Улучшение в SPAN модели поиска пиков с помощью модели линейной регрессии

Елена Картышева,
Институт биоинформатики

Научный руководитель: Алексей Диевский, JetBrains

ChIP-seq

ChIP-seq — метод анализа ДНК-белковых взаимодействий, для изучения модификаций гистонов по всему геному и поиска мест связывания транскрипционных факторов.

- Результаты ChIP-seq: трек — набор меток вида (хромосома, позиция).
- В местах связывания меток будет существенно больше.

Проблема: ChIP-seq — неточный метод: метки появляются и там, где связывания нет, то есть данные шумные.

Что такое SPAN?

SPAN (Semi-supervised Peak ANalyzer) — инструмент для поиска областей повышенного сигнала в геномных данных ChIP-seq.

В его основе лежит скрытая марковская модель с тремя состояниями:

- нулевое (отсутствие сигнала)
- два негативно-биномиальных (различные параметры для шумов и пиков)

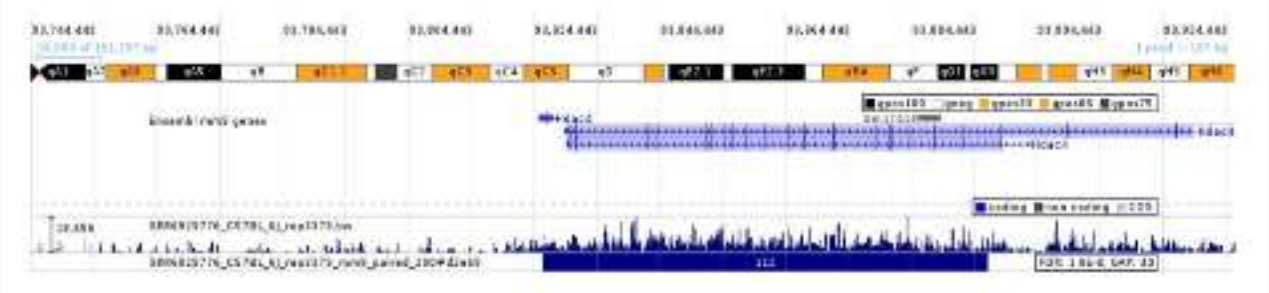


Рис. 1: Пример трека в SPAN

Задача

Улучшить SPAN, а именно добавить возможность дообучать модель, используя не только треки, но и какую-то информацию о данных, например:

- GC-content
- Mappability
- Local BG Estimate

Для этого решено использовать обобщенные линейные модели (GLM).

Generalized linear model

Generalized linear model

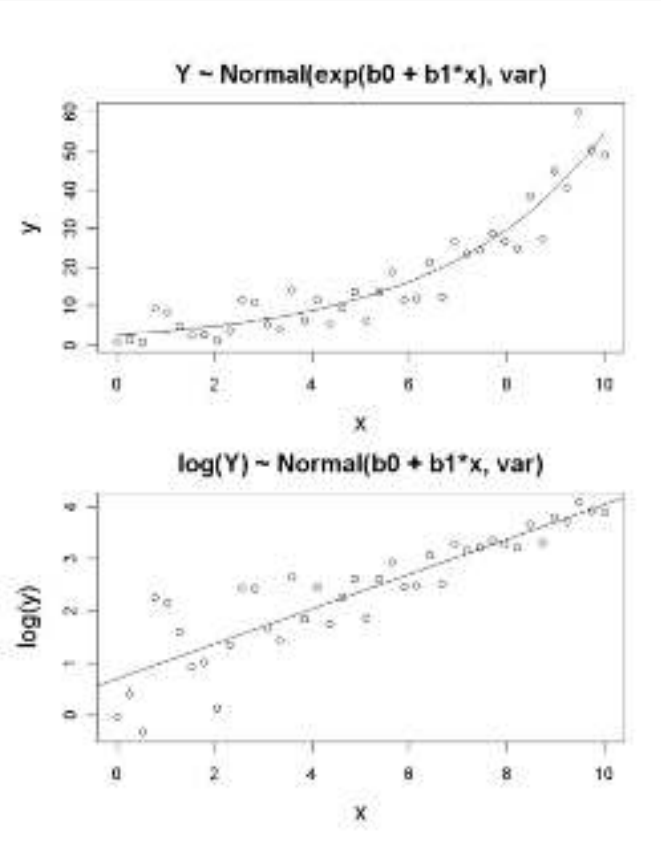
Линейная регрессия здорового человека:

