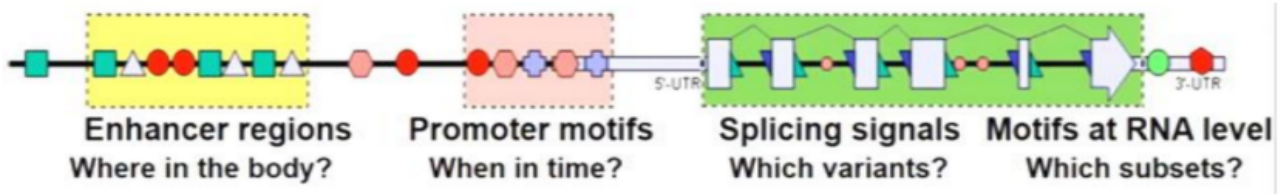
Поиск регуляторных мотивов

- набор генов фиксированный (~20-30 тыс, протеин-код, тРНК, микроРНК)
- при этом должно выполняться большое разнообразие биологических функций
- выполняются с помощью контроля: энхансеры, промоторы, сплайсинг

Регуляторный код: комбинаторный подход для кодирования меток

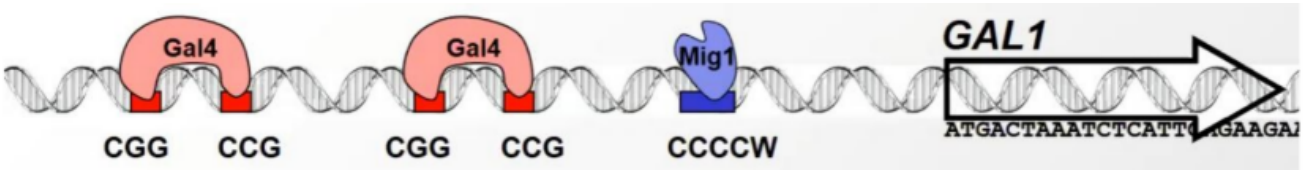


Регуляторные мотивы

- регулирование генов на изменения в окружающей среде
- нет прямого управления: гены содержат мотивы
- транскрипционные факторы (ТФ) распознают эти метки

Сложности обнаружения мотивов

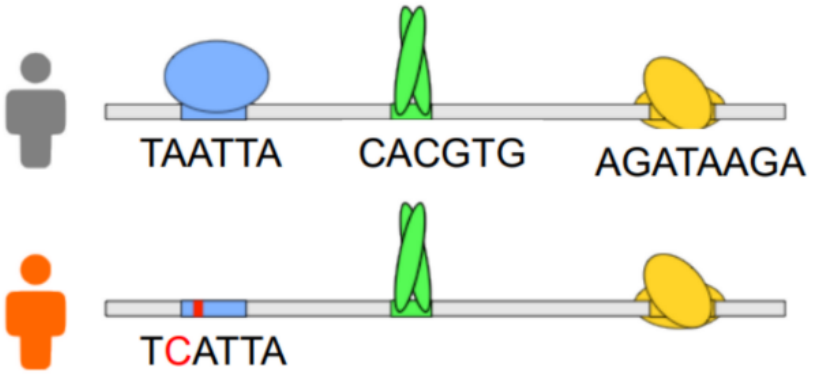
- короткие (6-8 bp), часто неоднозначные
- содержат любые последовательности нуклеотидов (нет старт/стоп кодонов)
- работают на разном расстоянии до или после целевого гена



Транскрипционные факторы и мотивы

Транскрипционные факторы (ТФ) – белки, которые используют ДНК-связывающие домены для распознавания участков двойной цепи ДНК Влияют на транскрипцию, что в итоге приводит к изменению количества РНК и в случае белок-кодирующих генов – белка

Последовательность мотива зависит от структуры ТФ



93% генетических вариантов полигенных особенностей находятся в некодирующих регионах

Распознавание мотивов по структуре

Белки распознают ДНК: определение химических свойств оснований, нет расплетания цепи ДНК (не смотрим на комплементарность оснований)

3D структура определяет специфичность: четко определенные позиции (каждый атом важен), неоднозначные позиции — слабые связи). Бывают и другие типы распознавания: микроРНК (комплементарность), нуклеосомы (GC-состав)

Мотив — некая последовательность в алфавите нуклеотидов, которая обобщает знание о последовательности отдельных генов или ТФ, регулирующих группу генов

Мотив → мера информации (энтропия) → объединяют много позиции

Важно: отличать мотив от экземляра мотива (мотив — усредненное)

Предположение о мотивах: независимость позиций и фиксированное расстояние

Задачи регуляторной геномики

Как изучать?

