Моделирование различий в данных ChIP-seq

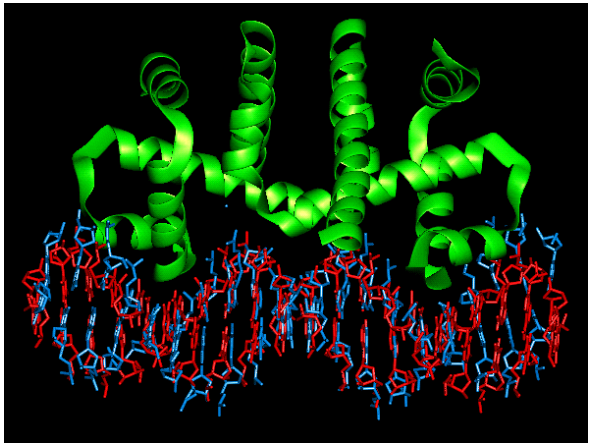
магистрант
Алексей Диевский, СПбАУ
руководитель
Олег Шпынов, JetBrains

СПбАУ НОЦНТ РАН
27.05.2013

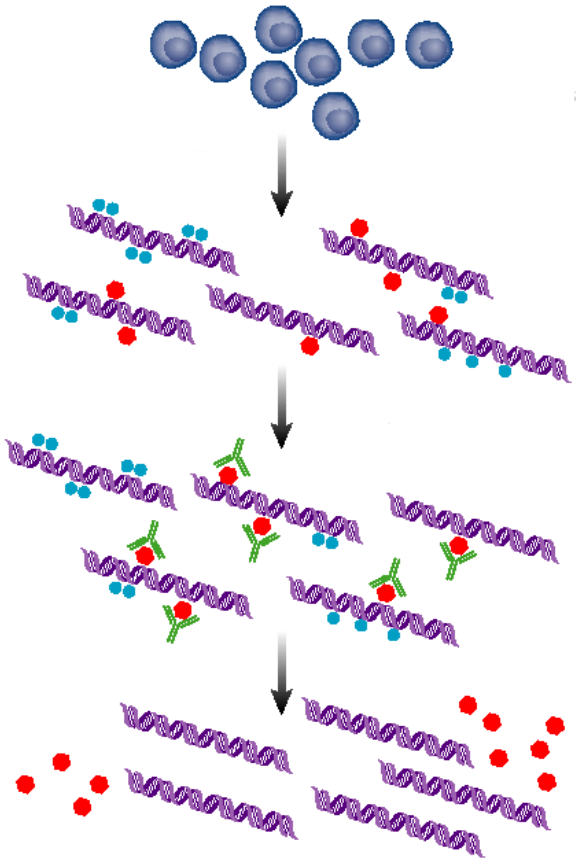
Мотивация

- Взаимодействие ДНК и белков играет важную роль.
- Нужно определить места связывания.

- Примеры взаимодействия:
  - полимеразы;
  - транскрипционные факторы;
  - активаторы и репрессоры;
  - модификации хроматина.



