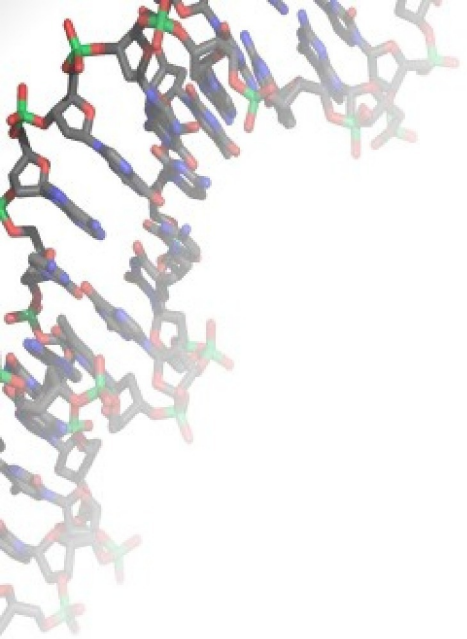
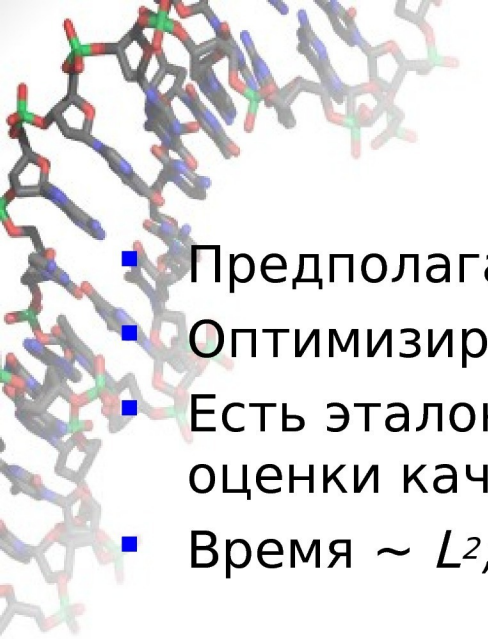


Часть 2

Как выравнивать очень длинные последовательности?



Выравнивание

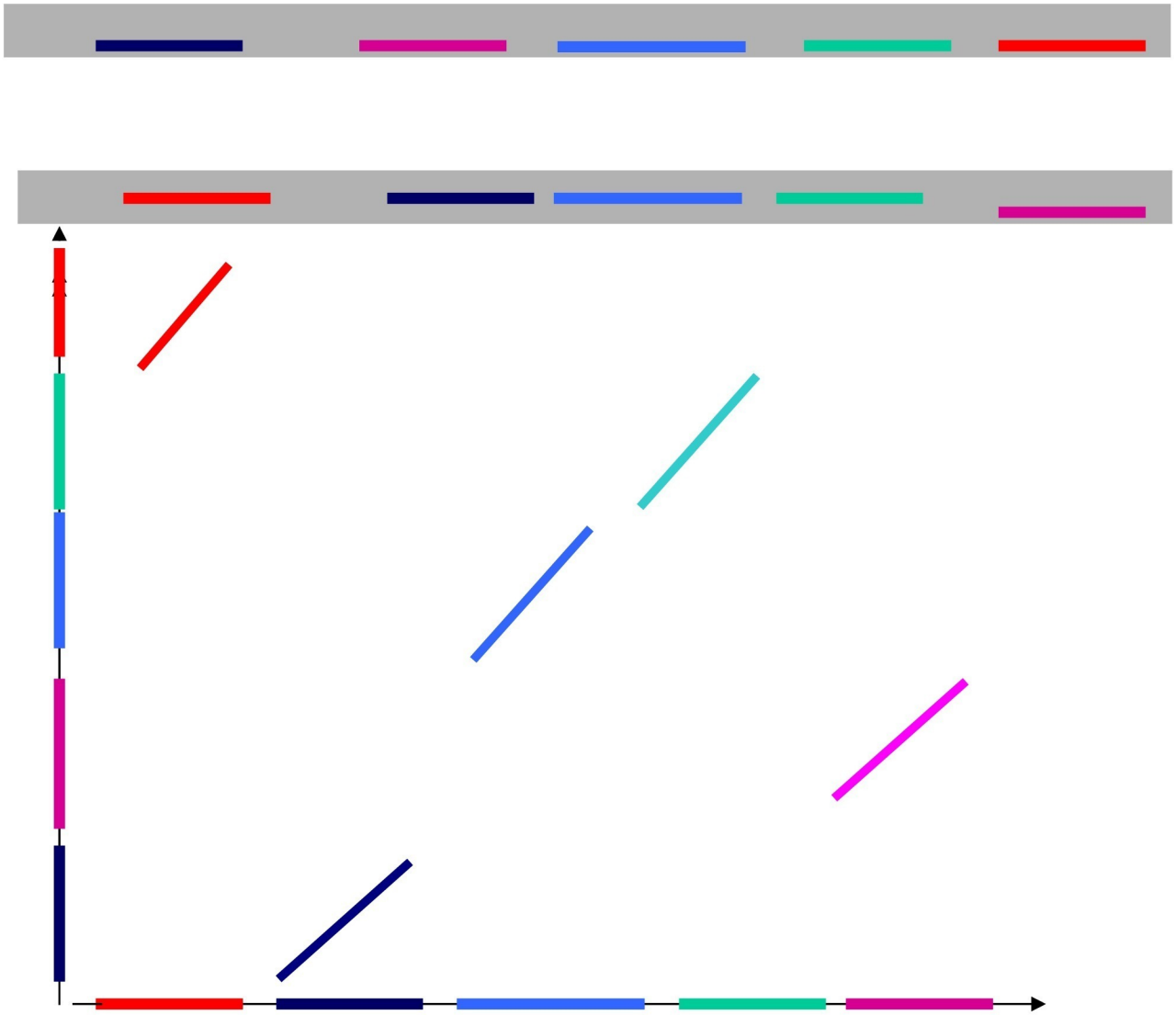
- Предполагаем наличие общего предка
- Оптимизируем вес выравнивания
- Есть эталонные выравнивания (для подбора весов и оценки качества)
- Время ~ L^2; $L \sim 10^2 - 10^3$

