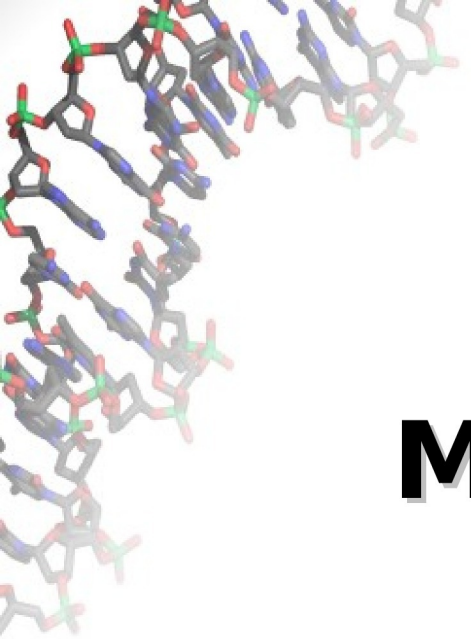


Часть *4.*

МНОЖЕСТВЕННОЕ ВЫРАВНЕНИЕ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

.

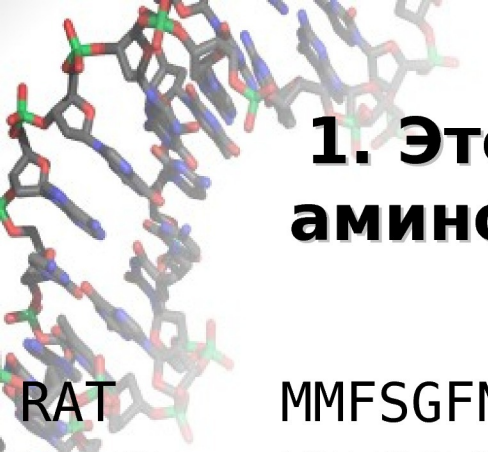
Вариации на тему парного сравнения последовательностей
Итоги

- 1. Эталонные выравнивания и качество выравниваний. Подбор параметров.
- 2. Попытка обойтись без штрафов за делеции. Парето-оптимальные выравнивания и выбор «правтльного» выравнивания в множестве Парето-оптимальных выравниваний.
- 3. Поиск выравнивания в полосе – заранее заданной или подбираемой по ходу. «Оптимистическое» выравнивание и алгоритм A*.
- 4. Выравнивание геномных (сверх-длинных) последовательностей. Иерархическое выравнивание.
- 5. Поиск локальных сходств. Затравки. Чувствительность и избирательность. Множественные разреженные затравки.
-

План занятия

- **1. Что такое "множественное выравнивание"?**
- **2. Как интерпретировать множественное выравнивание?**
- **3. Как строить множественное выравнивание?**

  Постановки задачи: 1) оптимизация веса выравнивания и 2) поэтапный подход.
- **4. Вес столбца. Вес выравнивания.**
- **5. Динамическое программирование.**
- **6. Поэтапный подход. "Путеводное" дерево. Филогения. PSWM.**
- *7. Локальные множественные сходства*
- *8. Иерархическое множественное выравнивание*



1. Это - множественное выравнивание аминокислотных последовательностей

```
RAT       MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQD-
MOUSE     MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDF
CHICK     MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDF
B_MOUSE   -MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-
B_HUMAN   -MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-
           *:..* .:*:: .***** **:.:*...* *..***.*  :.. :*:.
```

Вставляем ‘-’
в разные последовательности,
чтобы все они стали одной длины

1. Это - множественное выравнивание аминокислотных последовательностей

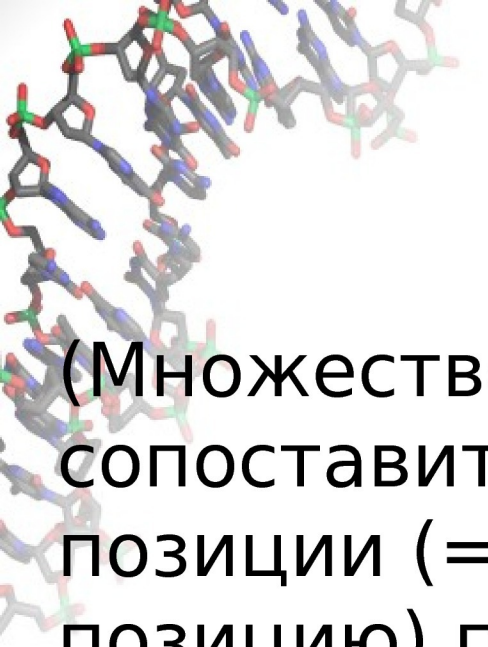
```
RAT       MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQD-
MOUSE     MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDF
CHICK     MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDF
B_MOUSE   -MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-
B_HUMAN   -MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-
           *:..* .:*:: .***** **:.:*...* *..***.*  :.. :*:.
```

* - все буквы в столбце одинаковые

: - есть одна «посторонняя» буква

Эволюция и множественное выравнивание

- **Предок:** ASVVLDFTGT
- **Потомки:**
  - ASVVLDFTGT AS-VVLDFTGT ASVVLDFTGT
    ATVVI--TGS GSMVVLEFSGT GSVVLEFTPT
  - AS-VVLDFTGT
  - AT-VVI--TGS
  - GSMVVLEFSGT
  - GS-LVLDFTPT
- **Консенсус:** *GS-VVL*FTGT*



Выравнивания и эволюция

(Множественное) выравнивание – попытка сопоставить друг другу **гомологичные** позиции (=имеющие общую предковую позицию) путем *удалений* и *вставок*

Цель: восстановить предковую последовательность («консенсус»)

Недостаток: не учитываем возможность перестановок.

Вес множественного выравнивания – сумма весов столбцов

- Обычно считают, что колонки в выравнивании независимы (что неверно!)
- Поэтому качество выравнивания оценивают как сумму качеств колонок.

```
-ИВАН---ОВ----
-ИВАН--ЦОВ----
 -ИВАН--КОВ----
 -ИВАНЧУКОВ----
ДИВАН---ОВ----
ДИВАНЧИКОВ----
 -ИВАН---ОВСКИЙ
 --ВАНЬ-КОВ----
```

?? Что делать с делециями?
Как оценивать вес столбца?

Вес множественного выравнивания

Учет делеций:

1) пробел – равноправный символ

2) делеции учитываем отдельно

$$S = G + \sum_{\text{columns}} S(m_k)$$

G – веса делеций,

$S(m_k)$ – вес колонки без учета делеций

Недостатки:

1) – только посимвольное удаление

2) – алгоритмические трудности при использовании ДП (преодолеваются при поэтапном подходе)

Как учитывать пофрагментные операции?

```
*++++***++****
-ИВАН---ОВ----
-ИВАН--ЦОВ---- вставка «Ц»
-ИВАН--КОВ---- вставка «К»
-ИВАНЧУКОВ---- вставка «ЧУК»
ДИВАН---ОВ---- вставка «Д»
ДИВАНЧИКОВ---- вставка «Д»,«ЧИК»
-ИВАН---ОВСКИЙ вставка «СКИЙ»
--ВАН-ЬКОВ---- удаление 1 симв.
                вставка «ЬК»
```

Вариант: выделяем «значимые» столбцы; все остальное учитываем отдельно в каждой последовательности.

