

Множественные выравнивания

Профили

Обобщение парного выравнивания

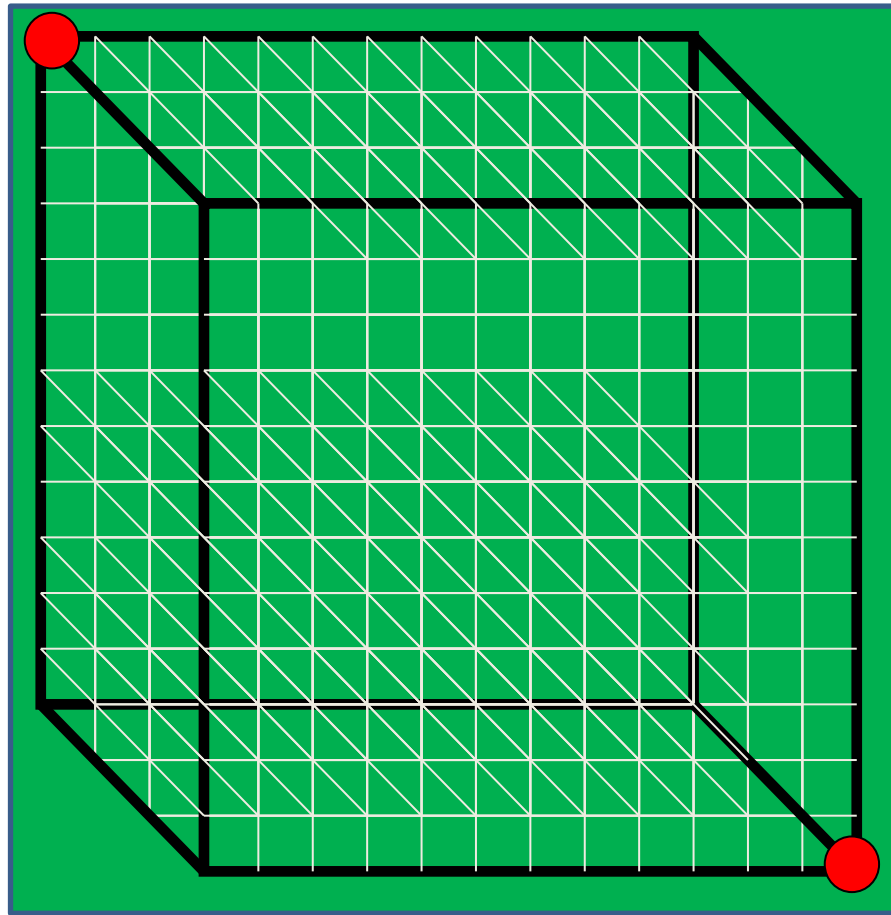
- Выравнивание 2-х последовательностей – двумерная матрица
- 3-х последовательностей – 3-х мерная.