ТФ: обнаружение гомологии

Мотив: сравнение *de novo*

Экземпляр: функциональный участок (эволюция, хроматиновые регионы)

Методы поиска регуляторных мотивов

На основе участков последовательностей:

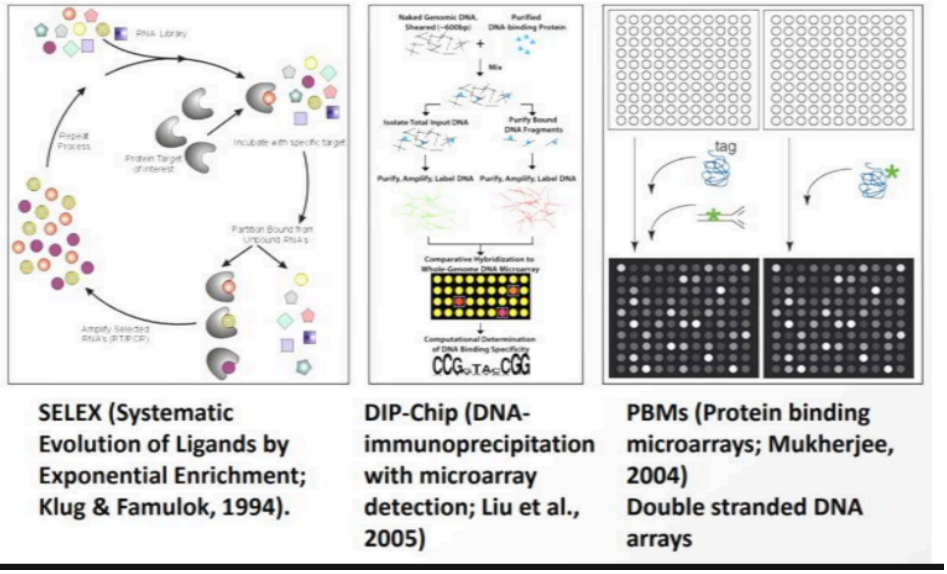
- Expectation-Maximization
- Gibbs sampling

На основе полных геномов:

- Эволюционный филогенетический анализ, оценка всего генома

Эксперименты in vitro:

- Идентификация мотивов на основе протеиновых доменов



SELEX (Systematic Evolution of Ligands by Exponential Enrichment; Klug & Famulok, 1994).

DIP-Chip (DNA-immunoprecipitation with microarray detection; Liu et al., 2005)

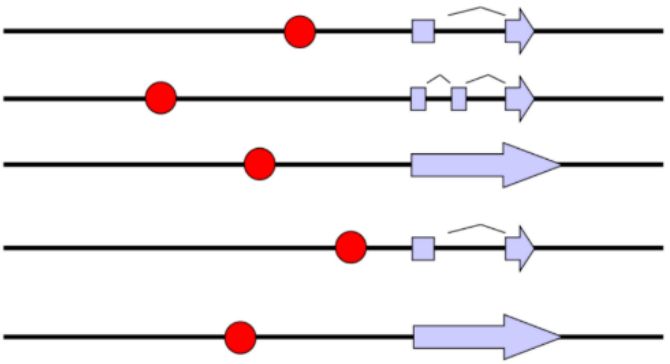
PBMs (Protein binding microarrays; Mukherjee, 2004)
Double stranded DNA arrays

Есть ТФ. *Задача*: определить, к какой последовательности прикрепляется ТФ
Библиотеки РНК: к каждой последовательности пытаемся прикрепить белок
Убираем белок, амплификация и секвенирование

Методы поиска на основе обогащения

По набору **совместно регулируемых** или **функционально связанных генов** найти общие мотивы в их промоторных регионах

локальное попарное выравнивание промоутеров
экспертный поиск участков, похожих на мотивы
поиск «усредненной» последовательности
начинать с консервативных участков (до гена)



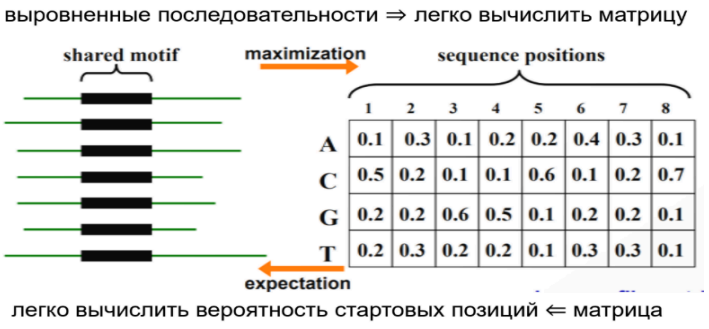
Стартовые позиции и матрица мотивов

На вход: некоторые последовательности, в которых мы хотим поискать общую часть

Идея: итеративно оценивать и мотив, и стартовые позиции

Мы предполагаем, что знаем длину мотива. И есть матрица с позициями (seq pos), где мы имеем на каждой позиции вероятность встретить каждый из нуклеотидов

А где находятся стартовые позиции в каждой из зеленых последовательностей?