Б. Сумма весов ВЫДЕЛЕННЫХ столбцов МИНУС штрафы за то, что между ДРУГИЕ ВАЖНЫЕ СТОЛБЦЫ

```
*++++***++****
-ИВАН---ОВ----  уд.1 симв.
-ИВАН--ЦОВ----
-ИВАН--КОВ----
-ИВАНЧУКОВ----  вст. «ЧУ»
ДИВАН---ОВ----  вст.«Д»; уд. 1 симв.
ДИВАНЧИКОВ----  вст «Д», «ЧИ»
-ИВАН---ОВСКИЙ  вст«СКИЙ»;уд.1 симв.
--ВАН-ЬКОВ----  уд. 1 симв.; вст.»Ь»
```

Выделяем «значимые» столбцы. Всё остальное учитываем отдельно в каждой последовательности.

Вариант Б позволяет использовать пофрагментные веса вставок и удалений

Вес множественного выравнивания

Сумма весов ВЫДЕЛЕННЫХ столбцов
МИНУС штрафы за вставки между столбцами
МИНУС штрафы за удаленные фрагменты в выделенных столбцах

!!! Каждый удаленный символ может отдельно учитываться еще и в весе столбца.

На практике:
– при оптимизации ВЕС ВЫРАВНИВАНИЯ – это сумма ВЕСОВ ВСЕХ столбцов [из-за алгоритмических трудностей];
- более сложные веса используются при уточнении построенного выравнивания

Вес множественного выравнивания – сумма весов столбцов

Вес столбца:

1) сумма попарных весов по расширенной матрице замен;
2) энтропия

Вес столбца множественного выравнивания

Сумма пар

$S(a_i, b)$ – вес сопоставления символов a, b (один из них может быть пробелом!)

$$S(m_i) = \sum_{k<l} s(x^k_i, x^l_i);$$

- Способ не совсем правильный. Более правильная оценка, например, для трех последовательностей:

$S(m_i) = \log (p_{abc} / q_a q_b q_c)$, а не

$\log(p_{ab}/q_a q_b) + \log(p_{bc}/q_b q_c) + \log(p_{ac}/q_a q_c)$;

(веса строятся по наиб. правдоподобию)

Качество множественного выравнивания
Энтропийная оценка

- Пусть c(i, a) – количество появлений a в колонке i.
- *(с поправкой на псевдоотсчеты,* можно учитывать априорные вероятности букв*)*
- $p_{i,a}$ - вероятность буквы a в колонке i:

$$p_{i,a} = c(i, a) / \sum_a c(i, a)$$

-ИВАН---ОВ----
-ИВАН--ЦОВ----
-ИВАН--КОВ----
-ИВАНЧУКОВ----
ДИВАН---ОВ----
ДИВАНЧИКОВ----
-ИВАН---ОВСКИЙ
--ВАНЬ-КОВ----

Качество множественного выравнивания
Энтропийная оценка

- Пусть c(i, a) – количество появлений a в колонке i.
- $p_{i,a}$ - вероятность буквы a в колонке i:

$$p_{i,a} = c(i, a) / \sum_a c(i, a)$$

Вероятность *i-й* колонки: $P(m_i) = \prod_a p_{ia}^{c(i,a)}$

- Вероятность выравнивания = $\prod_i P(m_i)$.
- Вес выравнивания: [минус] логарифм вероятности:

$$S = \sum_{columns} S(m_k);$$

$$S(m_k) = -\sum_a c_{ia} \log p_{ia} = (-\sum_a p_{ia} \log p_{ia})*N = H(m_i)*N$$

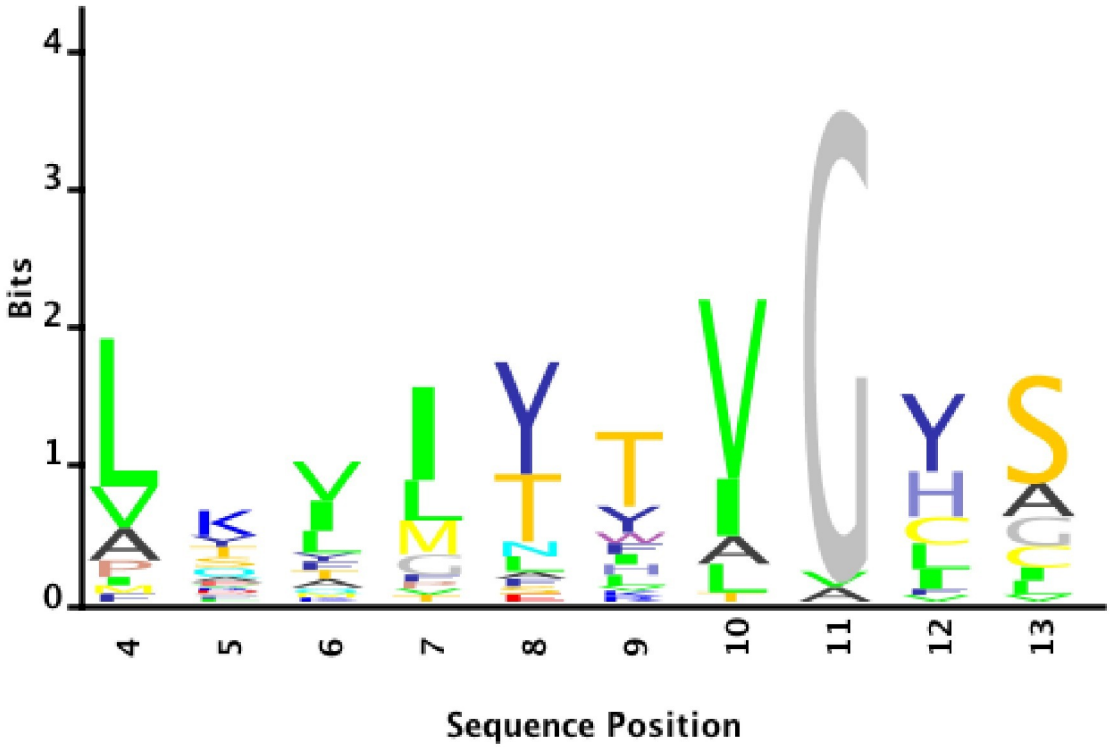
$H(m_i) = (-\sum_a p_{ia} \log p_{ia})$ – энтропия колонки;

