РАЗРАБОТКА МЕТОДОВ ДЛЯ ТОЧНОГО ПРЕДСКАЗАНИЯ ПОЛОЖЕНИЯ ПРОМОТЕРОВ НА ГЕНОМЕ

Код проекта: TT

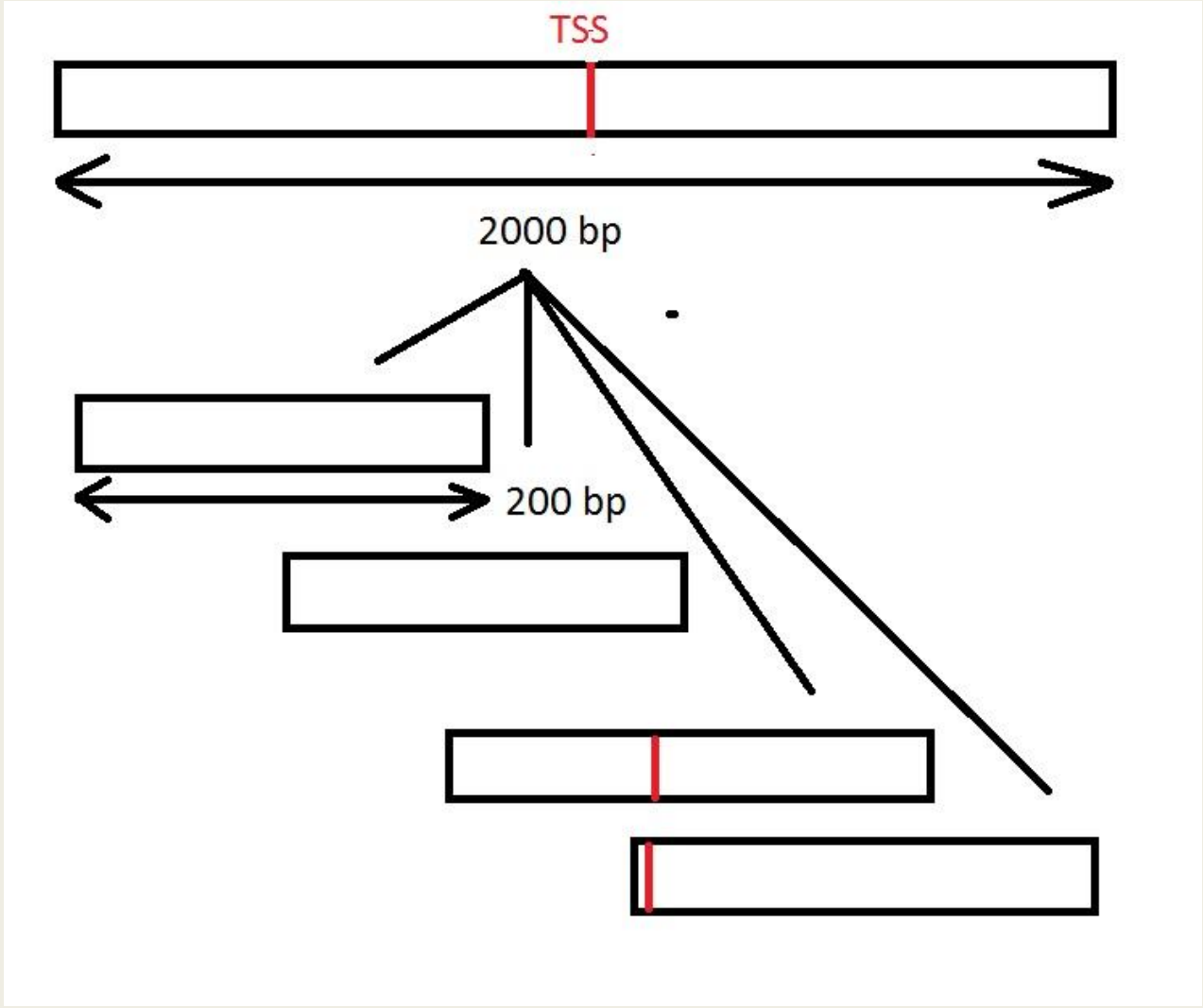
Команда №2

Доступные входные данные

- Участки последовательностей ДНК длиной 2000 нуклеотидов
- TSS находится ровно в середине каждого участка
- Разнообразные признаки для каждого нуклеотида из последовательности:
  - *Наличие метилирования*
  - *Наличие SNP*
  - *Наличие участков CA, CG*
  - *И др.*

