

Множественное выравнивание последовательностей

Множественное выравнивание последовательностей

Две главные составляющие автоматических методов множественного выравнивания:

- оценка качества выравнивания (функция оценки)
- алгоритм выравнивания

```
[Rattus_norvegicus_musculus] I STP--SDIPMVMEYVSG--GELFDYICRNGRLDERESRRLPQQIILSGVDYCHRMVVR
[Macaca_mulatta] I STP--SDIPMVMEYVSG--GELFDYICRNGRLDERESRRLPQQIILSGVDYCHRMVVR
[Gallus_gallus] I STP--TDIPMVMEYVSG--GELFDYICRNGRLDERESRRLPQQIILSGVDYCHRMVVR
[Homo_sapiens] I STP--SDIPMVMEYVSG--GELFDYICRNGRLDERESRRLPQQIILSGVDYCHRMVVR
[Bos_taurus] -----MVMEYVSG--GELFDYICRNGRLDERESRRLPQQIILSGVDYCHRMVVR
[Aedes_aegypti] I STP--TDIPMIMEYVSG--GELFDYIVNNGRLQESARRFPQQIILSGVDYCHRMIVR
[stk_Drosophila] LYNEERQRMYLIMEYCVGGLOEMID-YQPDARMPLFQAHGYFRQLVDGLEYLBSCRVIR
[stk_Aedes_aegy] LYNEERQRMYLIMEYCVGGLOEMID-SVPERKRLPMHQAHGYFVQLLDGLEYLBSGQIVR
[stk_Homo_sapie] LYNEERQRMYMVMEYCVCGMQEMLD-SVPERKRPVPCQAHGYFCQLIDGLEYLBSGQIVR
query LYNEERQRMYMVMEYCVCGMQEMLD-SVPERKRPVPCQAHGYFCQLIDGLEYLBSGQIVR
[stk_Gallus_gal] LYNEERQRMYMVMEYCVCGMQEMLD-SVPERKRPVPCQAHGYFCQLIDGLEYLBSGQIVR
[stk_Rattus_nor] LYNEERQRMYMVMEYCVCGMQEMLD-SVPERKRPVPCQAHGYFRQLIDGLEYLBSGQIVR
[stk_Mus_muscul] LYNEERQRMYMVMEYCVCGMQEMLD-SVPERKRPVPCQAHGYFRQLIDGLEYLBSGQIVR
[stk_Bos_taurus] LYNEERQRMYMVMEYCVCGMQEMLD-SVPERKRPVPCQAHGYFRQLVDGLEYLBSGQIVR
consensus lyneekqkmyvmeyavgglqemid-ivpdkrlpl-qahgyf-qlldgleylbr---ivhk

[Rattus_norvegicus_musculus] DLKPEENVLLDAHNNARIADPGLSNMMS---DGEFLRTSCGSPNYAAPEVISG-RLYAGPE
[Macaca_mulatta] DLKPEENVLLDAHNNARIADPGLSNMMS---DGEFLRTSCGSPNYAAPEVISG-RLYAGPE
[Gallus_gallus] DLKPEENVLLDAHNNARIADPGLSNMMS---DGEFLRTSCGSPNYAAPEVISG-RLYAGPE
[Homo_sapiens] DLKPEENVLLDAHNNARIADPGLSNMMS---DGEFLRTSCGSPNYAAPEVISG-RLYAGPE
[Bos_taurus] DLKPEENVLLDAHNNARIADPGLSNMMS---DGEFLRTSCGSPNYAAPEVISG-RLYAGPE
[Aedes_aegypti] DLKPEENVLLDAHNNARIADPGLSNMML---DGEFLRTSCGSPNYAAPEVISG-RLYAGPE
[stk_Drosophila] DIRPGNLLLSLDQTLRI SDPGVAEQDLDFAPDDTCTTGQSPAFQPPPEIANGHETVAGFR
[stk_Aedes_aegy] DIRPGNLLLSLDQTLRI SDPGVAEQDLDFAPDDTCTTGQSPAFQPPPEIANGHETVAGFR
[stk_Homo_sapie] DIRPVNLLLSLDQTLRI SDPGVAEQDLDFAPDDTCTTGQSPAFQPPPEIANGHETVAGFR
query DIRPGNLLLSLDQTLRI SDPGVAEQDLDFAPDDTCTTGQSPAFQPPPEIANGHETVAGFR
[stk_Gallus_gal] DIRPGNLLLSLDQTLRI SDPGVAEQDLDFAPDDTCTTGQSPAFQPPPEIANGHETVAGFR
[stk_Rattus_nor] DIRPGNLLLSLDQTLRI SDPGVAEQDLDFAPDDTCTTGQSPAFQPPPEIANGHETVAGFR
