PSSM и паттерны

А.С. Ершова
С.А. Спирин

Вспомним PWM, вес и информационное содержание

TTATGCC
ATCTTCA
GTATTAA

выравнивание

	1	2	3	4	5	6	7
G	0.26	-1.3	-1.3	-1.3	0.26	-1.3	-1.3
A	0.26	-1.3	0.74	-1.3	-1.3	0.26	0.74
T	0.26	1.18	-1.3	1.18	0.74	-1.3	-1.3
C	-1.3	-1.3	0.26	-1.3	-1.3	0.74	0.26

PWM для данного выравнивания

$$I_k = \sum_i f_{ki} \log_2 \frac{f_{ki}}{p_i}$$

$$I = \sum_k I_k$$

Информационное содержание (I) позволяет понять, как много похожих на мотив последовательностей мы найдем в наших данных по случайным причинам.

Элементы PWM: S_{ki} для основания i в позиции k,

p_i — фоновая частота основания i

f_{ki} — частота основания i в позиции k
(с учётом псевдоотсчётов)

λ — любое число (для удобства)

$$S_{ki} = \frac{1}{\lambda} \log \frac{f_{ki}}{p_i}$$

Применение PWM

Приложив позиционную весовую матрицу (PWM) к последовательности той же длины, можно понять, содержит ли последовательность сигнал, описываемый этой PWM.

Пусть есть последовательность b_1, b_2, …, b_L, и PWM (S_{ki}),

здесь k — позиция, $k = 1, \ldots, L$, а i — буква, например, i из {A, T, G, C}.

Тогда можно посчитать **вес** (score) последовательности относительно PWM:

$$S = \sum_k S_{kb_k}$$

Чем выше вес, тем более вероятно, что последовательность содержит сигнал.

Ещё можно искать вероятные вхождения мотива в длинную последовательность (например, геном), считая вес всех возможных отрезков нужной длины: где вес выше порога, там предсказывается мотив. Выбор порога — отдельная задача.

PSSM — position-specific scoring matrix

- По смыслу PSSM — это то же, что PWM, но термин PWM используется для мотивов в ДНК, а PSSM для мотивов в белках или для описания семейств родственных белков.
- Можно использовать гэпы, учитываются как 21 буква.
- PSSM применяется так же, как PWM: если вес последовательности белка относительно PSSM выше порога, предсказывается принадлежность белка семейству.

PSSM — position-specific scoring matrix

- Как создать PSSM по выравниванию?
- Базовая идея — та же, что для PWM:

$$S_{ki} = \frac{1}{\lambda} \log \frac{f_{ki}}{p_i}$$

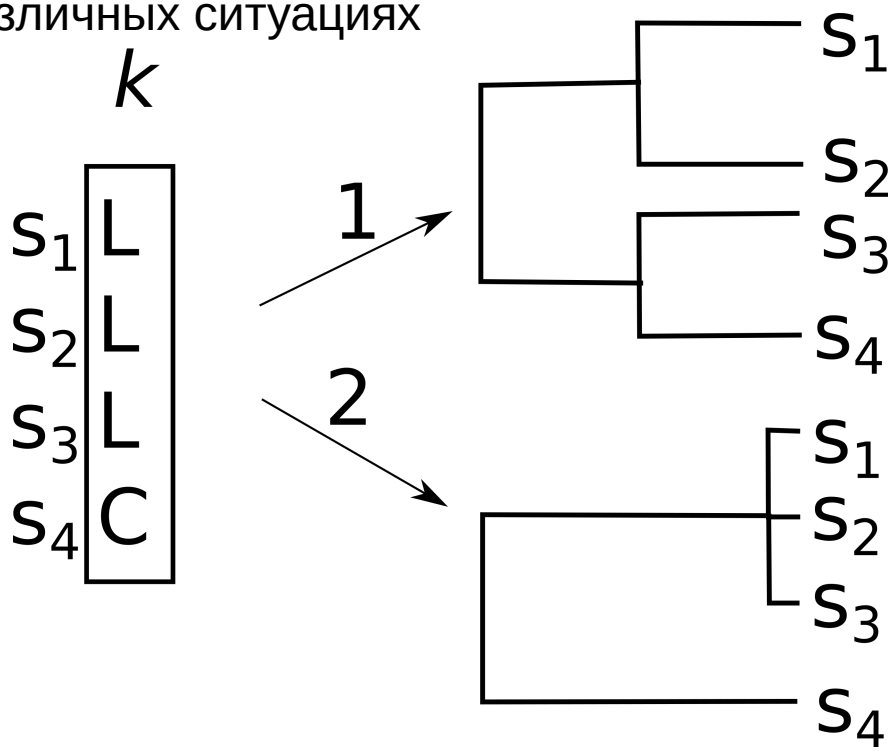
где S_{ki} — элемент позиционной весовой матрицы
(вес буквы i в позиции k),
p_i — фоновая частота остатка i
f_{ki} — частота остатка i в позиции k
(с учётом псевдоотсчётов)

Как оценить f_{ki}?