A	T	_	G	C	G	_
A	_	C	G	T	_	A
A	T	C	A	C	_	A

- Задача: больше консервативных столбцов, лучше выравнивание

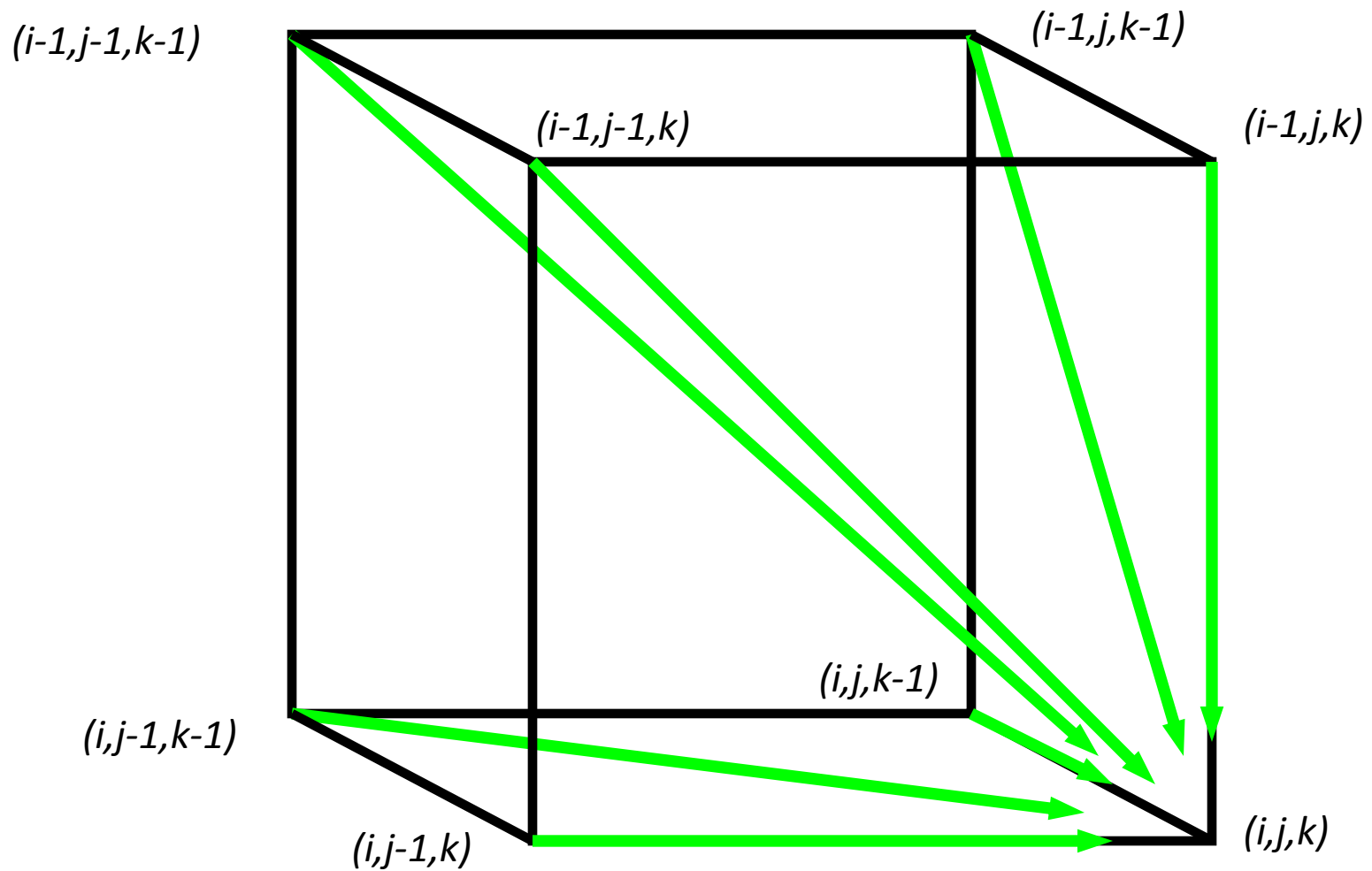
Глобальное выравнивание 3-х последовательностей

начало



конец

3-D архитектура



Алгоритм

- $s_{i,j,k} = \max \left\{ \begin{array}{ll} s_{i-1,j-1,k-1} + \delta(v_i, w_j, u_k) & \text{Нет гэпов} \\ s_{i-1,j-1,k} + \delta(v_i, w_j, _) & \\ s_{i-1,j,k-1} + \delta(v_i, _, u_k) & \text{Один гэп} \\ s_{i,j-1,k-1} + \delta(_, w_j, u_k) & \\ s_{i-1,j,k} + \delta(v_i, _, _) & \\ s_{i,j-1,k} + \delta(_, w_j, _) & \text{Два гэпа} \\ s_{i,j,k-1} + \delta(_, _, u_k) & \end{array} \right.$
- $\delta(x, y, z)$ – запись в трехмерной матрице весов

