

Множественное выравнивание.

Алгоритмы в биоинформатике

Дмитрий Мелешко
meleshko.dmitrii@gmail.com

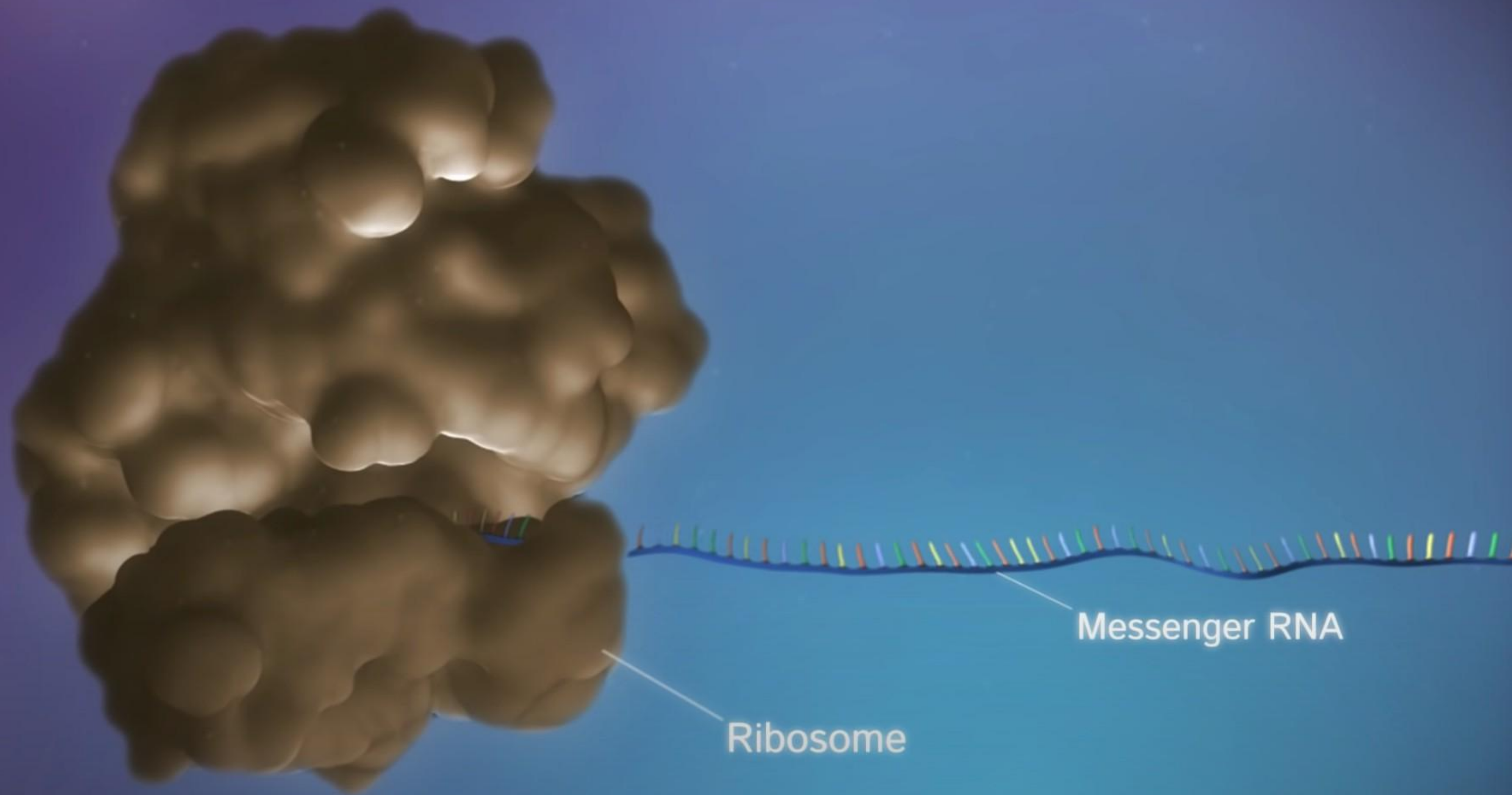
Что было на прошлой лекции?

- CpG островки как результат метилирования
- Марковская цепь для генерации последовательности из CpG островка
- НММ для генерации размеченной последовательности
- Алгоритм Витерби для восстановления последовательности скрытых состояний НММ
- Алгоритм прохода вперед-назад для восстановления распределения вероятности по скрытым состояниям на каждом шаге генерации НММ