$N = \sum_a c(i, a)$ – *количество последовательностей (с учетом псевдоотсчетов)*

Logo – представление

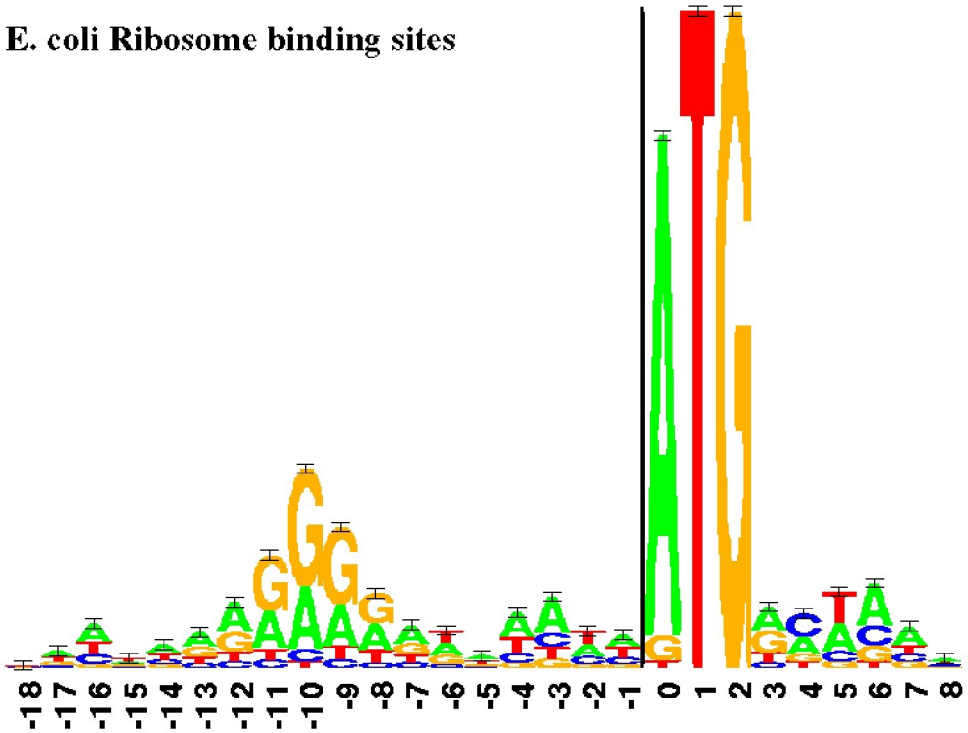
$$H_{col} = \log_2 N - \left(- \sum_{n=1}^{N} p_n \log_2 p_n \right)$$

$$h_a \sim p_a$$



Logo для сайта связывания рибосомы

- http://www.ccrnp.ncifcrf.gov/~toms/sequencelogo.html



- Free logo ey.edu/

Множественное выравнивание: как интерпретировать?

```
-ИВАН---ОВ----
-ИВАН--ЦОВ----
-ИВАН--КОВ----
-ИВАНЧУКОВ----
ДИВАН---ОВ----
ДИВАНЧИКОВ----
-ИВАН---ОВСКИЙ
--ВАНЬ-КОВ----
--ВАН- КРОЙФФ-
```

Консервативные и почти консервативные позиции

-ИВАН---ОВ----
-ИВАН--ЦОВ----
-ИВАН--КОВ----
-ИВАНЧУКОВ----
ДИВАН---ОВ----
ДИВАНЧИКОВ----
-ИВАН---ОВСКИЙ
--ВАНЬ-КОВ----
--ВАН- КРОЙФФ-

ВАН - консервативные позиции

ОВ – почти консервативные позиции

И – почти консервативная позиция

Данные мешают: посторонние

-ИВАН---ОВ----
-ИВАН--ЦОВ----
-ИВАН--КОВ----
-ИВАНЧУКОВ----
ДИВАН---ОВ----
ДИВАНЧИКОВ----
-ИВАН---ОВСКИЙ
--ВАНЬ-КОВ----
--*ВАН--КРОЙФФ-*

ВАН КРОЙФФ -
посторонний !!!

Консервативные и почти консервативные позиции

-ИВАН---ОВ----
-ИВАН--ЦОВ----
-ИВАН--КОВ----
-ИВАНЧУКОВ----
ДИВАН---ОВ----
ДИВАНЧИКОВ----
-ИВАН---ОВСКИЙ
--ВАНЬ-КОВ----

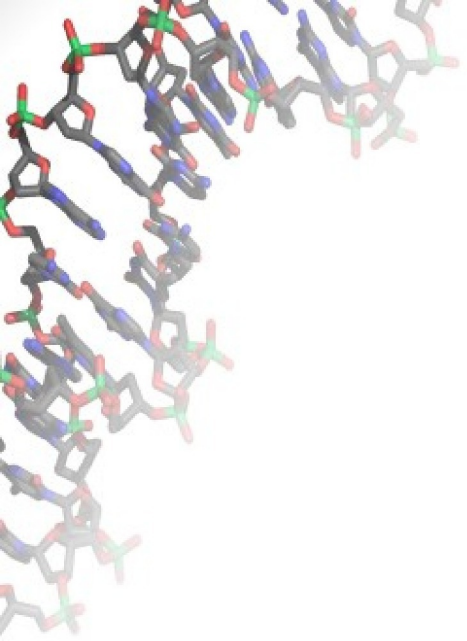
ВАН ОВ – консервативные позиции

И – почти консервативная позиция

Данные помогают:
НОВЫЕ ДАННЫЕ МЕНЯЮТ ВЫРАВНИВАНИЕ

```
ИВАН--КО          или    ИВА--НКО
ИВАНЕНКО                  ИВАНЕНКО
ИВАЩЕНКО                  ИВАЩЕНКО
```

```
ИВАН--КО
ИВАНЕНКО
ИВАЩЕНКО
ИВАНОВ
```



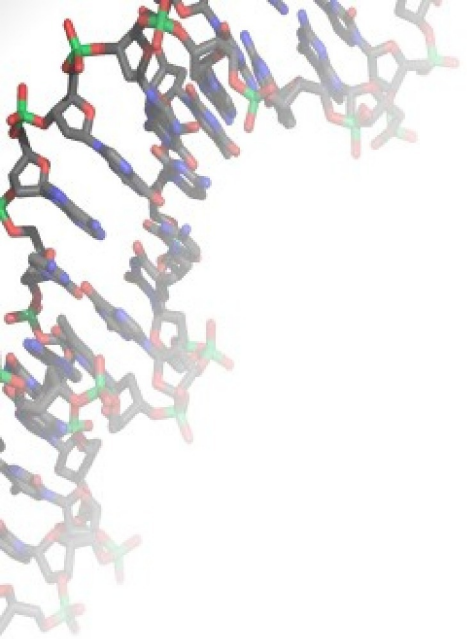
Редукция алфавита: Гласные, Согласные.