[stk_Mus_muscul] DIRPGNLLLSLDQTLRI SDPGVAEQDLDFAPDDTCTTGQSPAFQPPPEIANGHETVAGFR
[stk_Bos_taurus] DIRPGNLLLSLDQTLRI SDPGVAEQDLDFAPDDTCTTGQSPAFQPPPEIANGHETVAGFR
consensus DiRP-NLLS---tlRI SDPGvae-l--fad-d-art qGSPafqpPEianG-elfaGfk

[Rattus_norvegicus_musculus] VDIWSSGVILYALICGTLPPDDDBVPTLFRR-----ICDGIPTPQY
[Macaca_mulatta] VDIWSSGVILYALICGTLPPDDDBVPTLFRR-----ICDGIPTPQY
[Gallus_gallus] VDIWSSGVILYALICGTLPPDDDBVPTLFRR-----ICDGIPTPQY
[Homo_sapiens] VDIWSSGVILYALICGTLPPDDDBVPTLFRR-----ICDGIPTPQY
[Bos_taurus] VDIWSSGVILYALICGTLPPDDDBVPTLFRR-----ICDGIPTPQY
[Aedes_aegypti] VDIWSSGVILYALICGTLPPDDDBVPTLFRR-----IRSGIFPIPEY
[stk_Drosophila] VDIWSSGVILYALICGTLPPDDDBVPTLFRR-----IRSGIFPIPEY
[stk_Aedes_aegy] VDIWSTGVTLYNITGLYPFEGDNIYKLFEN-----ISRCMDVAPEN
[stk_Homo_sapie] VDIWSAGVTLYNITGLYPFEGDNIYKLFEN-----IGKGSYAIPGD
query VDIWSAGVTLYNITGLYPFEGDNIYKLFEN-----IGKGSYAIPGD
[stk_Gallus_gal] VDIWSAGVTLYNITGLYPFEGDNIYKLFEN-----IGKGSYAIPGD
[stk_Rattus_nor] VDIWSAGVTLYNITGLYPFEGDNIYKLFEN-----IGKGSYAIPGD
[stk_Mus_muscul] VDIWSAGVTLYNITGLYPFEGDNIYKLFEN-----IGKGSYAIPGD
[stk_Bos_taurus] VDIWSAGVTLYNITGLYPFEGDNIYKLFEN-----IGKGSYAIPGD
consensus VDIWSAGVTLYNITGLYPFEGDNIYKLFEN-----IGKGSYAIPGD

[Rattus_norvegicus_musculus] LNPS---VISLLKBMQLQVDPKRRATIKDIREHEWFRQDLPRYLFPEDPSYSSTMIDDEAL
[Macaca_mulatta] LNPS---VISLLKBMQLQVDPKRRATIKDIREHEWFRQDLPRYLFPEDPSYSSTMIDDEAL
[Gallus_gallus] LNPS---VISLLKBMQLQVDPKRRATIKDIREHEWFRQDLPRYLFPEDPSYSSTMIDDEAL
[Homo_sapiens] LNPS---VISLLKBMQLQVDPKRRATIKDIREHEWFRQDLPRYLFPEDPSYSSTMIDDEAL
[Bos_taurus] LNPS---VISLLKBMQLQVDPKRRATIKDIREHEWFRQDLPRYLFPEDPSYSSTMIDDEAL
[Aedes_aegypti] LNRQ---VVSLLCQMLQVDPKRRATIKDIREHEWFRQDLPRYLFPEDPSYSSTMIDDEAL
[stk_Drosophila] LYEMDADFANLILGMLQADPSKRLSLQELRBDTWF-----
[stk_Aedes_aegy] LETR---LADLLTNILQEDPANRFSLQQIRQHSEWFRKRPAPAEAPVIPPSPDTRDRWRS
[stk_Homo_sapie] CGPP---LSDLLKGMLEYEPARRFSIRQIRQHSEWFRKRPAPAEAPVIPPSPDTRDRWRS
query CGPP---LSDLLKGMLEYEPARRFSIRQIRQHSEWFRKRPAPAEAPVIPPSPDTRDRWRS
[stk_Gallus_gal] CGPP---LSDLLKGMLEYEPARRFSIRQIRQHSEWFRKRPAPAEAPVIPPSPDTRDRWRS
[stk_Rattus_nor] CAPP---LSDLLKGMLEYEPARRFSIRQIRQHSEWFRKRPAPAEAPVIPPSPDTRDRWRS
[stk_Mus_muscul] CGPP---LSDLLKGMLEYEPARRFSIRQIRQHSEWFRKRPAPAEAPVIPPSPDTRDRWRS
[stk_Bos_taurus] LTPP---PAGLLAGMLEYEPARRFSIRQIRQHSEWFRKRPAPAEAPVIPPSPDTRDRWRS
consensus l-p---v--Llk-mLq-dP-kR-sikdIr-heWfk---p---p-pv-■-----d---
```