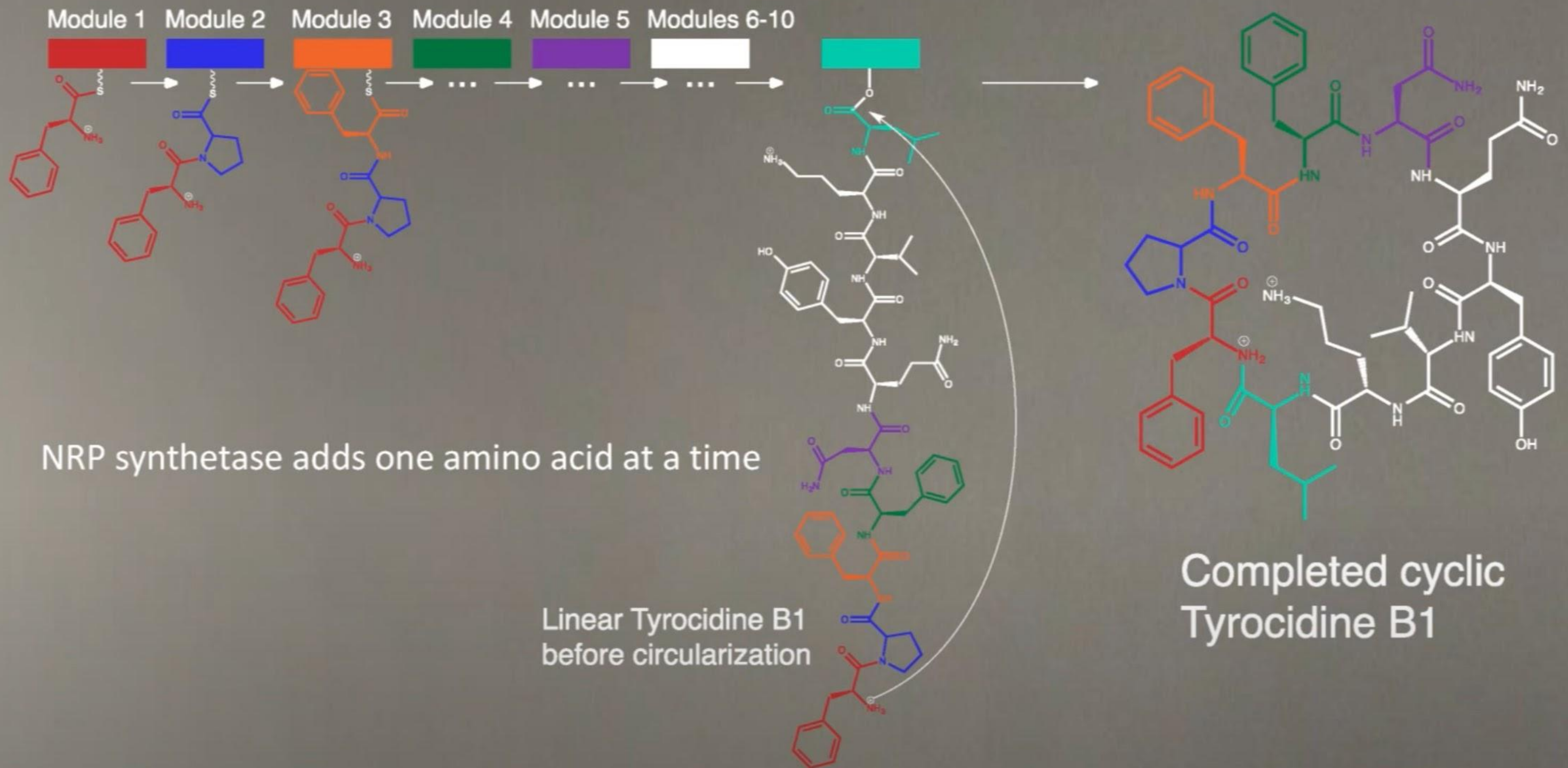
Что будет на этой лекции?

- Выравнивание многих последовательностей как средство поиска устойчивых и меняющихся участков генома.
- Определение множественного выравнивания и его профиля.
- Многомерная динамика и жадный алгоритм поиска множественного выравнивания, ClustalW.

Нерибосомные пептиды



Нерибосомные пептиды



Нерибосомные пептиды

YAFDLGYTCMFPVLLGGGELHIVQKETYTAPDEIAHYIKEHGITYIKLTPSLFHTIVNTASFAFDANFESLRLIVLGGEKIIPIDVIAFRKMYGHTEFINHYGPTEATIGA
AFDVSAGDFARALLTGGQLIVCPNEVKMDPASLYAIIKKYDITIFEATPALVIPLMEYIYEQKLDISQLQILIVGSDSCSMEDFKTLVSRFGSTIRIVNSYGVTEACIDS
IAFDASSWEIYAPLLNGGTVVCIDYYTTIDIKALEAVFKQHHRGAMLPPALLKQCLVSAPTMISSLEILFAAGDRLSSQDAILARRAVGSGVYNAYGPTENTVLS

Нерибосомные пептиды

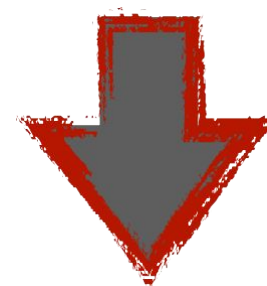
YAFDLGYTCMFPVLLGGGELHIVQKETYTAPDEIAHYIKEHGITYIKLTPSLFHTIVNTASFAFDANFESLRLIVLGGEKIIPIDVIAFRKMYGHTEFINHYGPTEATIGA
AFDVSAGDFARALLTGGQLIVCPNEVKMDPASLYAIIKKYDITIFEATPALVIPLMEYIYEQKLDISQLQILIVGSDSCSMEDFKTLVSRFGSTIRIVNSYGVTEACIDS
IAFDASSWEIYAPLLNGGTVCIDYYTTIDIKALEAVFKQHHRGAMLPPALLKQCLVSAPTMISSLEILFAAGDRLSSQDAILARRAVGSGVYNAYGPTENTVLS



YAFDLGYTCMFPVLLGGGELHIVQKETYTAPDEIAHYIKEHGITYIKLTPSLFHTIVNTASFAFDANFESLRLIVLGGEKIIPIDVIAFRKMYGHTEFINHYGPTEATIGA
AFDVSAGDFARALLTGGQLIVCPNEVKMDPASLYAIIKKYDITIFEATPALVIPLMEYIYEQKLDISQLQILIVGSDSCSMEDFKTLVSRFGSTIRIVNSYGVTEACIDS
IAFDASSWEIYAPLLNGGTVCIDYYTTIDIKALEAVFKQHHRGAMLPPALLKQCLVSAPTMISSLEILFAAGDRLSSQDAILARRAVGSGVYNAYGPTENTVLS