ChIP-seq



дробление

отбор

выравнивание

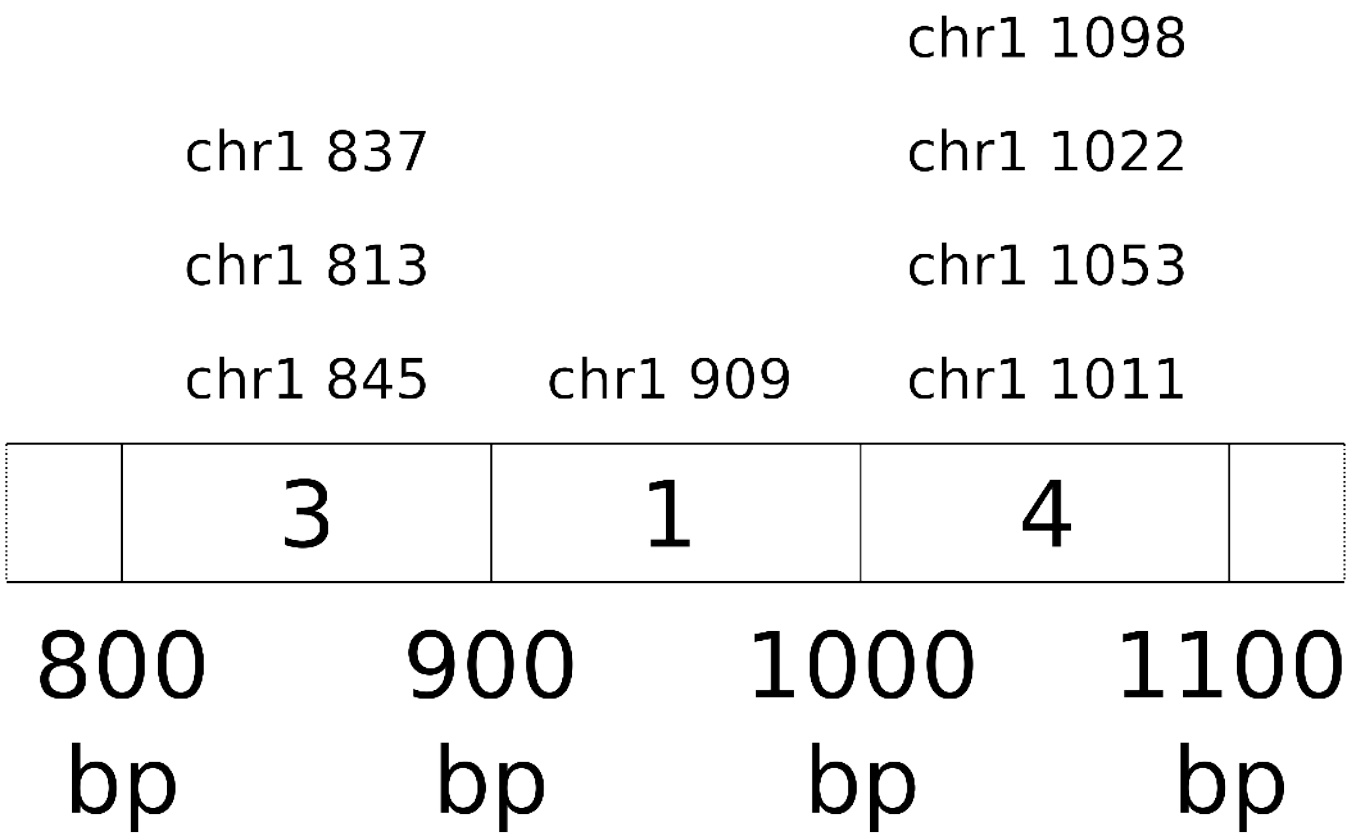
Результаты ChIP-seq

- Трек – набор меток вида: *(хромосома, позиция)*
- В местах связывания меток будет существенно больше.
- ChIP-seq – неточный метод: метки появляются и там, где связывания нет.

Обработка данных

- Неперекрывающиеся окна фиксированного размера.
- Суммарное количество меток, попавших в каждое окно – *покрытие:*
  - высокое покрытие – есть связывание;
  - низкое покрытие – техническая погрешность.
- Трек становится последовательностью натуральных чисел.

Обработка данных



Цель и задачи

- Цель – построить математическую модель для сравнения двух экспериментов ChIP-seq:
  - правдоподобную,
  - биологически интерпретируемую,
  - поддающуюся обсчёту.
- Задачи:
  - изучить аналоги;
  - построить несколько моделей и выбрать лучшую;
  - сравнить с аналогами.

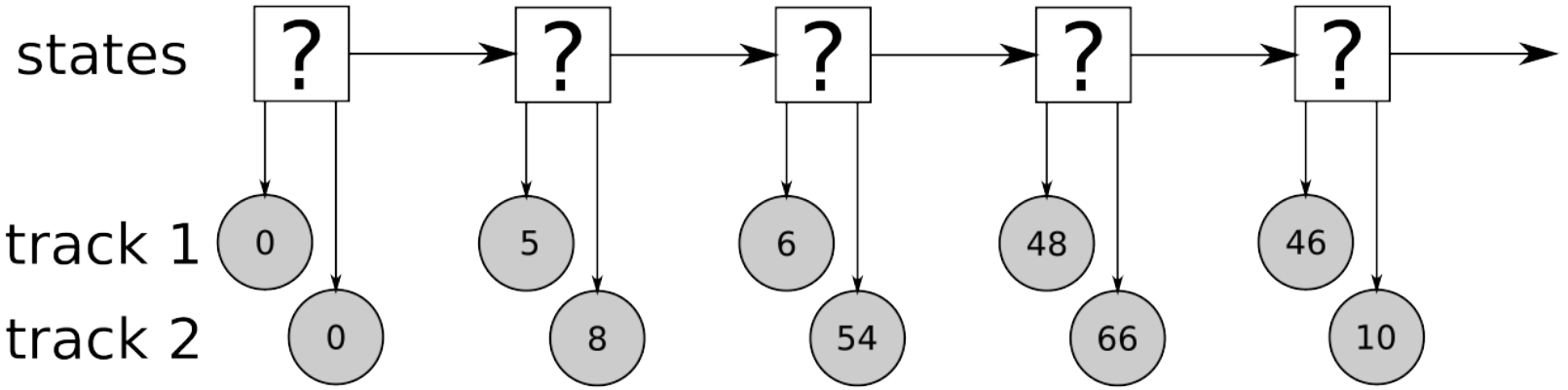
Аналог

- Алгоритм *ChIPDiff:*
  - критерий – кратное увеличение покрытия.
- Не подходит: нас интересует качественное изменение покрытия.