-ГСГС--сГС----

-ИВАН---ОВ----
-ИВАН--ЦОВ----
-ИВАН--КОВ----
-ИВАНЧУКОВ----
ДИВАН---ОВ----
ДИВАНЧИКОВ----
-ИВАН---ОВСКИЙ
--ВАНЬ-КОВ----

Перед ОВ – согласная

(или чередование «к-ц»)



Консенсус

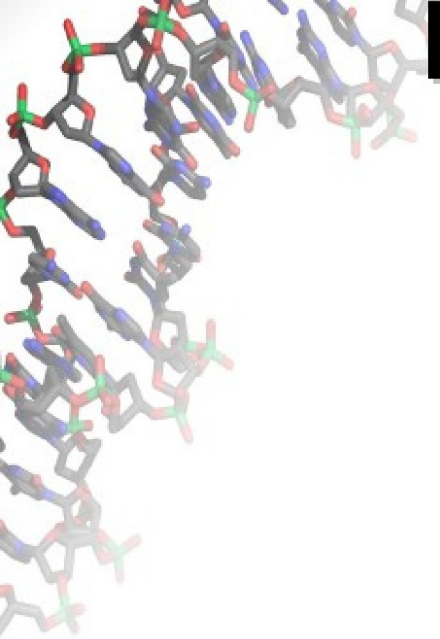
-ГСГС---ГС----
-ИВАН---ОВ----
-ИВАН--ЦОВ----
-ИВАН--КОВ----
-ИВАНЧУКОВ----
ДИВАН---ОВ----
ДИВАНЧИКОВ----
-ИВАН---ОВСКИЙ
--ВАН-ЬКОВ----

-ИВАН--сОВ-----

Убираем пробелы:

ИВАНсОВ

«с» - «согласная» или «к/ц»



Консенсус: ИВАНсОВ

```
-ГСГС---ГС----
-ИВАН---ОВ----
-ИВАН--ЦОВ----
-ИВАН--КОВ----
-ИВАНЧУКОВ----
ДИВАН---ОВ----
ДИВАНЧИКОВ----
-ИВАН---ОВСКИЙ
--ВАНЬ-КОВ----
```

Чего не поняли?

Анализ выравниваний. Выводы.

1. Выявить и убрать посторонних. Методика: граф попарных расстояний (с использованием выравнивания или без использования). Кластерные методы.

Проблемы: неоднородность по последовательности (в начале – посторонний, в конце – нет). См. п.3.

2. Редукция алфавита (возможно, своя в каждом столбце). Алфавит(ы) редукции может быть задан априорно, а может определяться по выравниванию.

3. Выделение достоверных блоков.

4. *Анализ выравниваний – неформальный процесс.*

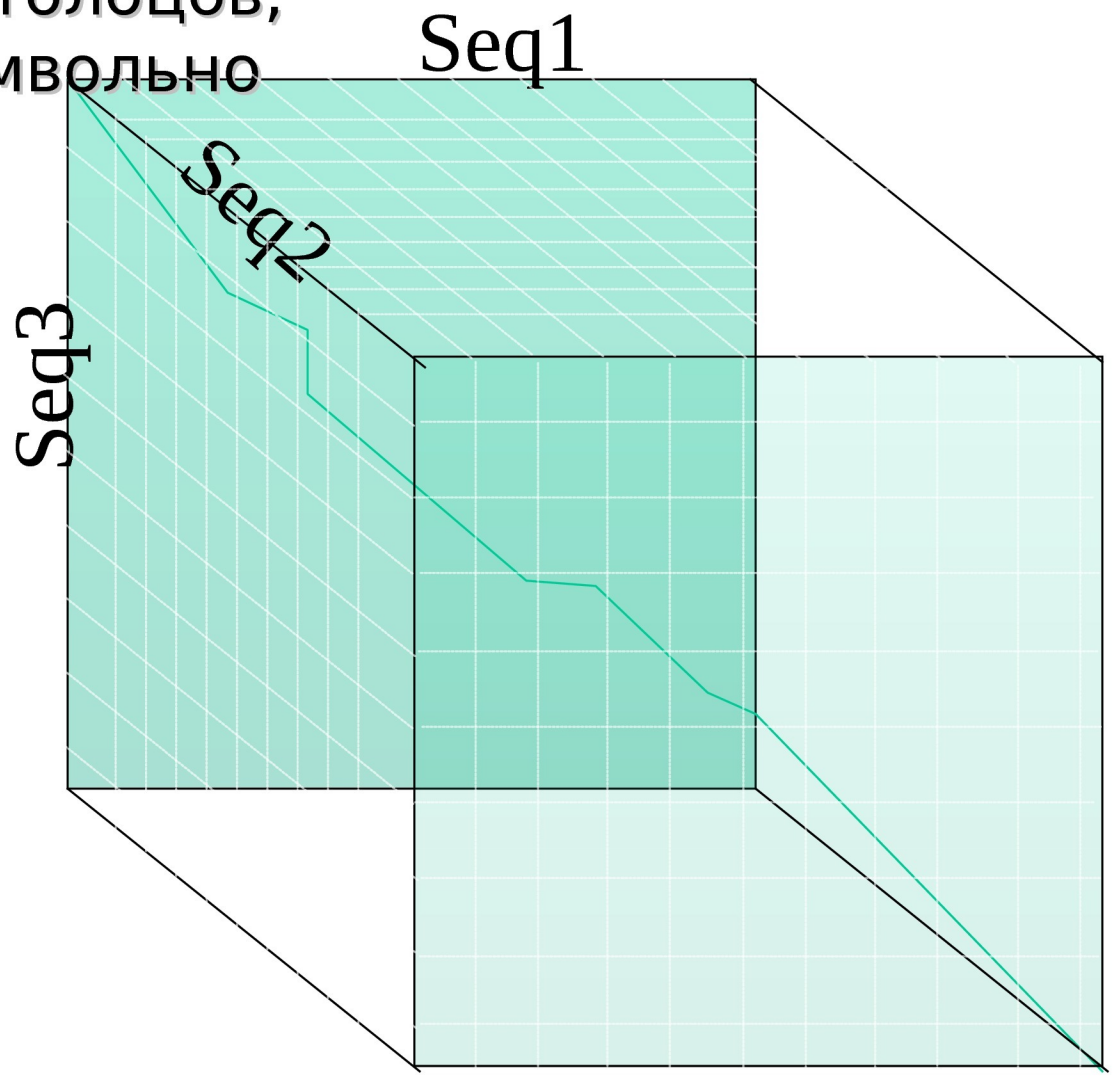
4. Как строить множественное выравнивание?