Нерибосомные пептиды

YAFDLGYTCMFPVLLGGGELHIVQKETYTAPDEIAHYIKEHGITYIKLTPSLFHTIVNTASFAFDANFESLRLIVLGGEKIIPIDVIAFRKMYGHTEFINHYGPTEATIGA
AFDVSAGDFARALLTGGQLIVCPNEVKMDPASLYAIIKKYDITIFEATPALVIPLMEYIYEQKLDISQLQILIVGSDSCSMEDFKTLVSRFGSTIRIVNSYGVTEACIDS
IAFDASSWEIYAPLLNGGTVCIDYYTTIDIKALEAVFKQHHRGAMLPPALLKQCLVSAPTMISSLEILFAAGDRLSSQDAILARRAVGSGVYNAYGPTENTVLS



YAFDLGYTCMFPVLLGGGELHIVQKETYTAPDEIAHYIKEHGITYIKLTPSLFHTIVNTASFAFDANFESLRLIVLGGEKIIPIDVIAFRKMYGHTE-FINHYGPTEATIGA
-AFDVSAGDFARALLTGGQLIVCPNEVKMDPASLYAIIKKYDITIFEATPALVIPLMEYI-YEQKLDISQLQILIVGSDSCSMEDFKTLVSRFGSTIRIVNSYGVTEACIDS
IAFDASSWEIYAPLLNGGTVCIDYYTTIDIKALEAVFKQHHRGAMLPPALLKQCLVSA----PTMISSLEILFAAGDRLSSQDAILARRAVGSGV-Y-NAYGPTENTVLS

Нерибосомные пептиды

YAFDLGYTCMFPVLLGGGELHIVQKETYTAPDEIAHYIKEHGITYIKLTPSLFHTIVNTASFAFDANFESLRLIVLGGEKIIPIDVIAFRKMYGHTE-FINHYGPTTEATIGA
-AFDVSAAGDFARALLTGGQLIVCPNEVKMDPASLYAIIKKYDITIFEATPALVIPLMEYI-YEQKLDISQLQILIVGSDSCSMEDFKTLVSRFGSTIRIVNSYGVTEACIDS
IAFDASSWEIYAPLLNGGTVVCIDYYTTIDIKALEAVFKQHHIRGAMLPPALLKQCLVSA---PTMISSLEILFAGDRLSSQDAILARRAVGSGV-Y-NAYGPTENTVLS



LTKVGHIG -> Asp (Aspartic acid)

VGEIGSID -> Orn (Ornithine)

AWMFAAVL -> Val (Valine)

Попарное vs множественное

- Парное выравнивание помогает найти различия между двумя геномами



Попарное vs множественное

- Парное выравнивание помогает найти различия между двумя геномами
- Множественное выравнивание необходимо для сравнения большого количества родственных геномов



Множественное выравнивание (3D)

Рассмотрим выравнивание трех последовательностей:
ATGCG, ACGTA, ATCACA

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

Как и в случае парного выравнивания не бывает столбца, состоящего из гэпов!

Множественное выравнивание (3D)

Рассмотрим выравнивание трех последовательностей:
ATGCG, ACGTA, ATCACA

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

Как и в случае парного выравнивания не бывает столбца, состоящего из гэпов!

Множественное выравнивание (3D)

01223455

AT_GCG_

01123445

A_CGT_A

01234556

ATCAC_A

Множественное выравнивание (3D)

