Профильные HMM

Способы описания множественного выравнивания

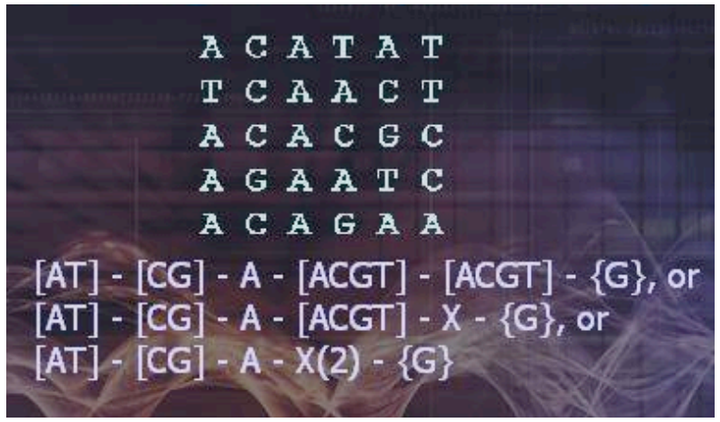
Дано: множественное выравнивание

Задача: определить принадлежит ли последовательность данному семейству

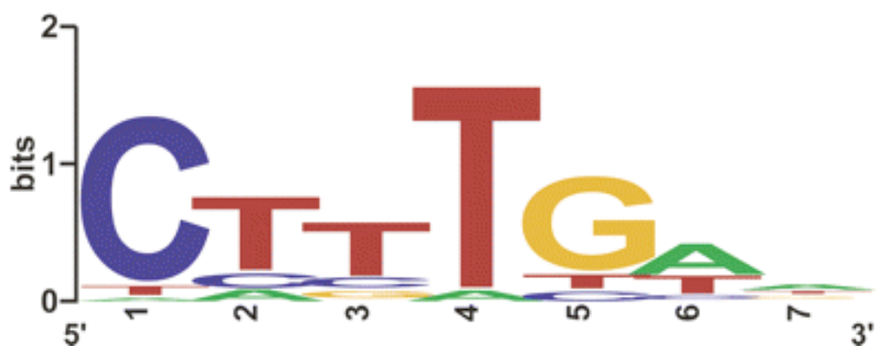
Варианты представления информации о семействе:

- Консенсус - доминирующий остаток в колонке
- Регулярное выражение
- Матрица частот встречаемости

```
AQP1.PRO     TLGLLLSCQISILRAVMYIIAQCVGAIVASAIL
AQP2.PRO     TVACLVGCHVSFLRAAFYVAAQLLGAVAGAAIL
AQP3.PRO     TFAMCFLAREPWIKLPIYTLAQTLGAFLGAGIV
AQP4.PRO     TVAMVCTRKISIAKSVFYITAQCLGAIIGAGIL
AQP5.PRO     TLALLIGNQISLLRAVFYVAAQLVGAIAGAGIL
consensus    T.A.l....iS.lravfY..AQ.lGAi.GAgIL
```