На вход: длина мотива, набор последовательности
do
заполнить матрицу мотивов начальными значениями
оценить стартовые позиции по мотиву оценить мотив по стартовым позициям
until сходимость (изменения < eps), критерий: матрица мотивов перестала изменяться
Выход: мотив, стартовые позиции

Способ записи мотива

фиксированная длина W
матрица вероятностей мотива p(c, k) c – нуклеотид {A, C, G, T} k – позиция в мотиве от 1 до W

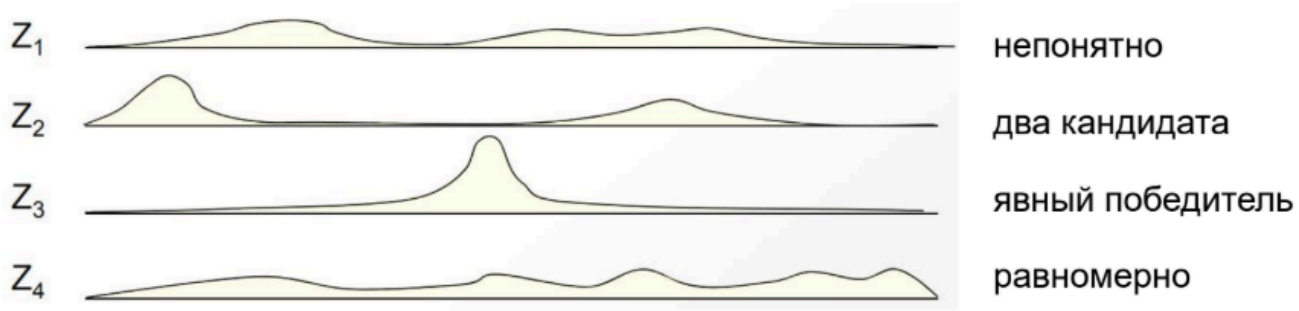
Базовые вероятности нуклеотидов в виде вектора p(c, 0)

- равномерное распределение
- можно учитывать особенности генома

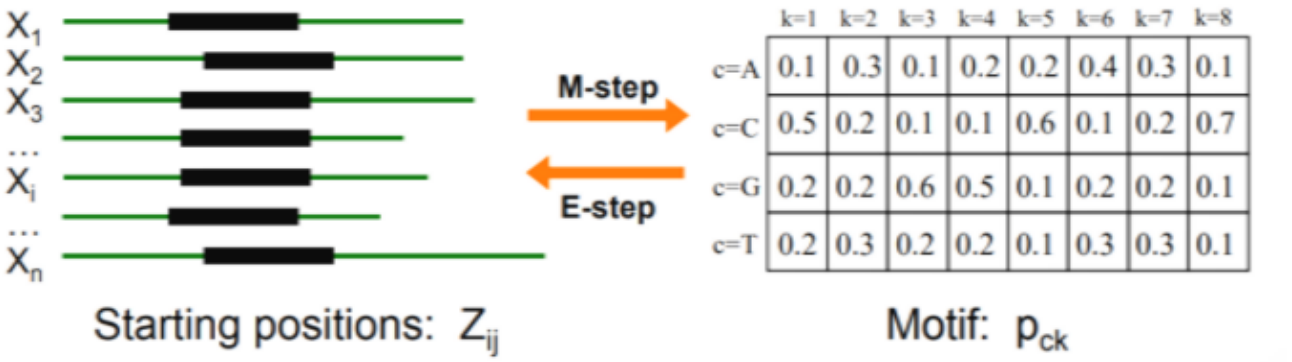
Способ записи стартовой позиции

Z(i, j) – вероятность, что в последовательности i мотив начинается с позиции j

$$Z = \begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ seq1 & 0.1 & 0.1 & 0.2 & 0.6 \\ seq2 & 0.4 & 0.2 & 0.1 & 0.3 \\ seq3 & 0.3 & 0.1 & 0.5 & 0.1 \\ seq4 & 0.1 & 0.5 & 0.1 & 0.3 \end{array}$$



по матрице старта пересчитываем матрицу мотивов, по матрице мотивов пересчитываем матрицу старта