01223455

AT_GCG_

01123445

A_CGT_A

01234556

ATCAC_A



(0, 0, 0) → (1, 1, 1) →

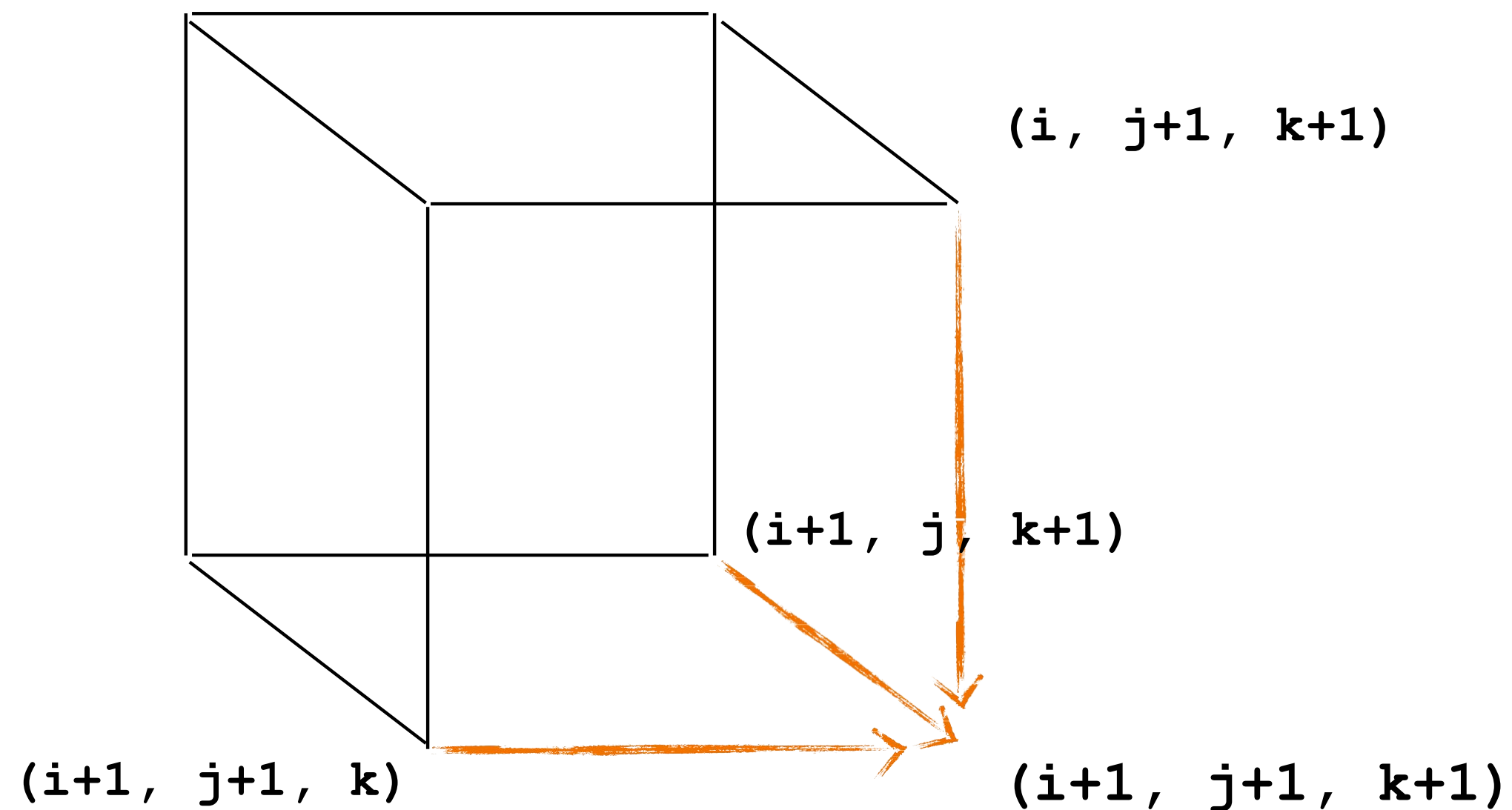
(2, 1, 2) → (2, 2, 3) →

(3, 3, 4) → (4, 4, 5) →

(5, 5, 6)

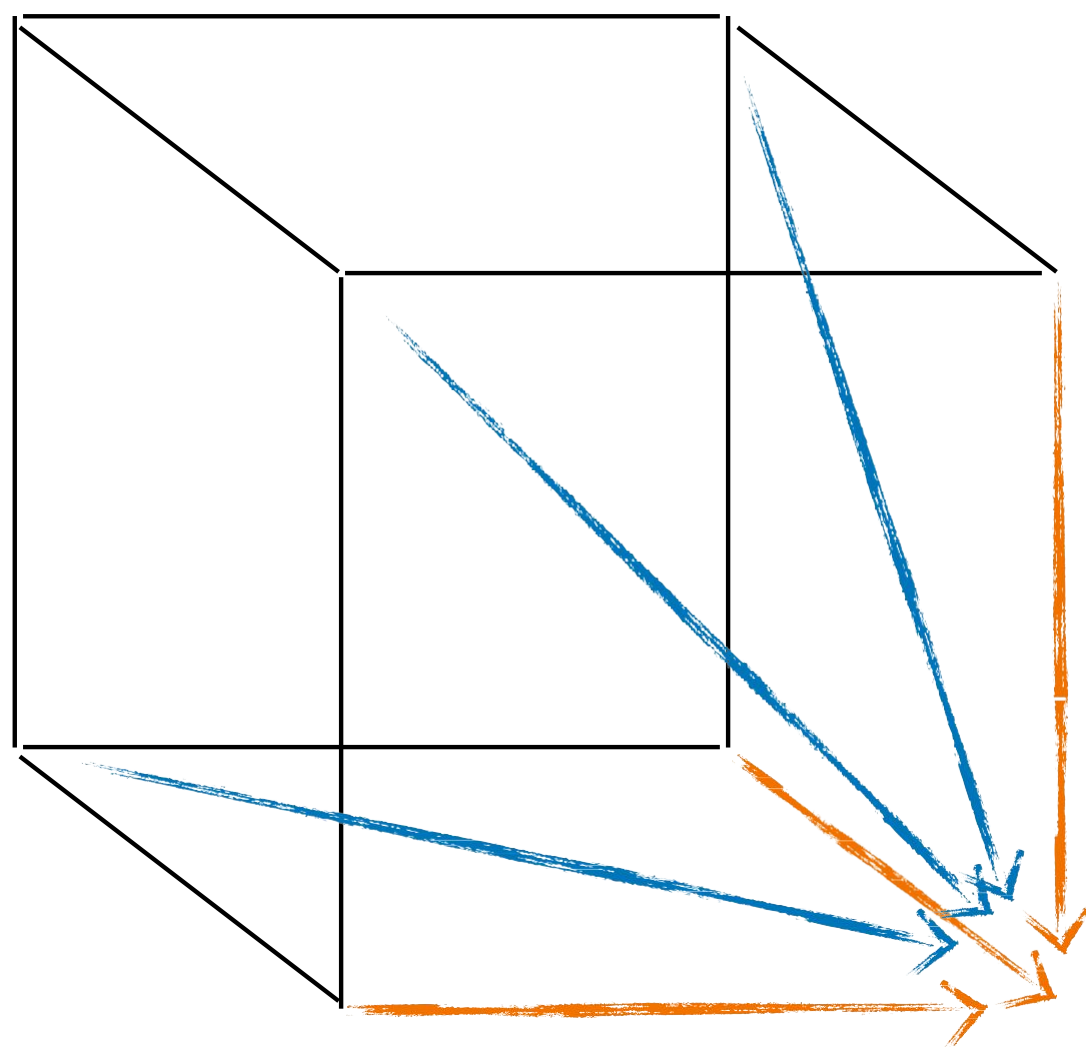
Множественное выравнивание (3D)

(i, j, k)



Множественное выравнивание (3D)