Основная идея

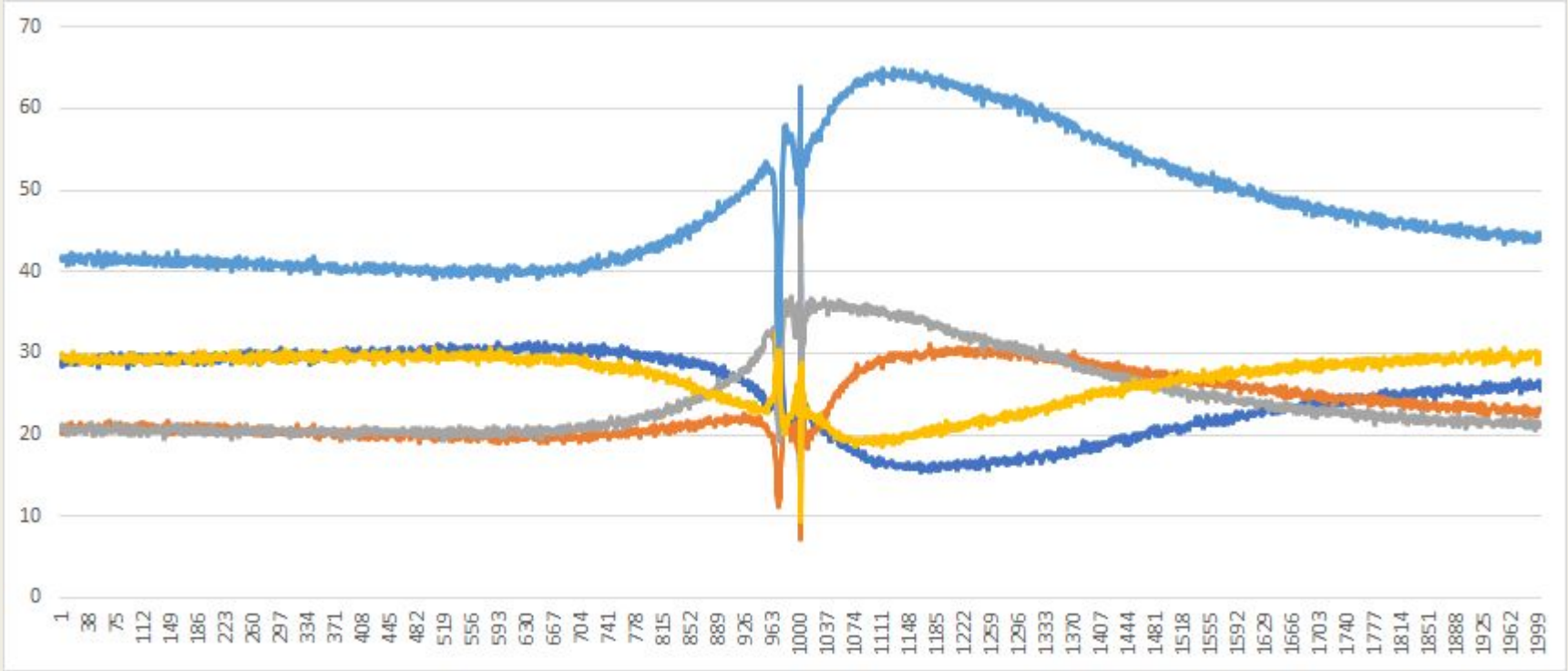
- Разбить последовательность на более мелкие последовательности по 200 пар нуклеотидов.
- Выбрать из получившейся последовательности только ту, в которой вероятность наличия TSS максимальна.
- Для выбранной последовательности предсказать точное положения точки начал транскрипции.