E-step: для каждой позиции оцениваем вероятность старта перемножая значения из матрицы

M-step: EM – усредняем все возможные значения с весами Z(i, j) Gibbs – для каждой последовательности Xi выбираем один старт из распределения Z(i, j) Greedy – для каждой последовательности Xi выбираем лучшее значение из Z(i, j)

Вероятность последовательности Xi

Вероятность увидеть последовательность Xi при старте в позиции j

$$\Pr(X_i \mid Z_{ij} = 1, p) = \underbrace{\prod_{k=1}^{j-1} p_{c_k,0}}_{\text{before motif}} \underbrace{\prod_{k=j}^{j+W-1} p_{c_k,k-j+1}}_{\text{motif}} \underbrace{\prod_{k=j+W}^{L} p_{c_k,0}}_{\text{after motif}}$$

Вероятность увидеть последовательность Xi при старте в позиции j

$$\Pr(X_i \mid Z_{ij} = 1, p) = \underbrace{\prod_{k=1}^{j-1} p_{c_k,0}}_{\text{before motif}} \underbrace{\prod_{k=j}^{j+W-1} p_{c_k,k-j+1}}_{\text{motif}} \underbrace{\prod_{k=j+W}^{L} p_{c_k,0}}_{\text{after motif}}$$

Пример:

$X_i =$ G C **T G T** A G $\quad p =$

	0	1	2	3
A	0.25	0.1	0.5	0.2
C	0.25	0.4	0.2	**0.1**
G	0.25	0.3	**0.1**	0.6
T	0.25	**0.2**	0.2	0.1

$$\Pr(X_i \mid Z_{i3} = 1, p) =$$

$$p_{G,0} \times p_{C,0} \times p_{T,1} \times p_{G,2} \times p_{T,3} \times p_{A,0} \times p_{G,0} =$$

$$0.25 \times 0.25 \times \boxed{0.2 \times 0.1 \times 0.1} \times 0.25 \times 0.25$$

Стартовые позиции (Формула Байеса)

$$\Pr(Z_{ij}=1 \mid X_i, p) = \frac{\Pr(X_i \mid Z_{ij}=1, p)\, P(Z_{ij}=1)}{\Pr(X_i)}$$

Вычисляем в момент времени t используя текущую матрицу p

- общая вероятность равна сумме по всем стартовым позициям
- мотив равновероятно начинается в любой позиции

$$Z_{ij}^{(t)} = \frac{\Pr(X_i \mid Z_{ij} = 1, p^{(t)})\, \Pr(Z_{ij} = 1)}{\sum_{k=1}^{L-W+1} \Pr(X_i \mid Z_{ik} = 1, p^{(t)})\, \Pr(Z_{ik} = 1)}$$

$X_i =$ G C T G T A G

$p =$

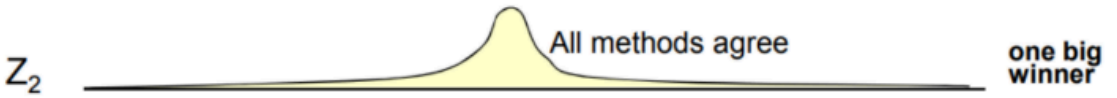
	0	1	2	3
A	0.25	0.1	0.5	0.2
C	0.25	0.4	0.2	0.1
G	0.25	0.3	0.1	0.6
T	0.25	0.2	0.2	0.1

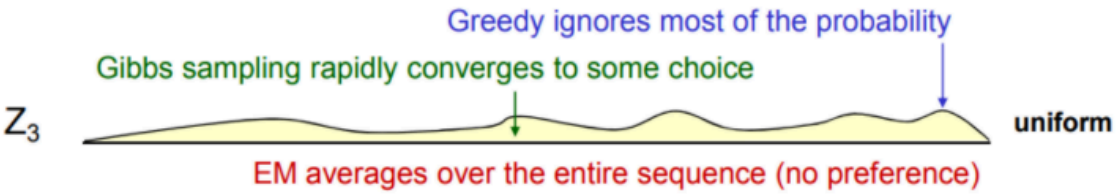
$$Z_{i1} = 0.3 \times 0.2 \times 0.1 \times 0.25 \times 0.25 \times 0.25 \times 0.25$$

$$Z_{i2} = 0.25 \times 0.4 \times 0.2 \times 0.6 \times 0.25 \times 0.25 \times 0.25$$

$\vdots$

- then normalize so that $\sum_{j=1}^{L-W+1} Z_{ij} = 1$





Как вычислить матрицу p?