- Очевидная идея: $f_{ki} = n_i/N$, где n_i — количество остатков i в данной колонке, N — общее число последовательностей в выравнивании.
- Однако есть две **проблемы**:
- (i) Частоты остатков, не встречающихся в выравнивании, будут равны 0 — это исправляют с помощью псевдоотсчетов.
- (ii) Если среди всех последовательностей есть большая группа близкородственных, их сходство может исказить истинный сигнал — чтобы учесть это искажение, применяют взвешивание последовательностей.

Чрезмерный вклад родственных последовательностей

Одинаковая с виду колонка в двух различных ситуациях



- Последовательности, из которых состоит выравнивание, могут быть очень близкородственны, как последовательности s_1-s_3 в случае 2. Частота L в данной колонке равна ¾, но можем ли мы в обоих случаях считать, что L более вероятна в данной позиции, чем C?
- Нет, это предположение верно для случая 1, а для случая 2 три L являются скорее следствием близости содержащих L белков.
- Если в нашем выравнивании много близкородственных последовательностей, они будут ухудшать качество нашей матрицы. Чтобы этого избежать, частоты остатков в колонке оценивают с учетом весов (weights) последовательностей, которые содержат эти остатки в данной колонке. Веса последовательностей рассчитывают так, чтобы суммарный вес группы близкородственных последовательностей был ненамного больше веса последовательности, не имеющей близких родственников.

Для PWM (сигнал в геноме) такой проблемы обычно нет, а для PSSM (принадлежность белка семейству) она почти всегда актуальна. Стоит подумать, почему.

Оценка частоты остатка в позиции с учетом веса последовательности

Придумали такой способ: присвоить последовательности вес (weight) так, чтобы у последовательностей, имеющих много родственников, он был маленьким, а у «одиноких» последовательностей — большой. При расчете частоты остатка i в позиции k используются веса последовательностей w_s:

$$f_{ki} = \frac{\sum\limits_{s:a_{sk}=i} w_s + \psi_i}{\sum\limits_{s} w_s + \sum\limits_{i} \psi_i}$$

Если все веса последовательностей равны, то получится обычная частота. Здесь a_{sk} — буква последовательности s в позиции k, ψ_i — псевдоотсчёт для остатка i.

Внимание: слово «вес» имеет два разных значения

- Вес = Score, вес выравнивания двух последовательностей или последовательности относительно профиля (PWM или PSSM или HMM), обычно обозначается S.
- Вес = Weight, вес последовательности, используемый при построении PSSM по множественному выравниванию, обычно обозначается w.

Дополнительная информация: как получается вес последовательности?

Как получить такой вес последовательности (w_s), чтобы он был маленьким у последовательностей, имеющим в нашем выравнивании близких родственников, и большим у одиноких последовательностей?

Например, так (а всего около десятка только популярных способов).

- Дано выравнивание: последовательность s содержит букву $a_{s,k}$ в позиции k.

  Сначала припишем вес каждой **букве** выравнивания:
  - Пусть в k-ой позиции выравнивания встречается $r(k)$ типов аминокислотных остатков. Сделаем так, чтобы суммарный вес каждого типа был равен $1/r(k)$ (то есть суммарный вес всех аланинов равнялся суммарному весу лейцинов и т.д.)
  - Для этого посмотрим, в скольких последовательностях содержится каждый тип остатка. Пусть такой же остаток, как $a_{s,k}$, встречается в нашей позиции k всего в $n(a_{s,k}, k)$ разных последовательностях. Припишем букве $a_{s,k}$ вес $w_{s,k} = 1/r(k)n(a_{s,k}, k)$.
- Вес последовательности будет равен сумме весов всех её букв:

$$w_s = \sum_k w_{s,k}$$

Добавление псевдоотсчётов

Используются по крайней мере три способа избавиться от 0 в матрице частот:

1. Добавление 1 к наблюдаемому количеству каждой буквы (правило Лапласа). Не учитываем разную частоту встречаемости разных остатков. Стандартный вариант для нуклеотидов (PWM).
2. Добавление фоновой частоты остатка в банке белковых последовательностей — лучше, но не учитываем свойства остатков.