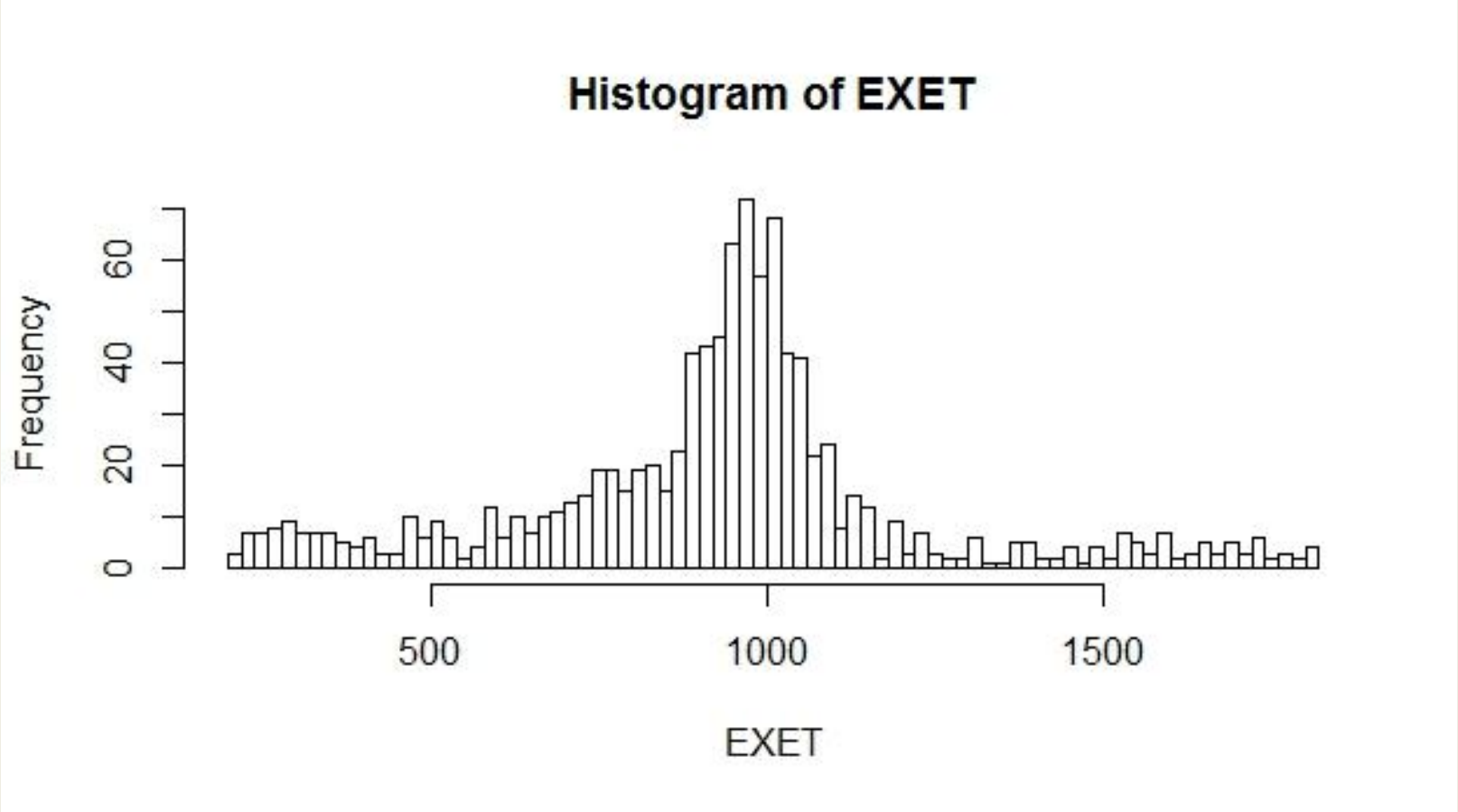
Предсказание положения TSS для подпоследовательности

- Random Forest. (Java, Weka library)
- Признаки:
  - *Закодированная последовательность нуклеотидов*
- Средняя ошибка:
  - *50 bp для последовательности длиной 200 bp*

Классификация подпоследовательностей: алгоритм



- Ближе к точке начала транскрипции доля GC увеличивается.



- Рассчитываем коэффициент корреляции между GC и позицией нуклеотида.
- Удаляем нуклеотиды с низким коэффициентом корреляции.
- Ищем области с наибольшим числом оставшихся нуклеотидов.
- Выбираем области с максимальным значением.

СПАСИБО ЗА ВНИМАНИЕ!