$$p_{c,k}^{(t+1)} = \frac{n_{c,k} + d_{c,k}}{\sum_b (n_{b,k} + d_{b,k})}$$

pseudo-counts

$$n_{c,k} = \begin{cases} \sum_i \sum_{\{j|X_{i,j+k-1}=c\}} Z_{ij} & k > 0 \quad \text{motif} \\ n_c - \sum_{j=1}^{W} n_{c,j} & k = 0 \quad \text{background} \end{cases}$$

total # of c's in data set

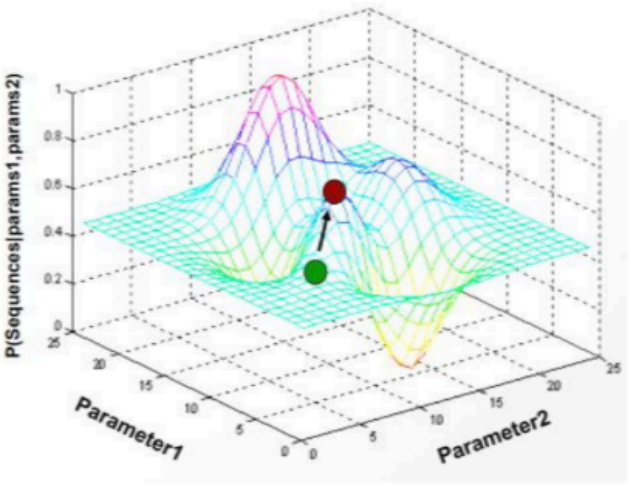
Сходимость EM алгоритма

Сходится к локальному максимуму увеличивая P(seqs|parameters)

Useful to think of P(seqs|parameters) as a function of parameters

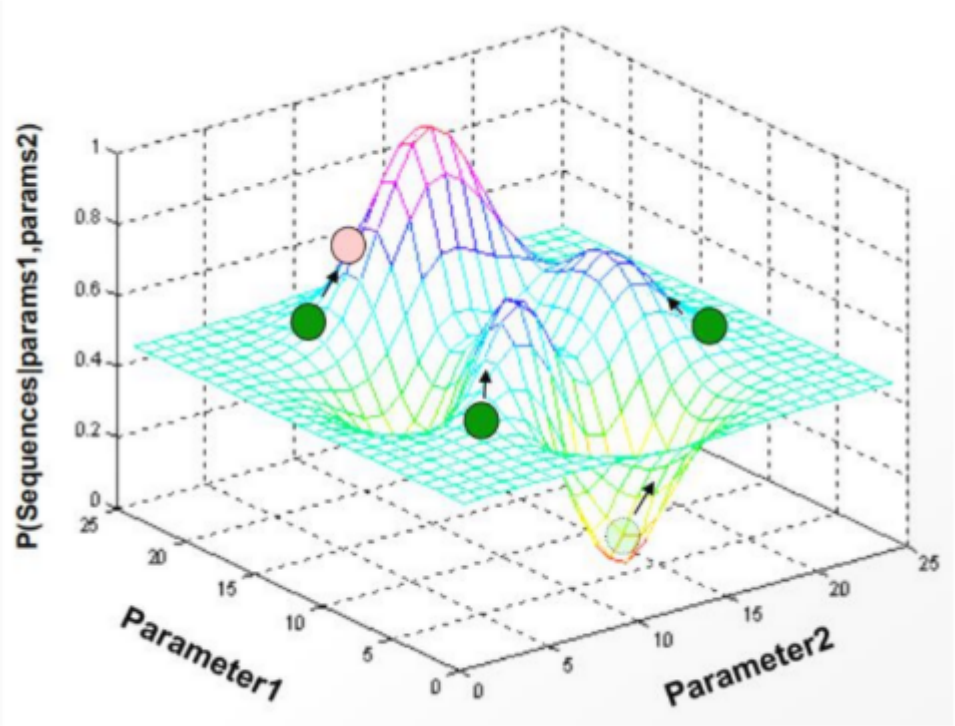
EM starts at an initial set of parameters

And then "climbs uphill" until it reaches a local maximum



Выбор стартовых значений критически важен

Оптимизация: стартовать из разных позиций



Gibbs sampling

Стохастический аналог алгоритма EM для поиска мотивов Менее подвержен застреванию в локальных минимумах

EM поддерживает распределение вероятности стартовых значений для каждой последовательности

Gibbs выбирает стартовое значение a_i для каждой последовательности, но итеративно обновляет значения стартовых позиций

Сравнение с EM

Очень похож на EM алгоритм

Преимущества: проще в реализации, меньше зависит от выбора начальных параметров, более универсален, возможны различные эвристики

Недостатки: больше зависит от совокупности последовательностей, менее систематическое исследование пространства поиска