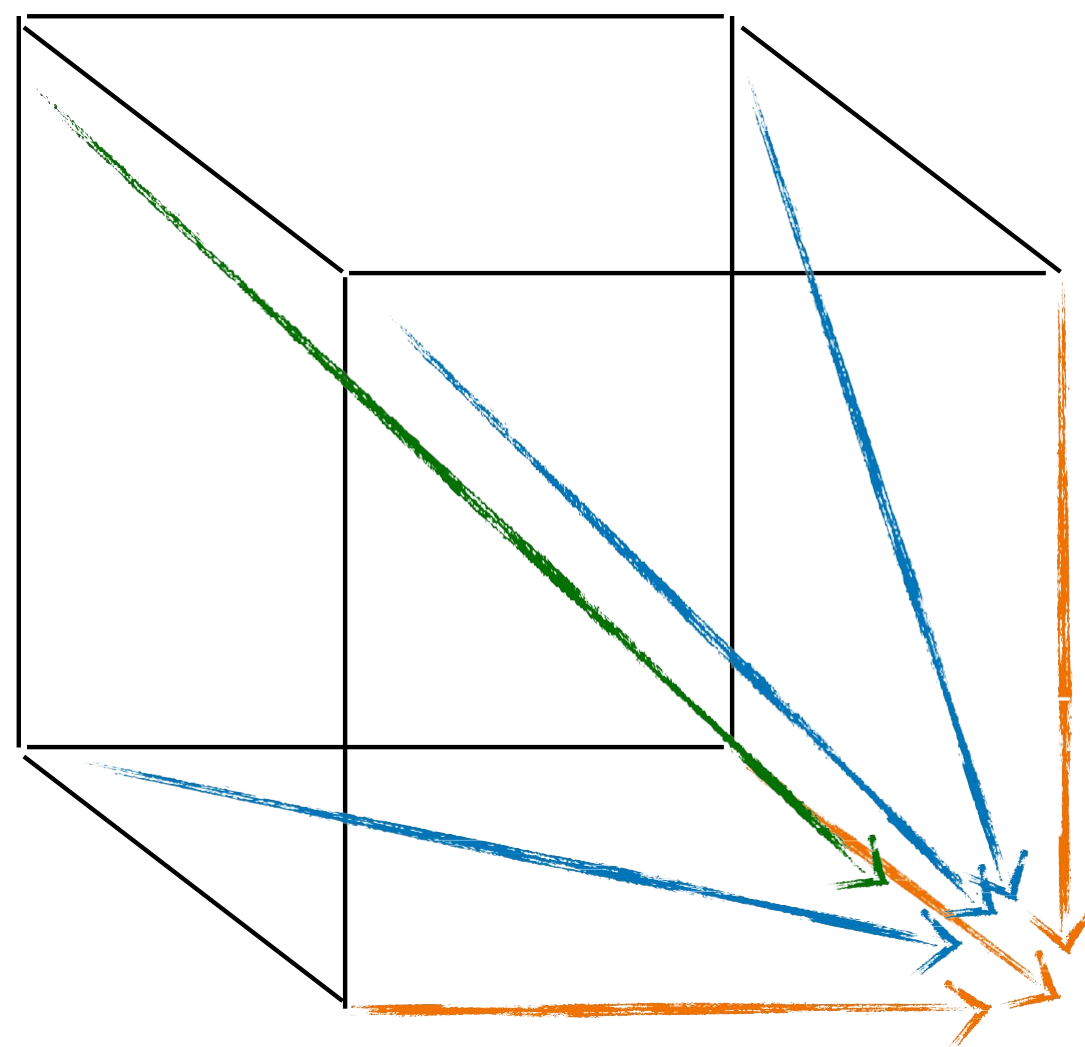
(i, j, k)



$(i+1, j+1, k+1)$

Множественное выравнивание (3D)

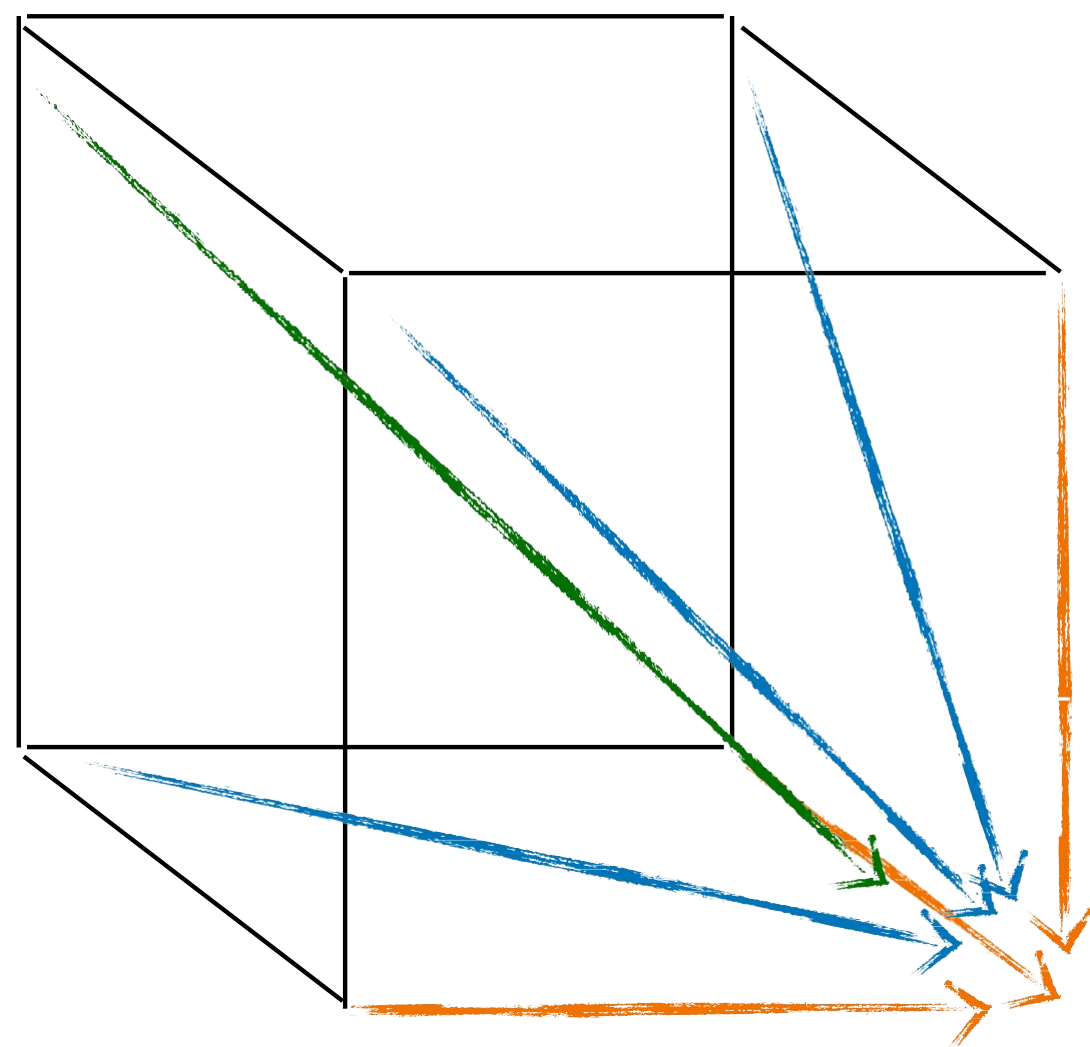
(i, j, k)