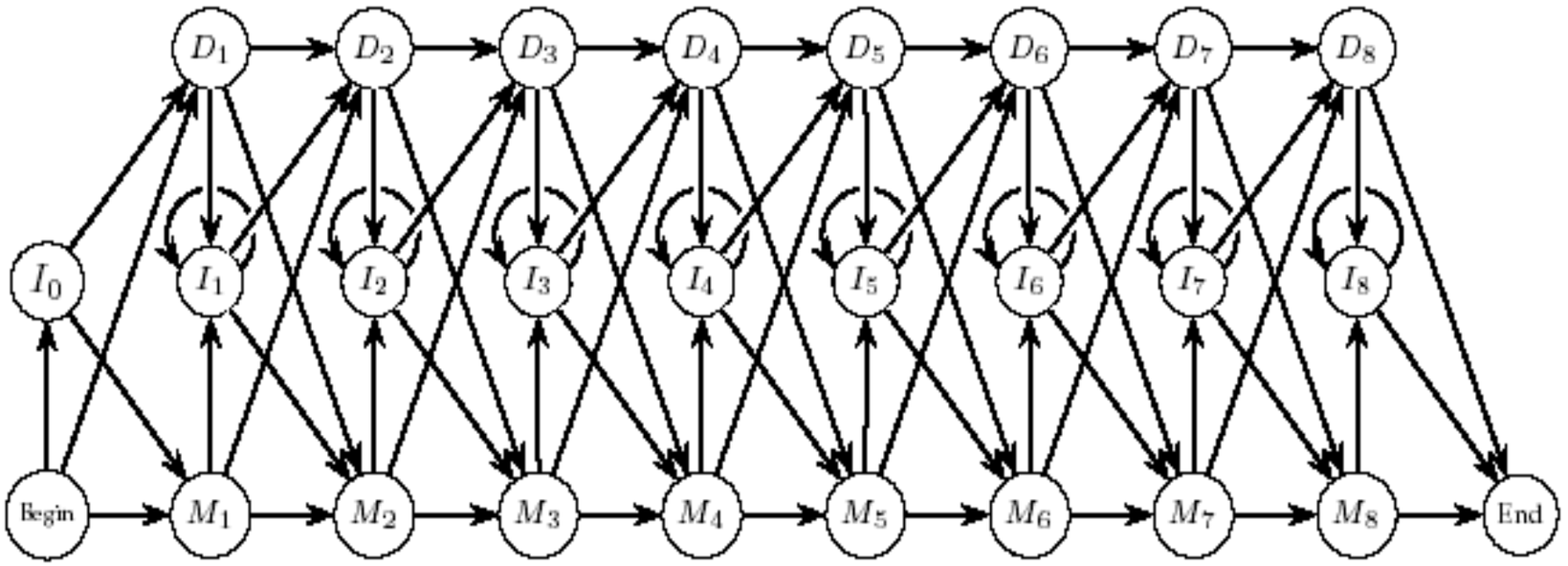


	1	2	3	4	5	6	7
A	1	4	1	2	0	17	13
C	28	5	5	0	3	3	2
G	0	0	4	0	25	1	7
T	2	22	21	29	4	10	9



Профильная HMM. Структура модели.

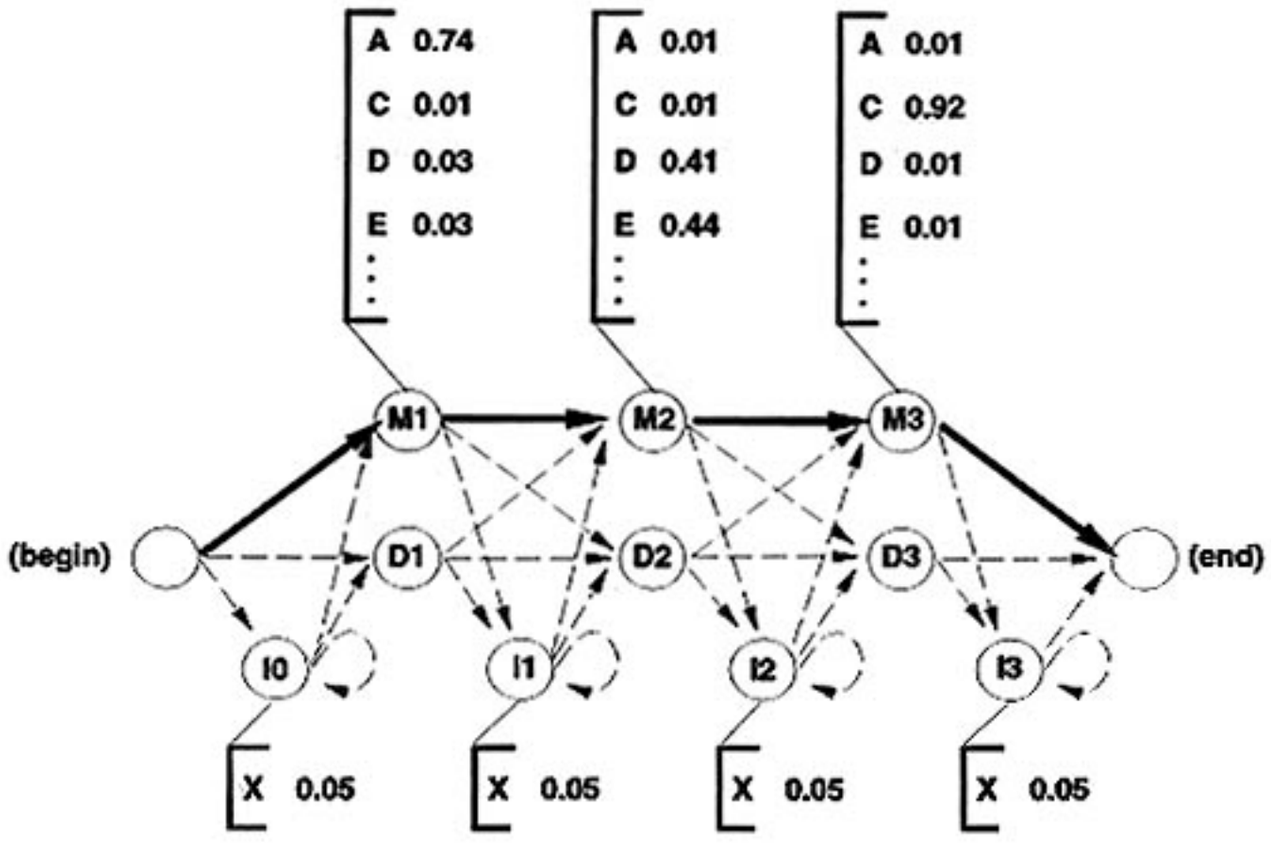
- Профильная HMM - вероятностное представление множественного выравнивания
- Для построения профильной HMM достаточно иметь множественное выравнивание последовательностей
- Построенная модель может быть использована для поиска новых представителей белкового семейства