Оценка качества выравнивания: наименьшая энтропия

Обычно считают, что колонки 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)

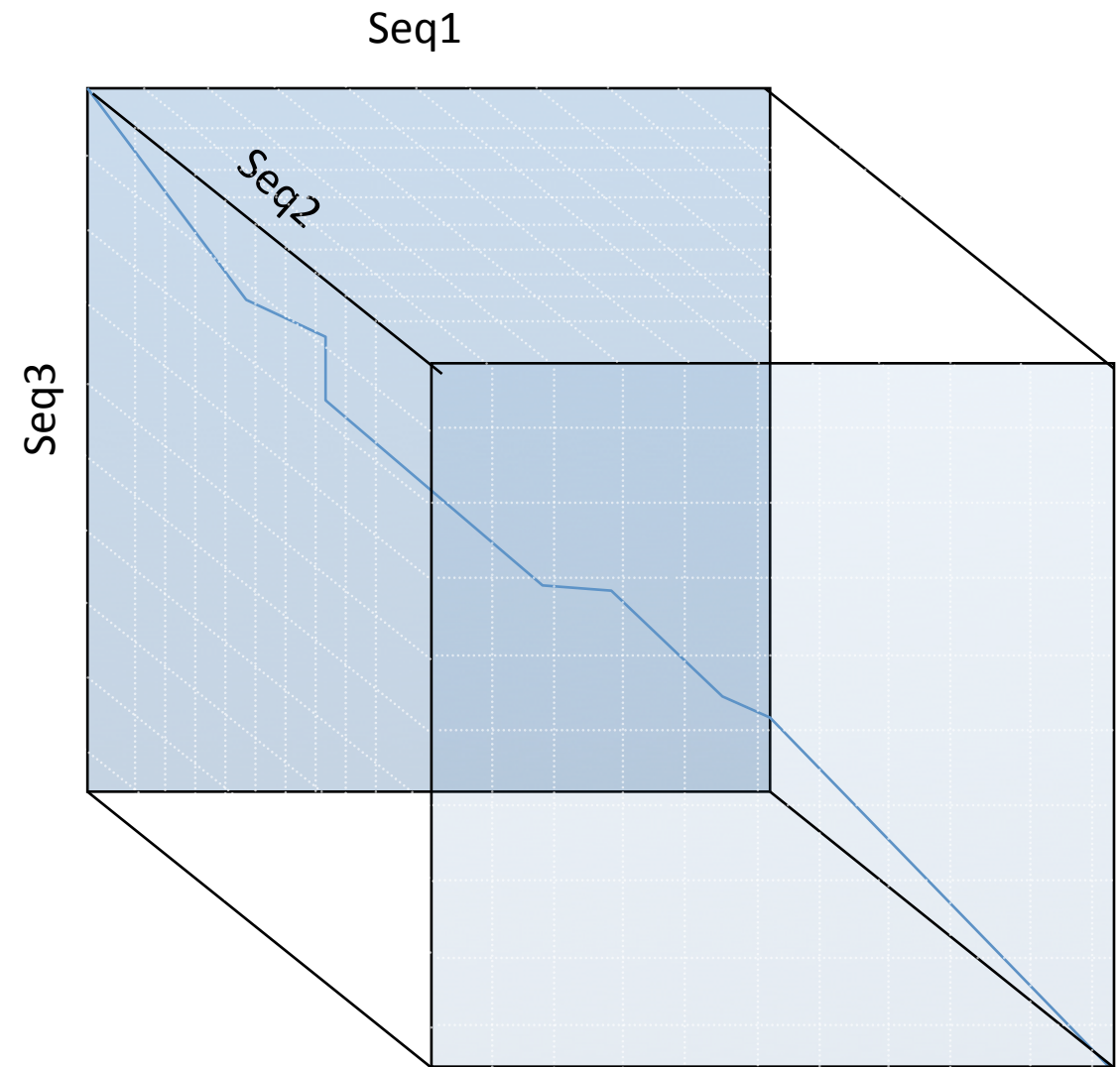
Многомерное динамическое программирование