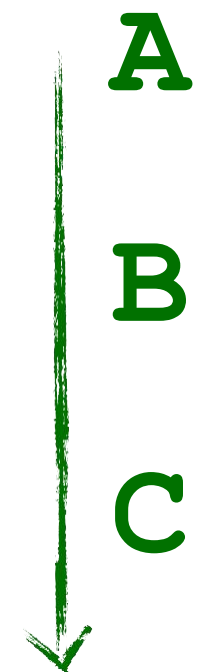
$(i+1, j+1, k+1)$

Множественное выравнивание (3D)

(i, j, k)



$(i+1, j+1, k+1)$



Множественное выравнивание (3D)

$$D_{i,j,k} = \max \left\{ \begin{array}{l} D_{i-1,j-1,k-1} + w(a_i, b_j, c_k) \\ D_{i-1,j-1,k} + w(a_i, b_j, _) \\ D_{i-1,j,k-1} + w(a_i, _, c_k) \\ D_{i,j-1,k-1} + w(_, b_j, c_k) \\ D_{i-1,j,k} + w(a_i, _, _) \\ D_{i,j-1,k} + w(_, b_j, _) \\ D_{i,j,k-1} + w(_, _, c_k) \end{array} \right.$$

Множественное выравнивание: весовая функция

Для попарного выравнивания, мы максимизировали некоторую функцию w

Для множественного выравнивания не очевидно как ее определить

Сумма пар (Sum-Of-Pairs) часто используемая метрика для подобных выравниваний

Множественное выравнивание: весовая функция

Сумма пар (Sum-Of-Pairs) - для каждой колонки складываются веса всех пар символов в этой колонке

Например, если $w(\text{match}) = 1$, $w(\text{mismatch}) = -1$, $w(\text{gap}) = -2$, то

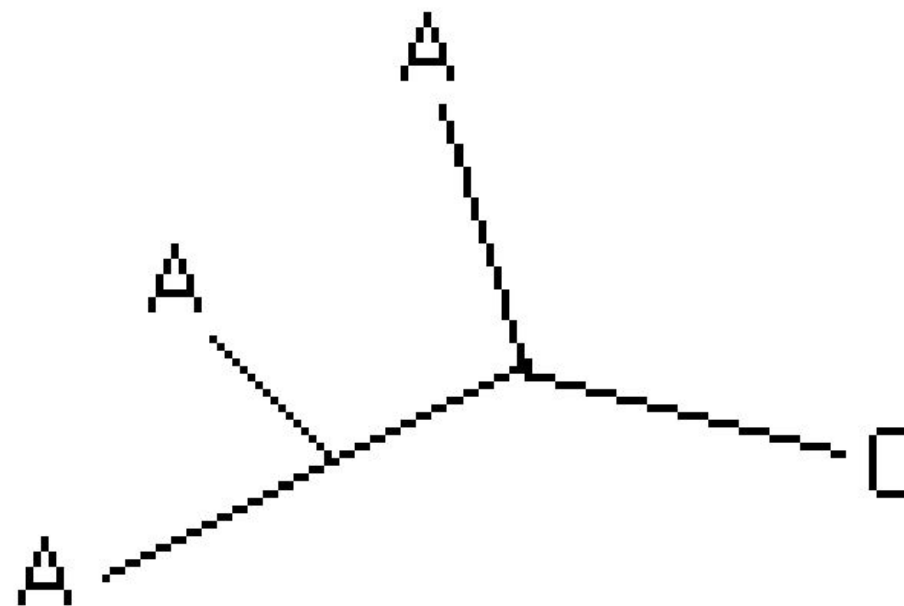
$$\begin{array}{l} \text{I} \\ - \\ \text{I} \\ \text{V} \end{array} = \text{score}(\text{I}, -) + \text{score}(\text{I}, \text{I}) + \text{score}(\text{I}, \text{V}) + \text{score}(-, \text{I}) + \text{score}(-, \text{V}) + \text{score}(\text{I}, \text{V})$$
$$= -2 + 1 + -1 + -2 + -2 + -1 = -7$$

Множественное выравнивание: весовая функция

Сумма пар (Sum-Of-Pairs) - хороша ли данная метрика?