$$b \sim N(Ax, \sigma^2),$$

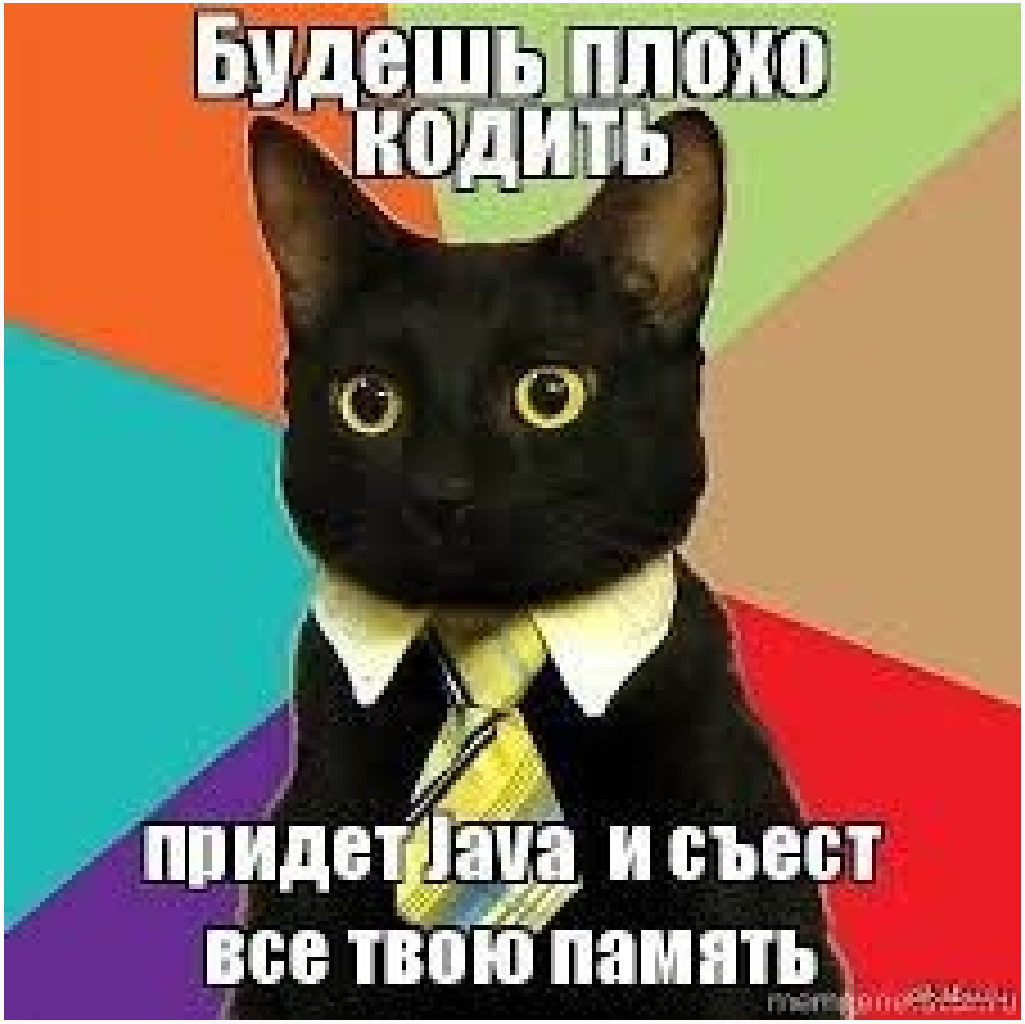
A — матрица данных, b — вектор наблюдений.

Линейная регрессия куриль... статистика:

$$b \sim NB(exp(Ax), f).$$



Хьюстон, кажется, у нас проблемы

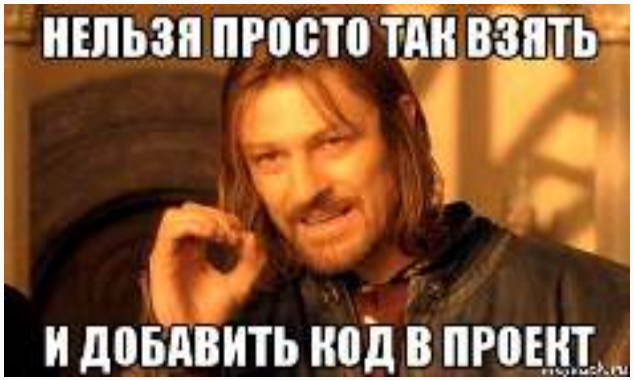


Что новенького?

Библиотека для биоинформатики на Kotlin — bioinf_commons

Добавлено:

1. Абстрактный класс регрессии
2. Класс пуассоновской регрессии
3. Класс смеси из 3х компонент: нулевая, две пуассоновские.



Что сделано (альтернативное название: тут у меня закончилась фантазия)

Что сделано:

- Написана эффективная версия GLM на Kotlin/Java.
- Имплементированы классы в bioinf_commons.
- Под это дело написан pull request к классу GLMRegression в Apache Commons, позволяющий эффективно применять взвешенную регрессию.

Спасибо за внимание!