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

$\alpha_{i_1, i_2, \dots, i_N}$ - вес оптимального выравнивания в ячейке $(i_1, i_2, \dots, i_N)$

Рекурсия для трехмерного случая:

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

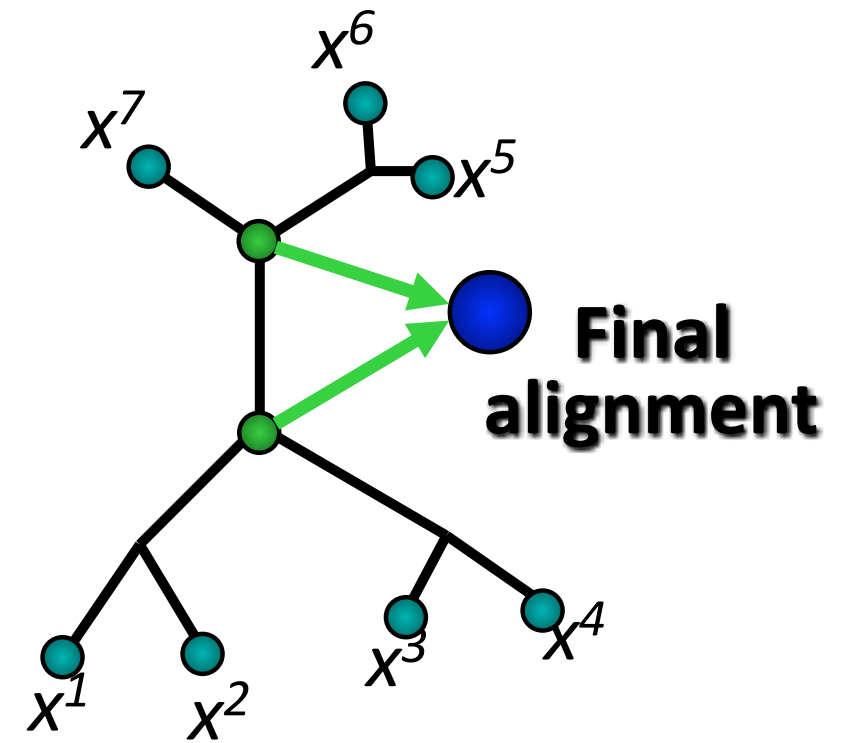


Многомерное динамическое программирование

- Количество ячеек равно $L_1 \times L_2 \times \dots \times L_N = O(L^N)$
- 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:

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

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

ClustalW - популярный метод множественного выравнивания

- Строится матрица расстояний с использованием попарных выравниваний
- Строится направляющее дерево с помощью метода ближайшего соседа
- 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)