Например, если в данной позиции выравнивания стоит Leu, то вероятность появления в этой позиции похожего по свойствам остатка (например, Met) должна быть больше фоновой, а непохожего (например, Asp) — меньше фоновой.

3. Добавление q_{ij} : частот замен из «образцовых» выравниваний (тех же, что использовались при создании матриц замен остатков, например BLOSUM62)

Матрица BLOSUM62

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

$$S_{ij} = \frac{1}{\lambda} \log \frac{q_{ij}}{p_i p_j}$$

здесь S_{ki} — элемент матрицы замен, $p_{i,} p_j$ — фоновые частоты остатков i и j, q_{ij} — частота замены остатка i на j в «образцовых» выравниваниях (из базы BLOCKS).

Расчет ожидаемой частоты остатка

$$g_i^k = \sum_j \frac{f_j^k}{p_j} q_{ij}$$

где g_i^k — псевдоотсчет для остатка i в позиции k, f_j^k — наблюдаемая частота остатка j в выравнивании, p_j — фоновая частота остатка j, q_{ij} — частота пары i, j в «образцовых» выравниваниях

$$Q_i^k = \frac{\alpha f_i^k + \beta g_i^k}{\alpha + \beta}$$

где Q_i^k — ожидаемая частота для остатка i в позиции k, g_i^k — псевдоотсчёт для остатка i

f_i^k — наблюдаемая частота остатка i в позиции k,

$\alpha = Ne - 1$, где Ne — эффективный размер выравнивания (среднее число **различных** остатков по позициям),

$\beta = 10$.

Окончательная формула для элемента PSSM

$$S_{ki} = \frac{1}{\lambda} \log \frac{Q_{ki}}{p_i}$$

где S_{ki} — элемент PSSM (вес остатка i в позиции k),
Q_{ki} — ожидаемая частота остатка i в позиции k, с учетом весов последовательностей и псевдоотсчетов,
p_i — фоновая частота остатка i ,
λ — константа (для удобства)

Использование PSSM

PSSM можно «выравнять» с белковой последовательностью и получить вес, аналогично весу выравнивания двух последовательностей.

PSSM используется при поиске в банке данных программой PSI-BLAST и программами пакета MEME.

PSI-BLAST (Position-Specific Iterative BLAST) — разновидность BLASTP, использующий PSSM, благодаря чему он способен находить дальних родственников заданного белка.

Алгоритм PSI-BLAST

- На входе — последовательность и порог по e-value, на выходе — набор найденных последовательностей и построенный по ним PSSM.
- 1. На первом этапе запускается обычный BLASTP входной последовательности против выбранного банка последовательностей
- 2. Для находок со значениями e-value лучше заданного порога строится множественное выравнивание.
- 3. Это выравнивание используется для получения PSSM.
- 4. На следующем шаге опять происходит запуск BLAST для исходной последовательности против того же банка последовательностей, но вместо матрицы замен остатков используется PSSM, полученная на предыдущем шаге.
- 5. Повторяем шаги 2-4, пока не перестанут добавляться новые последовательности.

Дополнительные возможности PSI-BLAST

- Можно вручную включать/исключать последовательности, которые используются для построения PSSM
- Можно использовать PSSM, созданную на основе поиска в одном банке, для поиска в другом банке.

Паттерны

- Запись выравнивания в виде регулярного выражения
- Правила записи:
  `http://www.hpa-bioinfotools.org.uk/ps_scan/PS_SCAN_PATTERN_SYNTAX.html`
- Пример паттерна

  < A-x-[ST](2)-x(0,1)-V

Банк ProSite

https://prosite.expasy.org/

Коллекция белковых семейств и доменов.

Аннотации эволюционных доменов.

Мотивы: функциональные участки и «подписи» семейств белков в виде паттернов и HMM-профилей.

Интерфейс (средства поиска, средства сохранения выравниваний и т. д.)

Можно:

1. Искать мотивы из коллекции ProSite в своем белке.
2. Искать свой мотив в коллекции последовательностей ProSite.
3. Искать свой мотив в своем белке или белках.

Поиск паттернов

- В пакете EMBOSS есть программы fuzznuc и fuzzpro для поиска паттернов в нуклеотидных и белковых последовательностях, соответственно.
- На вход — паттерн и последовательность, на выходе — позиция и вес найденных совпадений