Время работы алгоритма

- Для 3-х последовательностей длины n , время работы – $7n^3$; $O(n^3)$
- Для k последовательностей - $(2k-1)(n^k)$; $O(2kn^k)$

Множественное выравнивание порождает парные выравнивания

x: ACGCGG-C
y: AC-GC-GAG
z: GCCGC-GAG

Порождает:

x: ACGCGG-C ;	x: AC-GCGG-C ;	y: AC-GCGAG
y: ACGC-GAC ;	z: GCCGC-GAG ;	z: GCCGCGAG

Обратная проблема

Имея 3 субъективных парных варнивания:

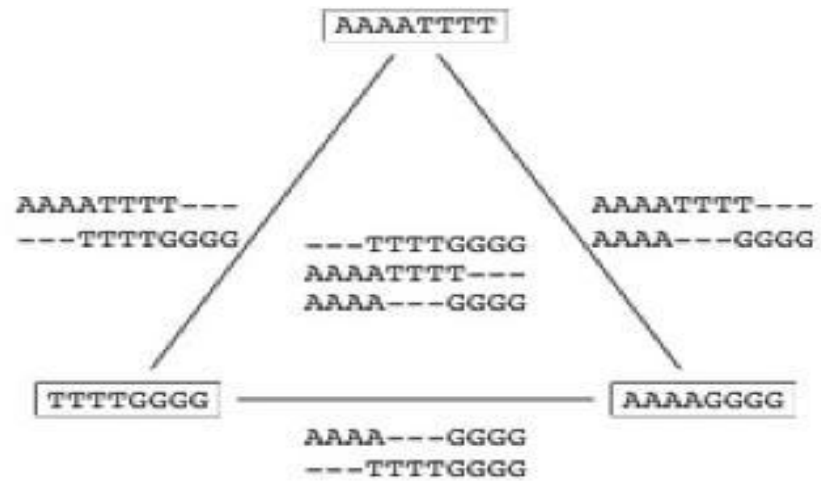
x: ACGCGG-C; **x:** AC-GCGG-C; **y:** AC-GCGAG
y: ACGC-GAC; **z:** GCCGC-GAG; **z:** GCCGCGAG

Обратная проблема

Имея 3 субъективных парных варнивания:

x: ACGCGG-C; **x:** AC-GCGG-C; **y:** AC-GCGAG
y: ACGC-GAC; **z:** GCCGC-GAG; **z:** GCCGCGAG

Хороший вариант



(a) Compatible pairwise alignments

Плохой вариант



(b) Incompatible pairwise alignments

Выравнивание выравниваний

x	GGGCACTGCAT	
y	GGTTACGTC--	Alignment 1
z	GGGAACCTGCAG	
w	GGACGTACC--	Alignment 2
v	GGACCT-----	

Описание выравнивания

$\left. \begin{array}{l} \text{GTC}^{\text{TA}}\text{GA} \\ \text{GTC}^{\text{AG}}\text{C} \end{array} \right\} \text{GTC}[\text{TA}]\text{G}[\text{AC}] - \text{профиль}$
 $[\text{X}][5\text{X}]$

x GGGCACTGCAT

y GGTTACGTC--

z GGGAACCTGCAG

w GGACGTACC--

v GGACCT-----

GGACACAGCAT - консенсус

Матрица частот – используется редко

НММ профиль

- Каждая колонка – отдельное состояние.
- Делеционные состояния – молчащие (не имеют эмиссии)
- Вероятность перехода в делеционное состояние зависит от позиции
- Значимость выравнивания с НММ профилем:

$$S = \log (P(x \mid M) / P(x \mid R)) = \sum \log \{b_i(x_i) / b_{random}(x_i)\}$$