- 1) Максимизируем вес
  : Динамическое программирование;
  вес – сумма весов всех столбцов
- 2) «Поэтапно»
  : сводим к серии парных выравниваний

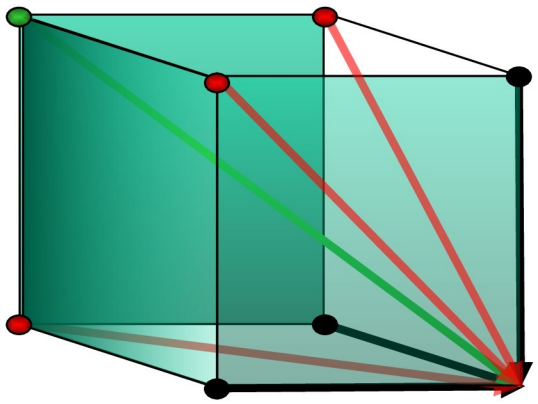
ДИНАМИЧЕСКОЕ ПРОГРАММИРОВАНИЕ

Неявно предполагается:

1) независимость столбцов;
2) удаления – посимвольно

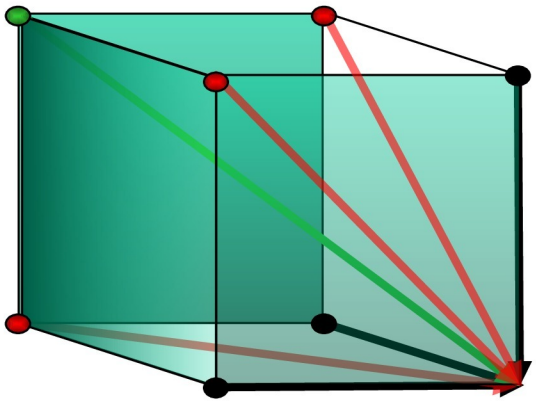


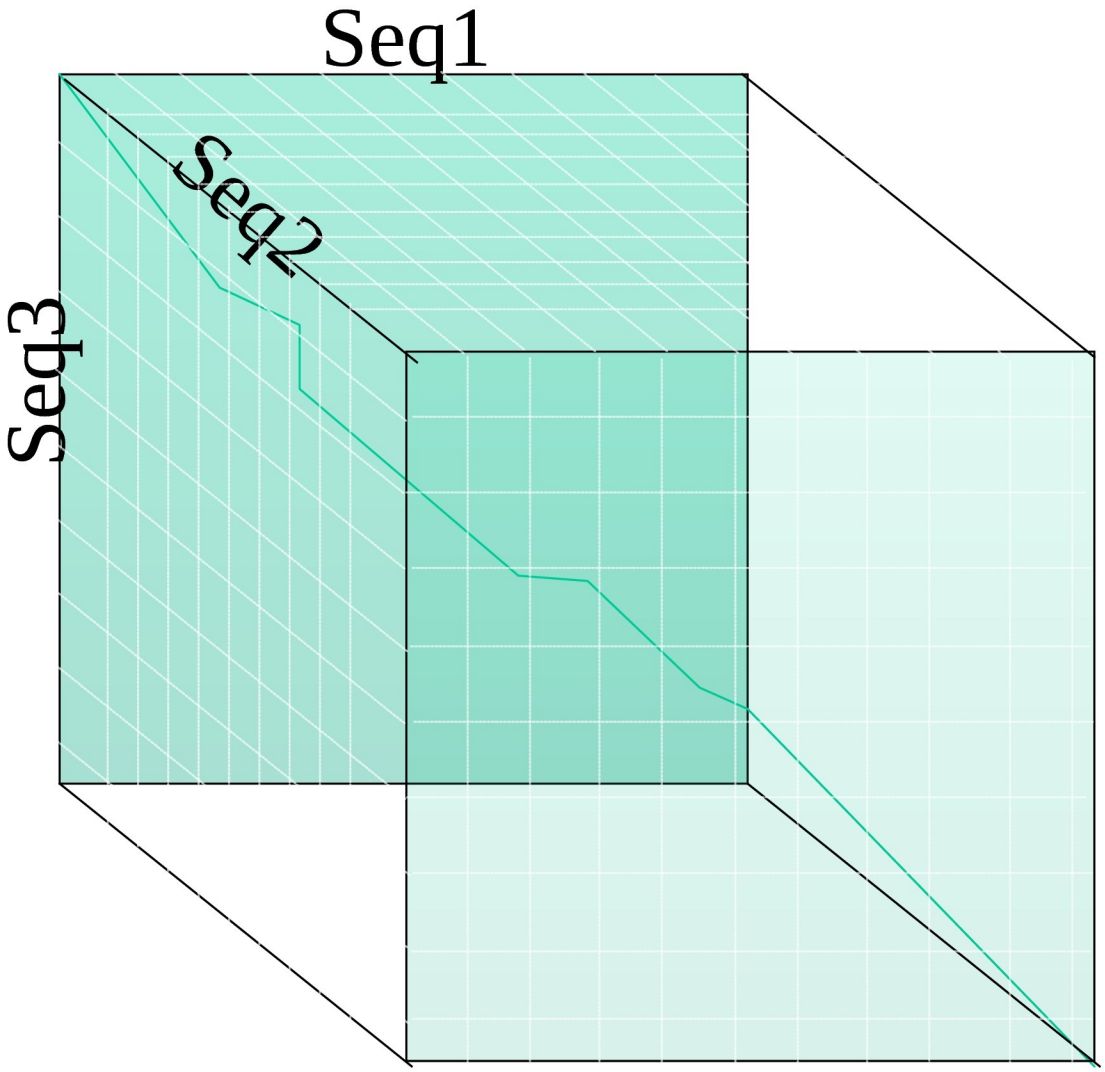
- ***Элементарные переходы:***



4.1. ДИНАМИЧЕСКОЕ ПРОГРАММИРОВАНИЕ

- Элементарные переходы:
  - ***Сопоставление трех***
  - ***Сопоставление двух и одна делеция***
  - ***Делеция в двух последователь-ностях***





ДИНАМИЧЕСКОЕ ПРОГРАММИРОВАНИЕ

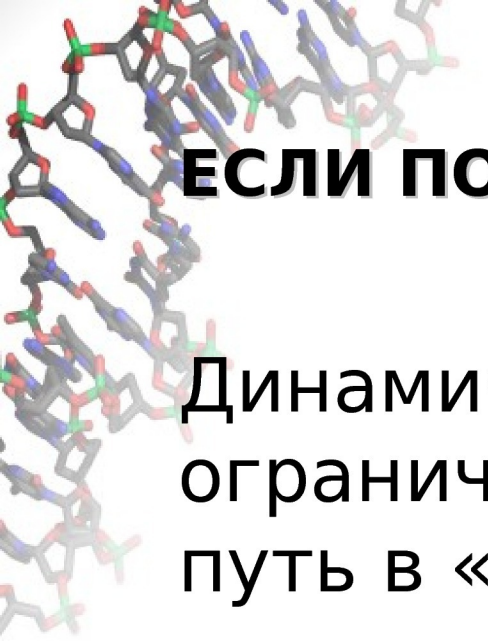
- **Количество вершин равно**

$$\prod_{посл.} L_i = O(L^N)$$

- **Количество ребер из каждой вершины = 2^N-1**
- **Количество операций равно**

$$T = O(L^N)$$

- **Надо запоминать обратные переходы в L^N вершинах.**
- **Если количество последовательностей > 4, то задача практически не разрешима.**



