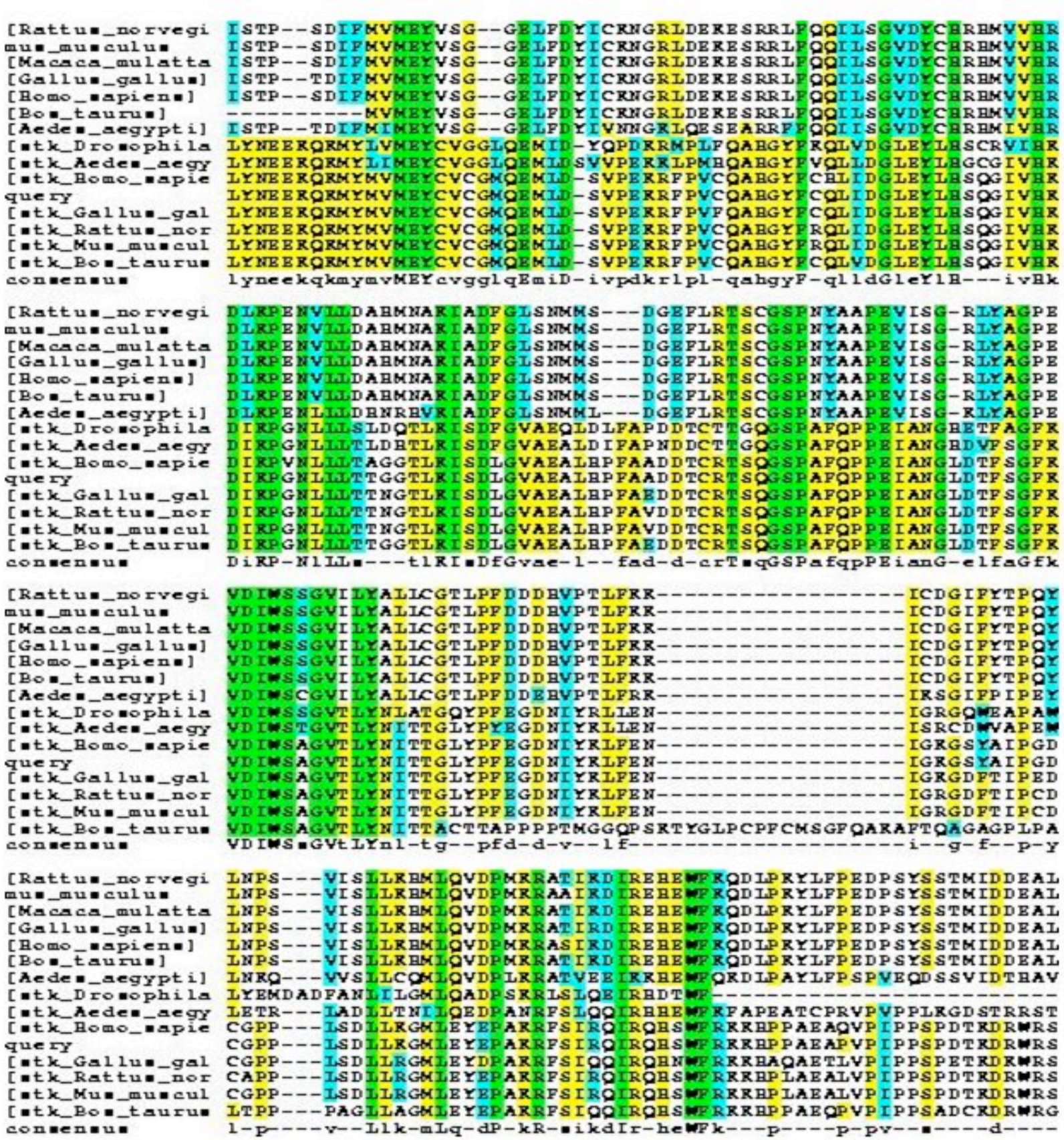
Множественное выравнивание последовательностей

Множественное выравнивание последовательностей

Две главные состовляющие автоматических методов множественного выравнивания:

- оценка качества выравнивания (функция оценки)
- алгоритм выравнивания



Оценка качества выравнивания: наименьшая энтропия

Обычно считают, что колонки $\{m_i\}$ в множественном выравнивании независимы. Тогда вес выравнивания:

$$S(m) = \sum_i S(m_i)$$

Информационная энтропия (мера неопределенности или мера беспорядка):

$$H(x) = -\sum_i^n p(i) \log p(i)$$

Можем задать вес колонки m_i как:

$$S(m_i) = -\sum_a p_{ia} \log p_{ia}$$

где $p_{ia} = \frac{c_{ia}}{\sum_b c_{ib}}$, а c_{ia} - количество остатков *a* в колонке *i*