- M_1,..,M_N, *Begin*, *End* - состояния совпадений
- I_0,...,I_N - состояния вставок
- D_1,...,D_N - состояния делеций

Структура профильной HMM

- Состояния совпадений и вставок имеют эмиссионные вероятности остатков
- Состояния делеций - молчащие - не порождают символы



Поиск при помощи профильной HMM

Вес соответствия последовательности модели - вероятность P(x|M) можем оценить одним из двух способов:

- Получая наиболее вероятное выравнивание (путь) π* последовательности *x* к модели *M* - *P(x,* π* |*M)*. Алгоритм Витерби.
- Вычисляя полную вероятность *P(x|M)*. Алгоритм просмотра вперед.

Построение профильной HMM по множественному выравниванию

1. Определение структуры (размера) профильной HMM: обычно используется правило назначения состояниями совпадений - колонок выравнивания, не имеющих более 50% делеций.

2. Вероятности переходов и эмиссионные вероятности можно определить используя обычные оценки:

$$a_{kl} = \frac{A_{kl}}{\sum_i A_{ki}} \qquad e_k(b) = \frac{E_k(b)}{\sum_c E_k(c)}$$

Псевдоотчеты

- Правило Лапласа: добавление ко всем счетчикам константы

$$e_i(a) = \frac{c_{ja} + C}{\sum_b (c_{jb} + C)}$$

где, $e_i(a)$ - эмиссионная вероятность остатка *a* в колонке *i*;
c_{ja} - счетчик аминокислоты a в колонке i;
C - константа

- Добавление значений, пропорциональных фоновому распределению остатков

$$e_i(a) = \frac{c_{ja} + Aq_a}{\sum_b c_{jb} + A}$$

где, q_a - фоновая частота остатка *a*;
A - константа

Построение профильной HMM по множественному выравниванию: пример

Использовать:
для псевдоотчетов - правило Лапласа,
для эмиссионных вероятностей состояний вставок - фоновые вероятности аминокислот, считая их равновероятными.

```
...VGA--HAGEY...
...V----NVDEV...
...VEA--DVAGH...
...VKG------D...
...VYS--TYETS...
...FNA--NIPKH...
...IAGADNGAGV...
```

Благодарности

- При подготовке слайдов использовались материалы лекций:
  - Михаила Гельфанда (ИППИ)
  - Андрея Миронова (МГУ)
  - Serafim Batzoglou (Stanford)
  - Manolis Kellis (MIT)
  - Pavel Pevzner (UCSD)