ЕСЛИ ПОСЛЕДОВАТЕЛЬНОСТИ ПОХОЖИ…

Динамическое программирование с ограничениями: строим оптимальный путь в «цилиндре» радиуса *d*

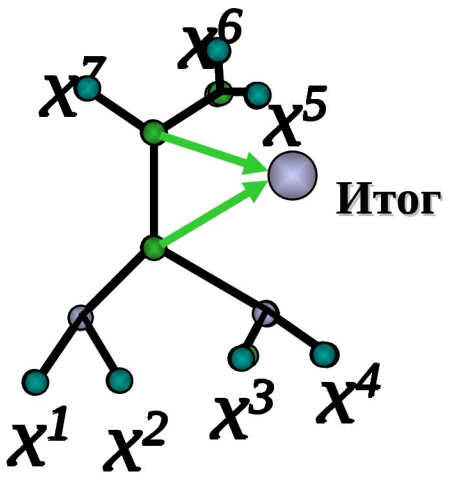
$$T \sim L \bullet d^{N-1}$$

!!! Величину *d* можно подбирать адаптивно (аналогия: алгоритм A*)

ДП применяется для уточнения выравнивания заведомо близких последовательностей

4.2 Поэтапное выравнивание (progressive alignment)

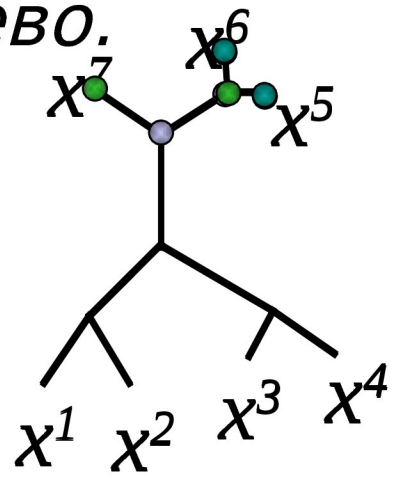
- Строится бинарное дерево (guide tree, путеводное дерево); листья – последовательности
- Дерево обходится начиная с листьев. При объединении двух узлов строится ***парное*** выравнивание суперпоследовательностей (профилей) и получается новая суперпоследовательность



Путеводное дерево строится приближенно – главное быстро. Обычно это кластерное дерево

Путеводные деревья

Термин: *филогенетическое дерево.*



Строится на основе матрицы попарных расстояний.

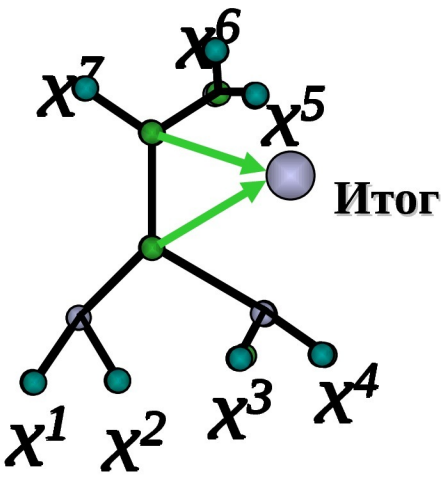
Методов – много (пример – Neighboour Joining).

Основное время – на построение матрицы попарных расстояний.

Идея: Сначала путеводное дерево строится приближенно (главное: быстро). Пример метрики: расстояние между векторами частот букв. Потом перевычисляем по построенным выравниваниям.

4.2 Поэтапное выравнивание (progressive alignment)

- Строится бинарное дерево (guide tree, путеводное дерево);
- листья –последовательности
- для внутренних узлов, двигаясь от листьев, строим множественные выравнивания посл-стей в листьях, для которых узел является предком



.

Рекурсия

- выбираем два узла, такие, что им приписаны последовательности, а их предку – нет;

- выравниваем посл-сти в выбранных узлах;

- приписываем предку профиль, построенный по полученному выравниванию.

Как выравнивать выравнивания?

Что такое «суперпоследовательность»? Вариант 1: «стопка выровненных последовательностей»

- Выравнивание двух стопок - ДП.
- Оптимизируется сумма парных весов:

$$\sum_i S(m_i) \rightarrow \max$$

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

- Если мы выравниваем две стопки – $0 < i \le n$ и $n < i \le N$, то сумму разбиваем на три части:

$$S(m_i) = \sum_{k< l\le n} s(x^k_i, x^l_i) + \sum_{n< k< l\le N} s(x^k_i, x^l_i) + \sum_{k\le n,\ n< l\le N} s(x^k_i, x^l_i)$$

- Две первые суммы являются внутренним делом стопок, последняя сумма отвечает за сравнение стопок (профилей)
- При сравнении используем расширенную матрицу сходства, добавив в нее сравнение символа удаления '-' :

$$s(-,-)=0,\ s(a,-) = -d\ ;$$

Что такое «суперпоследовательность»? Вариант 2: «профиль»

- Он же:

  PSSM =

  = Position Specific Scoring Matrix

- Он же:

  PSWM =

  = Position Specific Weight Matrix

- Он же:

  PWM = Position Scoring Matrix

Матрица частот
(с псевдоотсчетами, смесями и т.п.)

Пусть

- $p(a, i)$ – доля буквы a в i-м столбце выравнивания;
- $f(a)$ - априорная частота буквы a.

Тогда полагаем

$$m(a, i) = \beta * p(a, i) + (1-\beta) * f(a)$$

β – «коэффициент доверия»;

$m(a, i)$ – апостериорная вероятность буквы a в i-м столбце выравнивания

Матрица частот заполняется апостериорными вероятностями

Матрица частот
(с псевдоотсчетами, смесями и т.п.)