Наивная идея

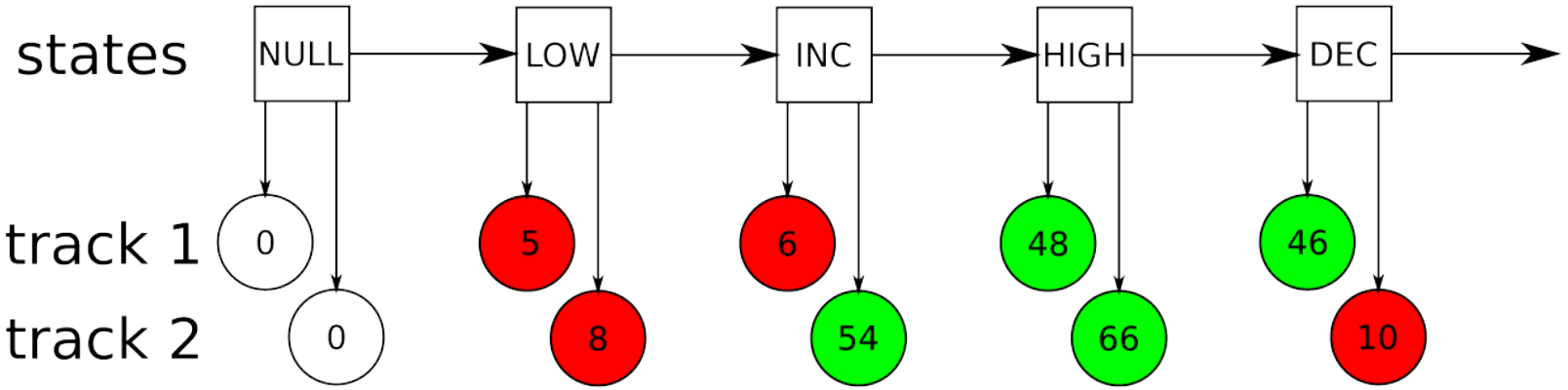
- Найти места связывания независимо для каждого трека и определить различия.
- Алгоритмы для исследования одного трека:
  - *HPeak*, *HHMM*, *BayesPeak...*
- Не подходит: места связывания на двух треках, скорее всего, зависимы.

Скрытая марковская модель



- Скрытые состояния – марковский процесс.
- Наблюдения – покрытие окна на двух треках:
  - подчиняются распределению Пуассона;
  - интенсивность зависит от скрытого состояния.

Скрытая марковская модель



- Пять возможных скрытых состояний.
- Обучение модели:
  - вероятности перехода;
  - интенсивности наблюдений.

Пример

- Данные:
  - белок – модифицированный гистон *H3K4me3;*
  - клетки – две клеточные линии *M. musculus;*
  - хромосома – *chr1.*
- При ширине окна в 500 нуклеотидов:
  - 1003 области в состоянии INCREASED;
  - 323 области в состоянии DECREASED;
  - время обучения около 10 минут.

Результаты

- Модель лучше, чем другие построенные нами модели, например:
  - две независимые модели для каждого трека;
  - модель без состояния NULL;
  - модель без зависимости близких состояний.
- Модель подходит для исследования данных ChIP-seq, поддаётся обсчёту, биологически интерпретируема.

Сравнение: правдоподобие

размер окна	полная модель	независимые треки	без NULL
50	**-3867261**	-5012256	-3874820
100	-3220008	-3822349	**-3121767**
500	**-1621935**	-1937743	-1762844
1 000	**-1226367**	-1442737	-1352232
5 000	**-675293**	-759971	-726937
10 000	**-521599**	-578464	-550646
50 000	**-288290**	-326742	-307852
100 000	**-242900**	-275045	-264421
500 000	**-181984**	-203096	-195168
1 000 000	**-163032**	-180921	-172443
5 000 000	-91068	-91247	**-86401**

Сравнение: области

размер окна	ChIPDiff	наша модель	пересечение
50	76108	10632	5076
100	7496	7490	3226
500	1479	2580	812
1 000	692	1452	418
5 000	134	307	98

- *ChIPDiff* использует другой критерий, но пересечение существенное.
- Наши результаты более стабильны.

Спасибо за внимание!

- Контактный адрес:
  - dievsky@gmail.com
- Благодарности:
  - JetBrains BioLabs