Методы вычисления вероятности эмиссий

1. Нативный метод

$$b_k(i) = E_k(i) / \sum_j E_k(j)$$

2. Метод Лапласа

$$b_k(i) = (E_k(i) + 1) / (\sum_j E_k(j) + N)$$

3. Метод Байеса

$$b_k(i) = (E_k(i) + Aq_i) / (\sum_j E_k(j) + A)$$

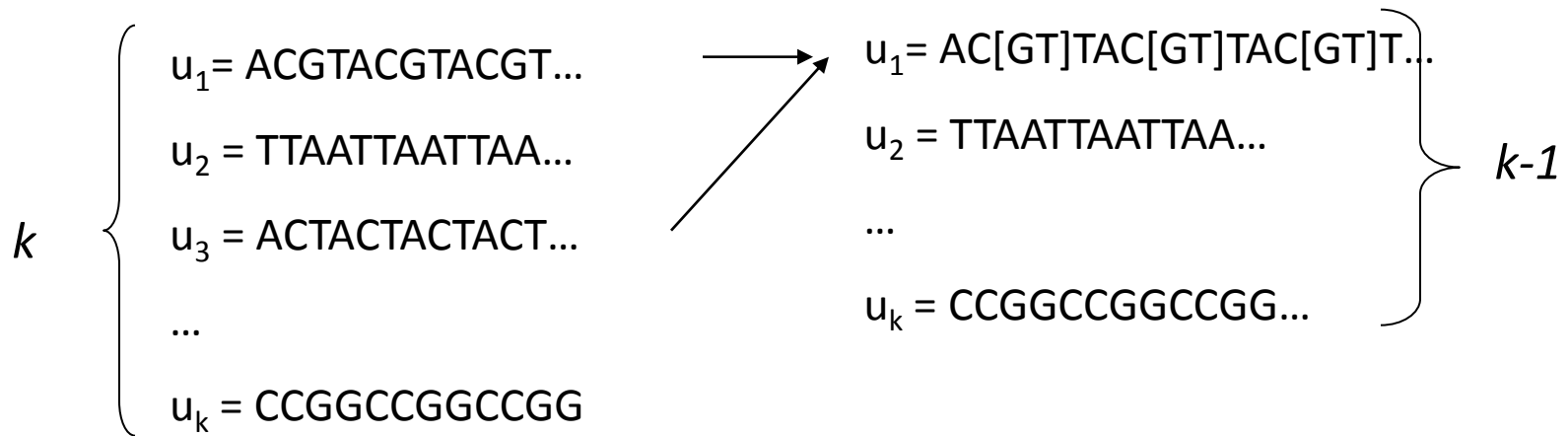
4. Метод матриц замен

$$E_{k(i)} = A \sum_j f_{kj} P(i \rightarrow j)$$

5. Метод общего предка

$$b_k(i) = \sum_j P_k(i \rightarrow j) P(pred_k = j \mid \text{alignment})$$

Множественное выравнивание – жадный алгоритм



Время работы алгоритма на k последовательностях длины n – $O(n^2k^2)$

Прогрессивное выравнивание ClustalW

- Прогрессивное выравнивание – жадный алгоритм с более «умным» способом выбора пар.
- Три шага
 - 1.) Построить парные выравнивания
 - 2.) Построить дерево-подсказку
 - 3.) Прогрессивное выравнивание по дереву-подсказке

Шаг 1: Парные Выравнивания

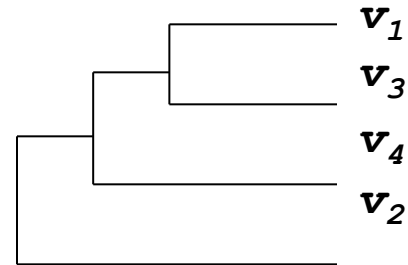
- Выравнивания пар порождают матрицу identity

	v_1	v_2	v_3	v_4
v_1				
v_2	.17	–		
v_3	.87	.28	–	
v_4	.59	.33	.62	–

(.17 значит идентичны на 17 %)