	1	2	3	4	5	6	7	8	9	10
A	0.125	0.062	0.062	0.000	0.031	0.000	0.094	0.031	0.000	0.156
R	0.000	0.031	0.000	0.000	0.000	0.031	0.000	0.000	0.000	0.000
N	0.000	0.000	0.000	0.000	0.062	0.000	0.000	0.000	0.000	0.000
D	0.000	0.031	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.125	0.094
Q	0.000	0.094	0.031	0.000	0.000	0.000	0.000	0.000	0.000	0.000
E	0.000	0.031	0.000	0.000	0.031	0.000	0.000	0.000	0.000	0.000
G	0.000	0.031	0.000	0.094	0.000	0.000	0.000	0.938	0.000	0.125
H	0.000	0.000	0.000	0.000	0.000	0.062	0.000	0.000	0.219	0.000
I	0.031	0.000	0.219	0.438	0.000	0.062	0.188	0.000	0.094	0.062
L	0.562	0.031	0.156	0.188	0.062	0.062	0.094	0.000	0.125	0.062
K	0.000	0.281	0.031	0.000	0.000	0.031	0.000	0.000	0.000	0.000
M	0.031	0.000	0.031	0.156	0.000	0.000	0.000	0.000	0.000	0.000
F	0.031	0.031	0.062	0.031	0.031	0.062	0.000	0.000	0.031	0.000
P	0.062	0.000	0.000	0.031	0.000	0.000	0.000	0.000	0.000	0.000
S	0.000	0.094	0.000	0.000	0.031	0.000	0.000	0.000	0.000	0.469
T	0.000	0.125	0.062	0.031	0.281	0.438	0.031	0.000	0.000	0.000
W	0.000	0.000	0.000	0.000	0.000	0.062	0.000	0.000	0.000	0.000
Y	0.000	0.125	0.062	0.000	0.469	0.156	0.000	0.000	0.375	0.000
V	0.156	0.031	0.281	0.031	0.000	0.031	0.594	0.031	0.031	0.031

Переход от частот к PSSM

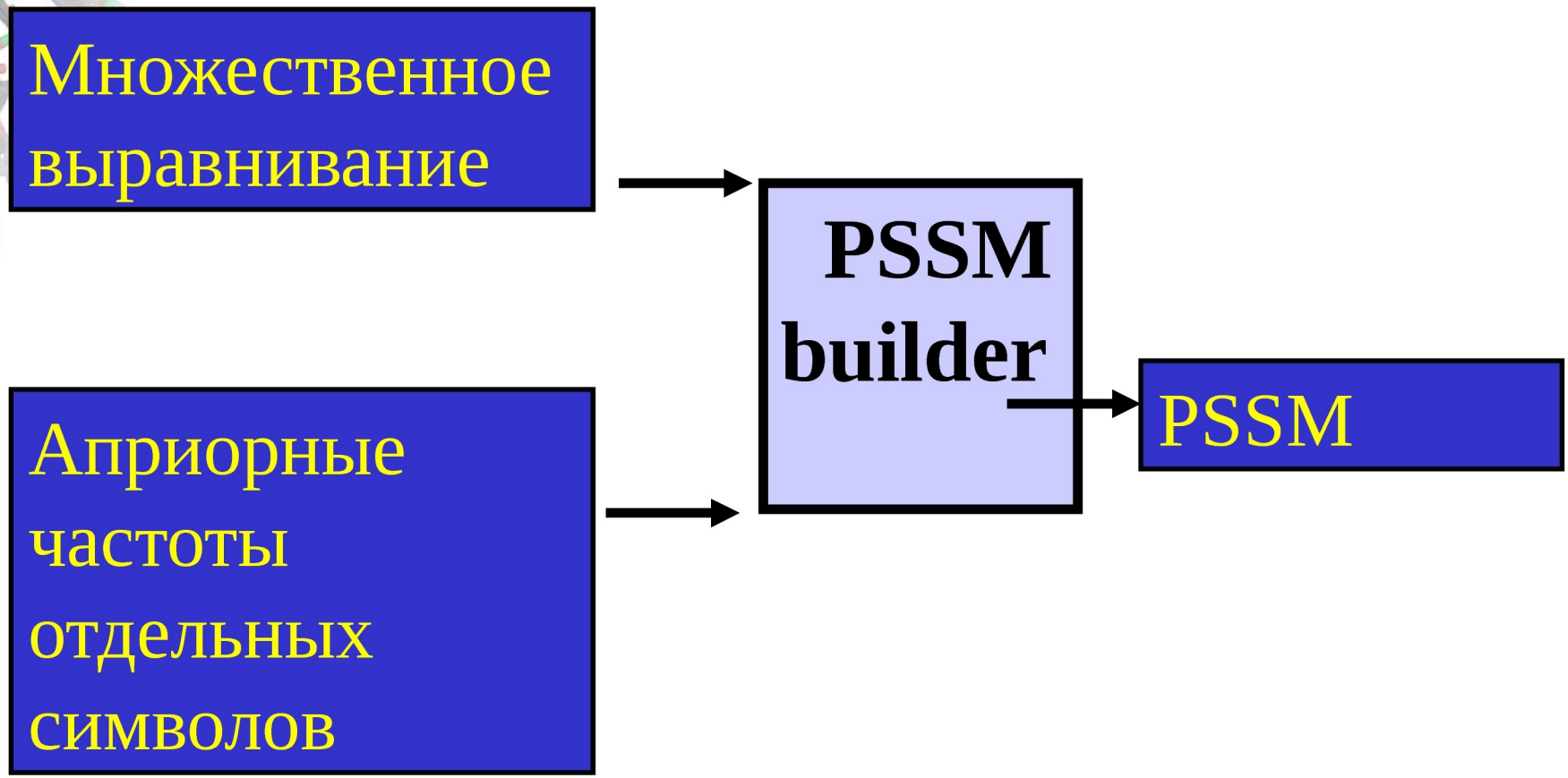
- Log-отношение правдоподобия:

$$score(j,i) = \log m(j,i) / f(j)$$

  - *m(j,i)* – частота символа *j* в позиции *i;*
  - *f(j)* – частота символа *j* в корпусе

  - Таким образом, вместо «универсальных» весов символов получаем вес, привязанный к месту.

Построение PSSM



Вес сопоставления позиции весового профиля и символа

- **N** – длина профиля; **A** – размер алфавита

F[NxA] - профиль

Вес сопоставления символа *a* с *j*-м столбцом профиля:

F[*j, a*]

Вес сопоставления весового профиля и позиции множественного выравнивания

- **N** – длина профиля; **A** – размер алфавита

F[NxA] – профиль

$P_k(a_i)$ – доля буквы a_i в k-м столбце выравнивания

Вес сопоставления j-го столбца профиля с k-м столбцом выравнивания:

$$\sum_i (P_k(a_i) * \mathbf{F}[j, a_i])$$

Вес сопоставления позиции множественного выравнивания и позиции частотного профиля

- 1. Строим по каждой «стопке» = (Мн.выр-ние) частотную матрицу.
- 2. Строим по одной из «стопок» PSSM.
- 3. Выравниваем PSSM с частотной матрицей (вес сопоставления – взвешенная сумма элементов столбца PSSM).
- === Асимметрия:

  а) игнорируем;

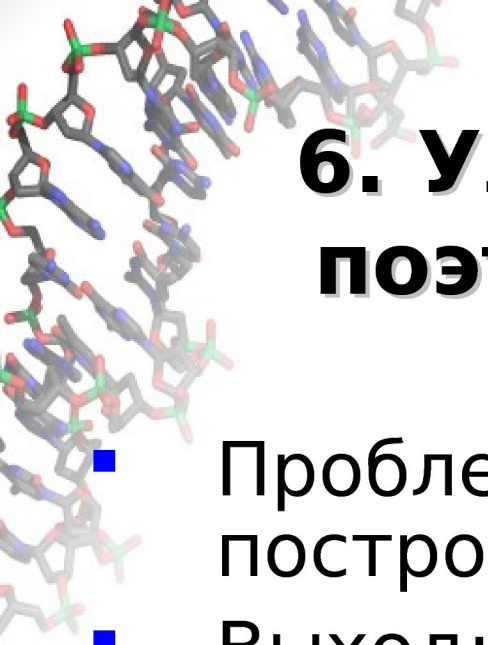
  б) делаем дважды и согласовываем.