Рассмотрим следующую колонку: A A A C

$$w(A, A, A, C) = 0$$



Эволюционную историю можно описать одной мутацией,
возможно функция слишком строга

Множественное выравнивание (3D)

- о Размер матрицы:

Множественное выравнивание (3D)

- Размер матрицы:

$$O(n^3)$$

- Количество переходов:

7

Множественное выравнивание (3D)

- Размер матрицы:

$$O(n^3)$$

- Количество переходов:

7

- Общая сложность:

$$7n^3$$

Множественное выравнивание

Рассмотрим множество строк (x_1, x_2, \dots, x_k) где $x_{i,j} \in \mathbb{A}$

Множественное выравнивание — такое множество строк $(x_1^*, x_2^*, \dots, x_k^*)$ где $x_{i,j}^* \in (\mathbb{A} \cup \{__\})$, что

1. $|x_i^*| = N$
2. $x_{1,i}^* \neq _\text{ или } \dots, x_{k,i}^* \neq _\text{}$
3. При удалении всех гэпов из x_i^* получаем x_i

Множественное выравнивание

Сложность динамического алгоритма для (x_1, x_2, \dots, x_k) , одинаковой длины

- Размер матрицы:

Множественное выравнивание

Сложность динамического алгоритма для (x_1, x_2, \dots, x_k) , одинаковой длины

- Размер матрицы:

$$n^k$$

Множественное выравнивание

Сложность динамического алгоритма для (x_1, x_2, \dots, x_k) , одинаковой длины

- Размер матрицы:

$$n^k$$

- Количество переходов:

Множественное выравнивание

Сложность динамического алгоритма для (x_1, x_2, \dots, x_k) , одинаковой длины

- Размер матрицы:

$$n^k$$

- Количество переходов:

$$2^k - 1$$

-

Сложность динамики:

$$O((2n)^k)$$

Множественное выравнивание

Сложность динамики:

$$O((2n)^k)$$

Даже для 20 последовательностей длины 4 придется совершить порядка 10^{18} операций!

Что же делать!? :(

Прогрессивное выравнивание!

Основная идея — использование жадности и парного выравнивания

Рассмотрим выравнивание

```
AT_GCG_  
A_CGT_A  
ATCAC_A
```

Прогрессивное выравнивание!

Основная идея — использование жадности и парного выравнивания

Рассмотрим выравнивание

```
AT_GCG_  
A_CGT_A  
ATCAC_A
```

Оно содержит все парные выравнивания!

```
AT_GCG_  
A_CGT_A
```

```
AT_GCG_  
ATCAC_A
```

```
A_CGT_A  
ATCAC_A
```

Прогрессивное выравнивание

Восстановим множественное выравнивание из парных? :)

Прогрессивное выравнивание

Восстановим множественное выравнивание из парных? :)

В общем случае не можем

AAAATTTT
____TTTGGGG

____AAAATTTT
GGGGAAAA_____

TTTTGGGG____
_____GGGGAAAA

Прогрессивное выравнивание

Профиль множественного выравнивания

	–	A	G	G	C	T	A	T	C	A	C	C	T	G
T	A	G	–	C	T	A	C	C	A	–	–	–	–	G
C	A	G	–	C	T	A	C	C	A	–	–	–	–	G
C	A	G	–	C	T	A	T	C	A	C	–	G	G	G
C	A	G	–	C	T	A	T	C	G	C	–	G	G	G
A	0	1	0	0	0	0	1	0	0	.8	0	0	0	0
C	.6	0	0	0	1	0	0	.4	1	0	.6	.2	0	0
G	.0	0	1	.2	0	0	0	0	0	.2	0	0	.4	1
T	.2	0	0	0	0	1	0	.6	0	0	0	0	.2	0
–	.2	0	0	.8	0	0	0	0	0	0	.4	.8	.4	0

Прогрессивное выравнивание

Профиль множественного выравнивания

A	0	1	0	0	0	0	1	0	0	.8	0	0	0	0
C	.6	0	0	0	1	0	0	.4	1	0	.6	.2	0	0
G	.0	0	1	.2	0	0	0	0	0	.2	0	0	.4	1
T	.2	0	0	0	0	1	0	.6	0	0	0	0	.2	0
-	.2	0	0	.8	0	0	0	0	0	0	.4	.8	.4	0

Консенсус

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

Прогрессивное выравнивание

Если мы имеем 2 профиля выравнивания, то построим консенсус обоих и выровняем.

		G	A	T	T	A	C	A
	0	1	2	3	4	5	6	7
A	1	1	1	2	3	4	5	6
A	2	2	1	2	3	3	4	5
G	3	2	2	2	3	4	4	5
A	4	3	2	3	3	3	4	4
G	5	4	3	3	4	4	4	5
T	6	5	4	3	3	4	5	5
A	7	6	5	4	4	3	4	5
C	8	7	6	5	5	4	3	4

GAT^TTAC^A
AAGAG^TTAC_—

Прогрессивное выравнивание

Жадный алгоритм на основе парных выравниваний

- Посчитаем попарные расстояния выравнивания
-
- Заменим 2 самые близкие последовательности их консенсусом
- Начнем все заново
-

Будем продолжать пока не останется 1 последовательность

Прогрессивное выравнивание

GATTCA, GTCTGA, GATATT, GTCAGC

Вставки, и удаления и мутации -1

Совпадения $+1$

Прогрессивное выравнивание

GATTCA, GTCTGA, GATATT, GTCAGC

Вставки, и удаления и мутации -1

Совпадения +1

GAT-TCA 1
G-TCTGA

GAT-TCA 1
GATAT-T

GATTCA-- 0
G-T-CAGC


G-TCTGA -1
GATAT-T

GTCTGA 2
GTCAGC

GAT-ATT -1
G-T-CAGC

Прогрессивное выравнивание

GTCTGA
GTCAGC



GTC [A, T] GA

Таким образом задача свелась к подзадаче меньшего размера

GATTCA,
GATATT,
GTC [A, T] GA

Прогрессивное выравнивание

Сложность по
времени:

Прогрессивное выравнивание

Сложность по времени:

На каждом шаге нужно считать попарные выравнивания $O(k_i^2 l^2)$ где l – оценка сверху на длину последовательности $l \leq nk$, а k_i – количество последовательностей на данном шаге