Оценка качества выравнивания: сумма пар

Оценка качества колонки функцией “сумма пар”:

$$S(m_i) = \sum_{k<l} s(m_{ik}, m_{il})$$

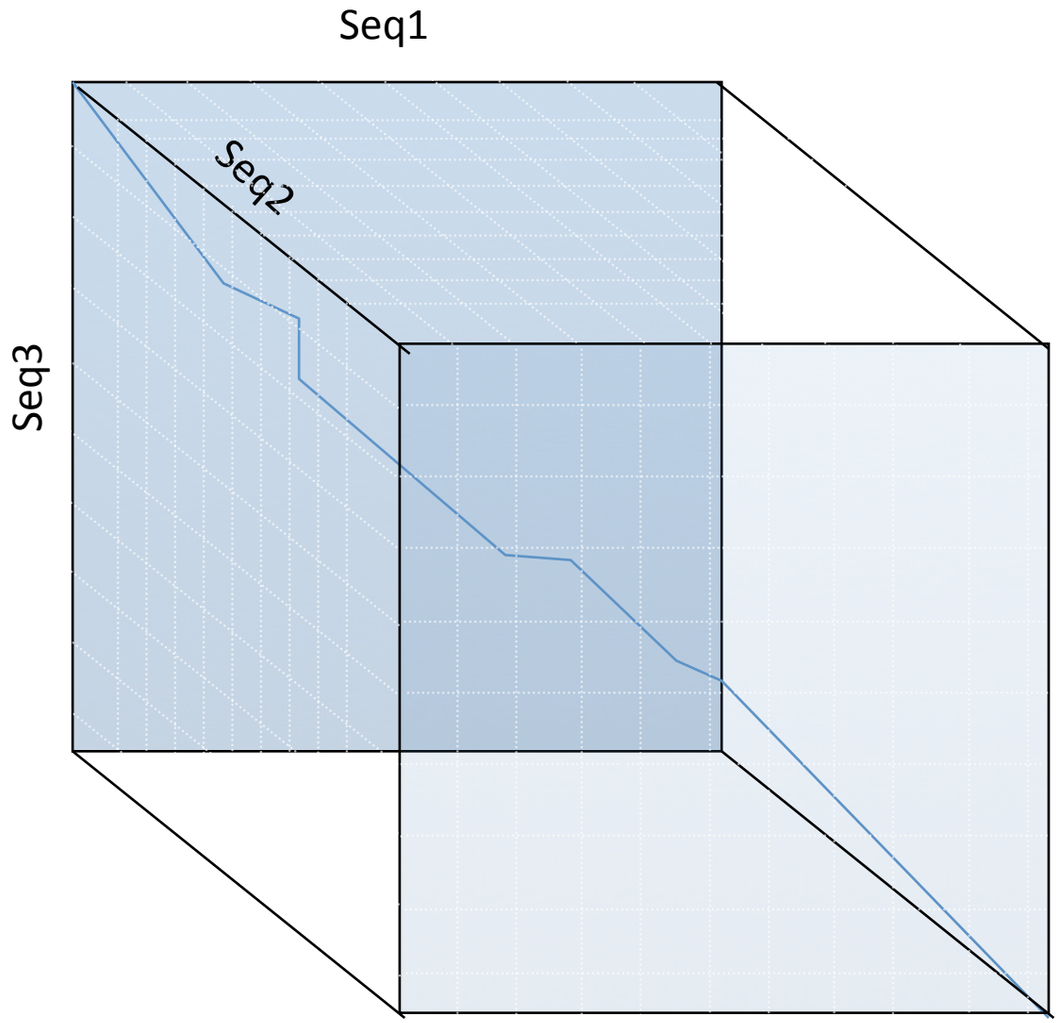
где m_{ik} - остаток в i-й колонке и k-ой строке,

а $s(a,b)$ - вес замены остатка a на остаток b, вычисленный на основе матрицы замен (BLOSUM или PAM)

Многомерное динамическое программирование

$S(x_{i_1}^1, x_{i_2}^2, \ldots, x_{i_N}^N)$ - вес колонки, составленной из $i_1, i_2, .., i_N$-ых символов последовательностей $x^1, x^2, .., x^N$

