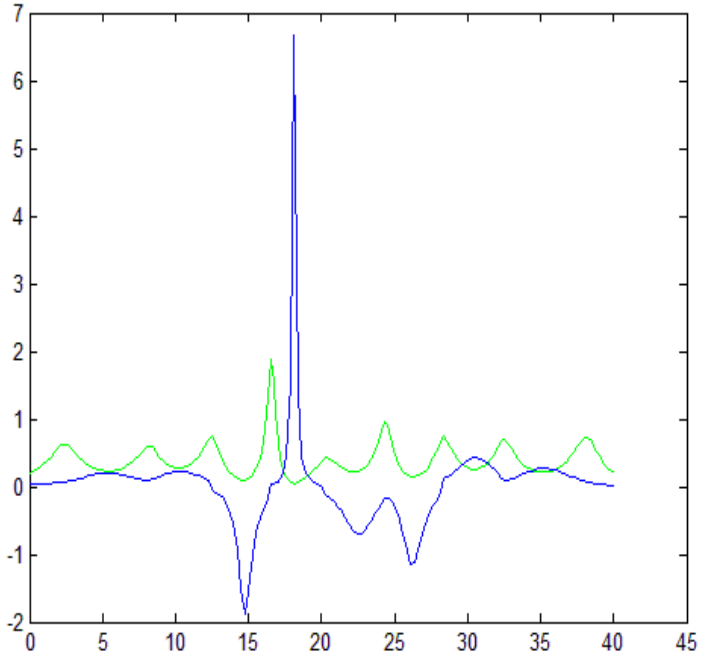


Pyatkov M. I., Pankratov A. N. (2014) SBARS: fast creation of dotplots for DNA sequences on different scales using GA,GC-content // Bioinformatics, Vol. 30, No. 14, pp. 1765 - 1766

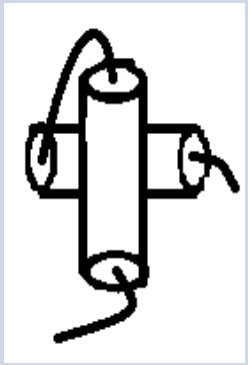
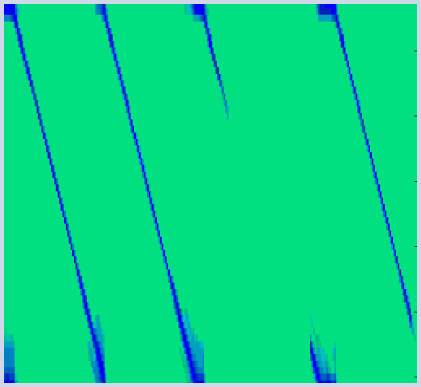
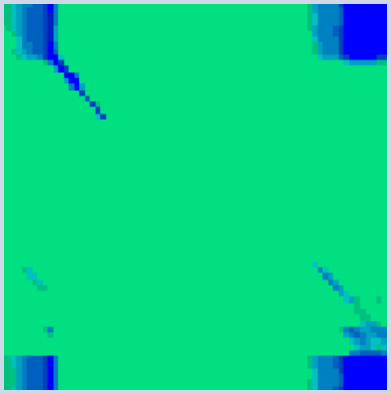
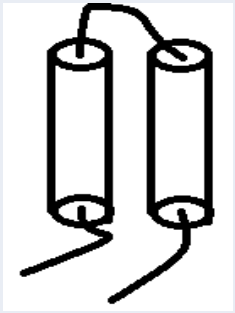
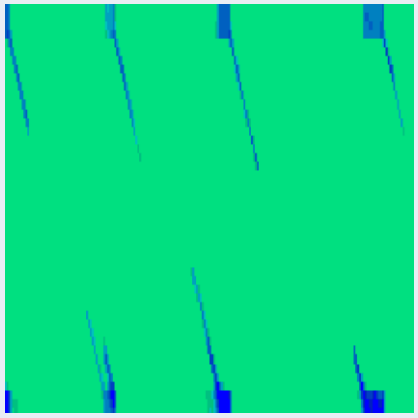
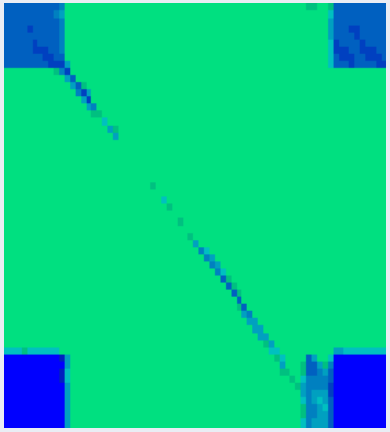
Реализация алгоритма поиска паттерна а-а уголка в белках

Паттерн – 1D1L.pdb{15 - 37}

Получение кривых



Сравнение пространственных структур белков

	1R69 (3 уголка)	1ROP (1 шпилька)
1D1L 		
256B 		

Выводы

Разработан новый метод распознавания близких по структуре элементов геномов и пространственных структур белков

1. Распознавание повторов на разных масштабах
2. Выравнивание последовательностей или структур
3. Быстрое построение точечной матрицы

Благодарю за внимание!