Прогрессивное выравнивание

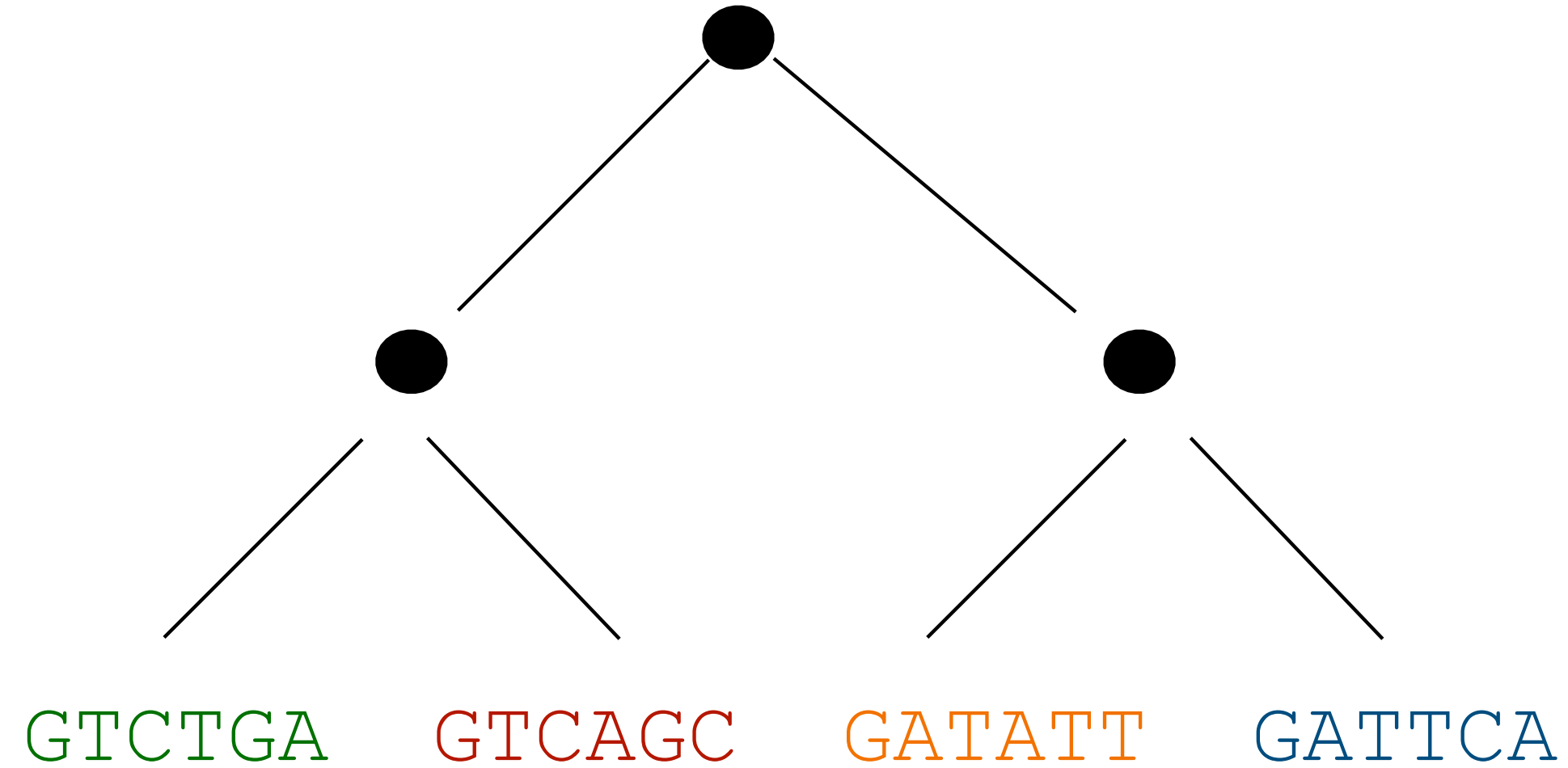
Сложность по времени:

На каждом шаге нужно считать попарные выравнивания $O(k_i^2 l^2)$ где l – оценка сверху на длину последовательности $l \leq nk$, а k_i – количество последовательностей на данном шаге

$$k \begin{cases} k^2(nk)^2 \\ (k-1)^2(nk)^2 \\ \dots \\ (1)^2(nk)^2 \end{cases} = (kn)^2 \sum_{i=1}^k i^2 = (kn)^2 \frac{k(k-1)(2k+1)}{6} = O(k^5 n^2)$$

Замечания

- Зависит от порядка жадности. Мы можем построить дерево 1 раз и следовать ему в дальнейшем



Замечания

- Зависит от порядка жадности. Мы можем построить дерево 1 раз и следовать ему в дальнейшем
- Если у нас есть финальное глобальное выравнивание и скоринг функция то можем улучшить выравнивание

Замечания

- Зависит от порядка жадности. Мы можем построить дерево 1 раз и следовать ему в дальнейшем
- Если у нас есть финальное глобальное выравнивание и скоринг функция то можем улучшить выравнивание
- Начиная с какой то глубины можем использовать динамику

Замечания

- Зависит от порядка жадности. Мы можем построить дерево 1 раз и следовать ему в дальнейшем
- Если у нас есть финальное глобальное выравнивание и скоринг функция то можем улучшить выравнивание
- Начиная с какой то глубины можем использовать динамику
- На самом деле, все попарные расстояния пересчитывать не нужно на каждом шаге, достаточно только расстояний до нового консенсуса. С учетом этого, сложность алгоритма $O(k^4 n^2)$

Резюмируем

- Множественное выравнивание позволяет находить консервативные домены в генома
- Хорошее но не оптимальное выравнивание можно искать за полиномиальное время
- Полученное выравнивание можно улучшать оптимизационными методами (имитация отжига, генетические алгоритмы)