Пример: ClustalW

1. Строится матрица расстояний с использованием попарных выравниваний
2. Строится NJ дерево (метод ближайшего соседа)
3. Строится поэтапное выравнивание

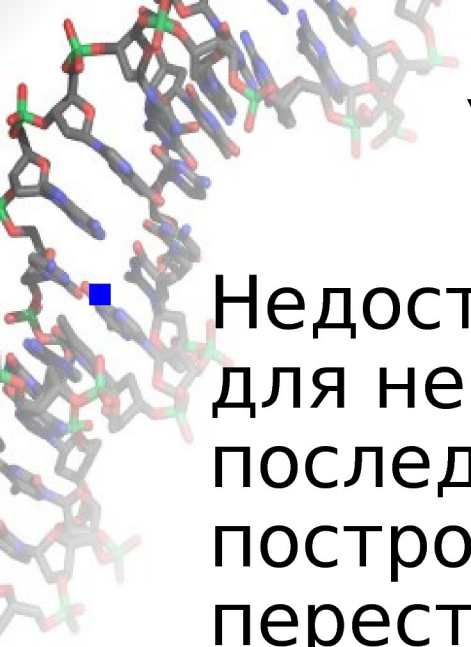
Пример: ClustalW

1. поэтапное выравнивание
- Дополнительные эвристики при поэтапном выравнивании
  - Взвешивание последовательностей (с учетом только топологии дерева)
  - На разных уровнях дерева используются разные матрицы сходства
  - Используется контекстно-зависимые штрафы за открытие делеции
  - Если при построении выравнивания появляются очень низкие веса, то дерево корректируется



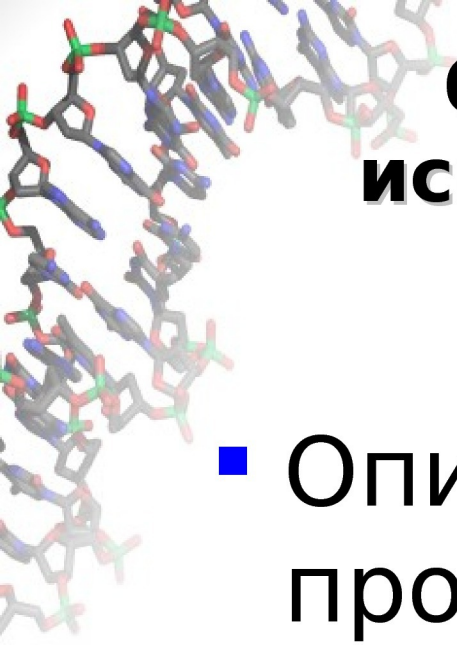
6. Улучшение результатов поэтапного выравнивания

- Проблема: результат зависит от построенного дерева
- Выход: пересчет расстояний по построенному выравниванию и повторение процедуры



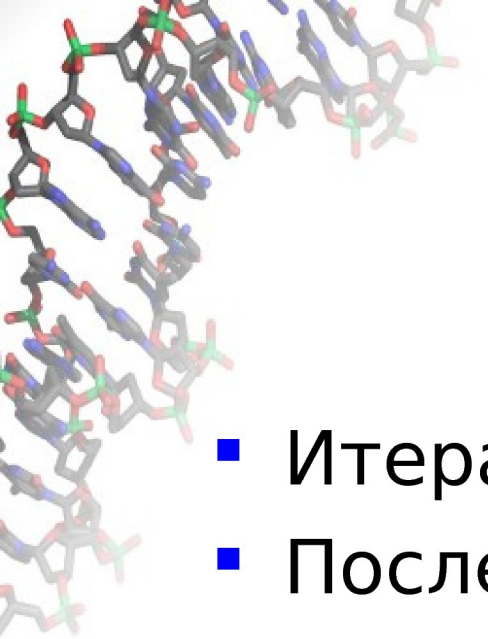
Улучшение выравнивания

- Недостаток поэтапных методов: если для некоторой группы последовательностей выравнивание построено, то оно уже не перестраивается.
- Алгоритм итеративного улучшения
  1. Вынимаем из выравнивания одну последовательность
  2. По оставшимся последовательностям строим профиль
  3. Выравниваем вынутую последовательность с профилем
  4. Переходим к этапу 1.



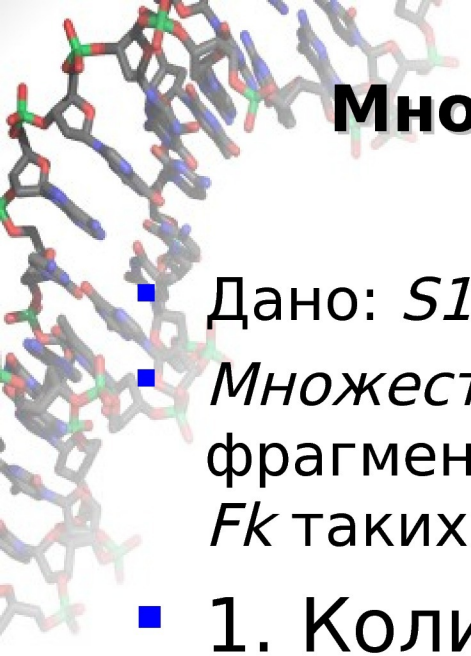
6. Отступление в сторону: использование профилей для поиска мотивов

- Описываем семейство слов профилем
- Ищем вхождения (неточные!) профиля в текст



PSI-BLAST

- Итеративный поиск в базе текстов.
- После каждого этапа уточняем PSSM.
- При построении новой PSSM используется множественное выравнивание с предыдущей PSSM в качестве «мастер-последовательности»



Множественные локальные сходства. Уточнение.

- Дано: *S1, …, Sn*
- *Множественное локальное сходство* – это набор фрагментов этих последовательностей *F1, …, Fk* таких, что фрагменты *похожи.*
- 1. Количество фрагментов в тексте:

  - ровно 1 в каждом,

  - не менее 1 в каждом,

    - сколько получится (= неточные повторы в объединенном тексте)