Шаг 2: Дерево-подсказка

	v_1	v_2	v_3	v_4
v_1				
v_2	.17	–		
v_3	.87	.28	–	
v_4	.59	.33	.62	–




Далее вычислить:

- $V_{1,3}$ = выравнивание (v_1, v_3)
- $V_{1,3,4}$ = выравнивание $((V_{1,3}), v_4)$
- $V_{1,2,3,4}$ = выравнивание $((V_{1,3,4}), v_2)$

Шаг 3: Прогрессивное выравнивание

- Выравниванием 2 наиболее близких последовательности.
- Следуя дереву - подсказке, довыравниваем следующую последовательность к имеющемуся выравниванию

```
FOS_RAT      PEEMSVTS-LDLTGGLPEATTPESSEEAFTLPLLNDPEPK-PSLEPVKNISNMELKAEPFD
FOS_MOUSE    PEEMSVAS-LDLTGGLPEASTPESEEAFTLPLLNDPEPK-PSLEPVKSISNVELKAEPFD
FOS_CHICK     SEELAAATALDLG----APSPAAAEFAFALPLMTEAPPAVPPKEPSG--SGLELKAEPFD
FOSB_MOUSE    PGPGPLAEVRDLPG-----STSAKEDGFGWLLPPPPPPP-----LPFQ
FOSB_HUMAN    PGPGPLAEVRDLPG-----SAPAKEDGFSWLLPPPPPPP-----LPFQ
.             . : ** . :.. *:.* * . * **:
```



Точки и звезды отображают насколько консервативны столбцы.

Множественные Выравнивания: Взвешивание

- Количество полных совпадений
- Сумма по парам (SP-Score)
- Энтропия

Количество полных совпадений

ААА
ААА
ААТ
АТС

- Хорошо только для очень близких последовательностей

Сумма по парам (SP-Score)

- Построим парное выравнивание по множественному
- Посчитаем веса всех этих парных выравниваний – $s(a_i, a_j)$
- Просуммируем:

$$s(a_1, \dots, a_k) = \sum_{i,j} s(a_i, a_j)$$

Энтропия

- Определим вероятности букв в столбцах
 - $p_A = 1, p_T=p_G=p_C=0$ (1-ый столбец)
 - $p_A = 0.75, p_T = 0.25, p_G=p_C=0$ (2-ый столбец)
 - $p_A = 0.50, p_T = 0.25, p_C=0.25, p_G=0$ (3-ий столбец)
- Энтропия столбца будет равна

$$- \sum_{X=A,T,G,C} p_X \log p_X$$

AAA
AAAT
AAT
ATC

Энтропия: Пример

Лучший вариант $\text{entropy} \begin{pmatrix} A \\ A \\ A \\ A \end{pmatrix} = 0$

Худший вариант $\text{entropy} \begin{pmatrix} A \\ T \\ G \\ C \end{pmatrix} = -\sum \frac{1}{4} \log \frac{1}{4} = -4(\frac{1}{4} * -2) = 2$

Энтропия: Пример

Энтропия столбца:

$$-(p_A \log p_A + p_C \log p_C + p_G \log p_G + p_T \log p_T)$$

A	A	A
A	C	C
A	C	G
A	C	T

- Столбец 1 = $-[1 \cdot \log(1) + 0 \cdot \log 0 + 0 \cdot \log 0 + 0 \cdot \log 0]$
= 0

- Столбец 2 = $-[(1/4) \cdot \log(1/4) + (3/4) \cdot \log(3/4) + 0 \cdot \log 0 + 0 \cdot \log 0]$
= $-[(1/4) \cdot (-2) + (3/4) \cdot (-.415)] = +0.811$

- Столбец 3 = $-[(1/4) \cdot \log(1/4) + (1/4) \cdot \log(1/4) + (1/4) \cdot \log(1/4) + (1/4) \cdot \log(1/4)]$
= $4 \cdot -[(1/4) \cdot (-2)] = +2.0$

- Энтропия выравнивания = $0 + 0.811 + 2.0 = +2.811$