$\alpha_{i1,i2,\ldots,iN}$ - вес оптимального выравнивания в ячейке $(i_1, i_2, \ldots, i_N)$



Рекурсия для трехмерного случая:

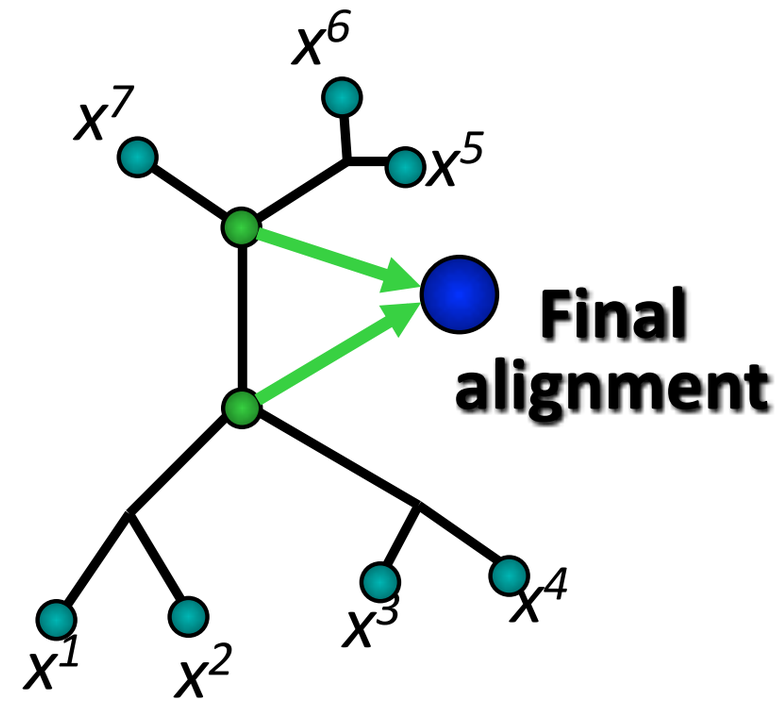
$$\alpha_{i,j,k} = \max \begin{cases} \alpha_{i-1,j-1,k-1} + S(x_i^1, x_j^2, x_k^3) \\ \alpha_{i,j-1,k-1} + S(-, x_j^2, x_k^3) \\ \alpha_{i-1,j,k-1} + S(x_i^1, -, x_k^3) \\ \alpha_{i-1,j-1,k} + S(x_i^1, x_j^2, -) \\ \alpha_{i,j,k-1} + S(-, -, x_k^3) \\ \alpha_{i-1,j,k} + S(x_i^1, -, -) \\ \alpha_{i,j-1,k} + S(-, -, -) \end{cases}$$

Многомерное динамическое программирование

- Количество ячеек равно $L_1 \times L_2 \times ... \times L_N = O(L^N)$
- Для вычисления веса в каждой ячейке необходимо найти максимум из 2^N-1 вариантов
- Вычислительная сложность:
  - по памяти - $O(L^N)$
  - по времени - $O(2^N L^N)$

Методы прогрессивного выравнивания

- Методы построения множественного выравнивания путем последовательного объединения попарных выравниваний
- Как правило, первоначально строится приближенное филогенетическое дерево
- С помощью дерева выбирается порядок построения попарных выравниваний
- Далее последовательности добавляются поочередно к главному выравниванию, либо подвыравнивания (профили) выравниваются друг с другом



Выравнивание профилей

- Выравнивание профилей (стопок последовательностей) выполняется обычным алгоритмом динамического программирования

- Оптимизируется сумма парных весов:

$$\sum_i S(m_i) = \sum_i \sum_{k<l\le N} s(m_i^k, m_i^l) =$$

$$\sum_i \sum_{k<l\le n} s(m_i^k, m_i^l) + \sum_i \sum_{n<k<l\le N} s(m_i^k, m_i^l) + \sum_i \sum_{k<n;n<l\le N} s(m_i^k, m_i^l)$$

ClustalW - популярный метод множественного выравнивания

- Строится матрица расстояний с использованием попарных выравниваний
- Строится направляющее дерево с помощью метода соединения соседей
- Выравнивание строится в порядке убывания сходства последовательностей. Выполняются выравнивания последовательности к последовательности, последовательности к профилю и профиля к профилю
- Дополнительные эвристики:
  - взвешивание последовательностей
  - использование различных матриц замен (BLOSUM80/50)
  - корректировка дерева при низком весе выравнивания

Благодарности

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