ОЧЕНЬ ДЛИННЫЕ ПОСЛЕДОВАТЕЛЬНОСТИ

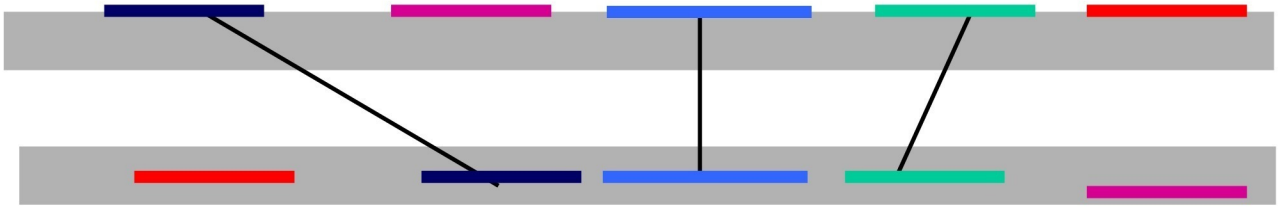
- $L \sim 10^7$
- Эталонных выравниваний нет

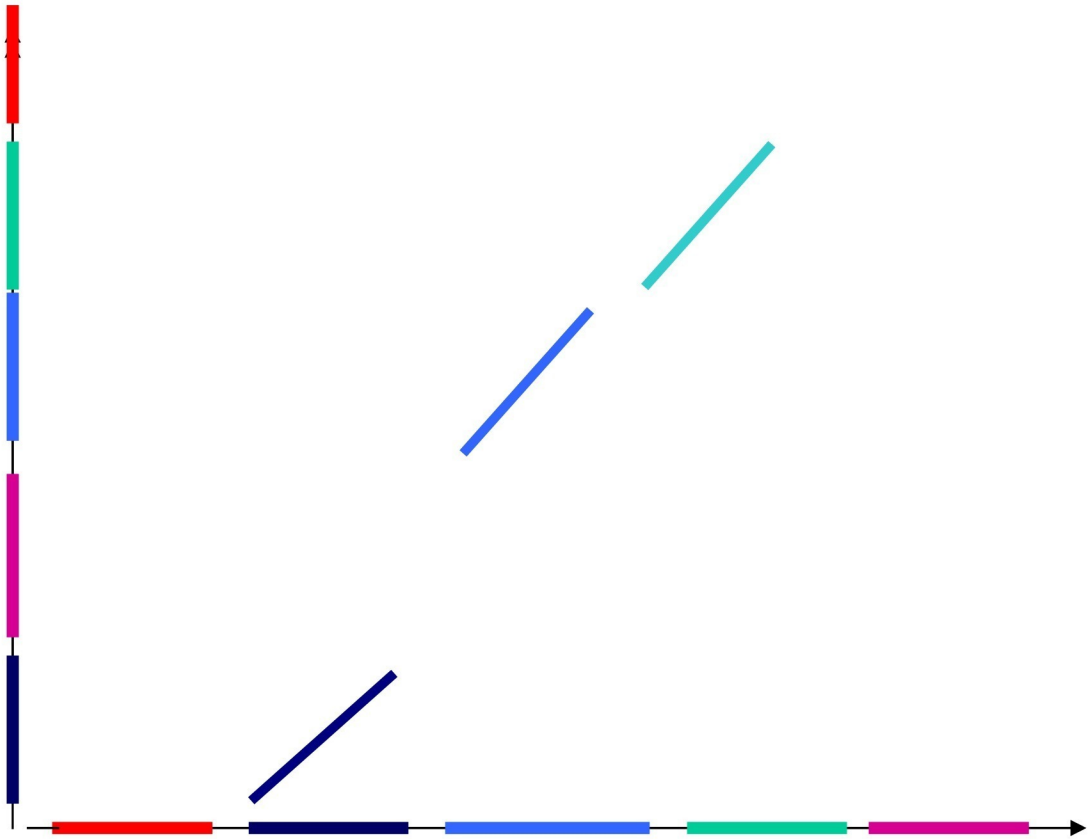
ЧТО ДЕЛАТЬ?

Выравнивание на основе локальных сходств

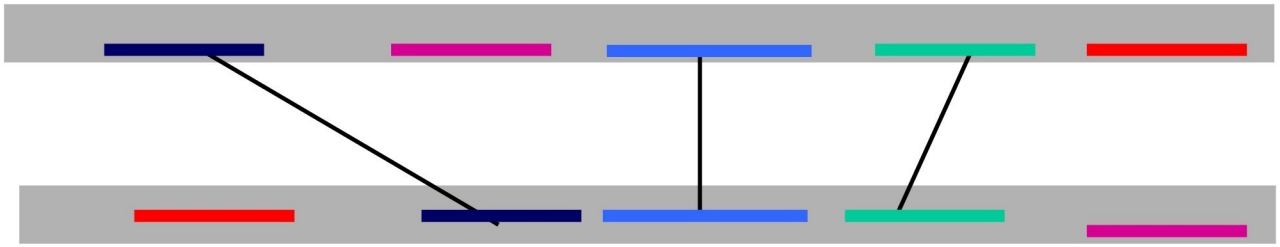


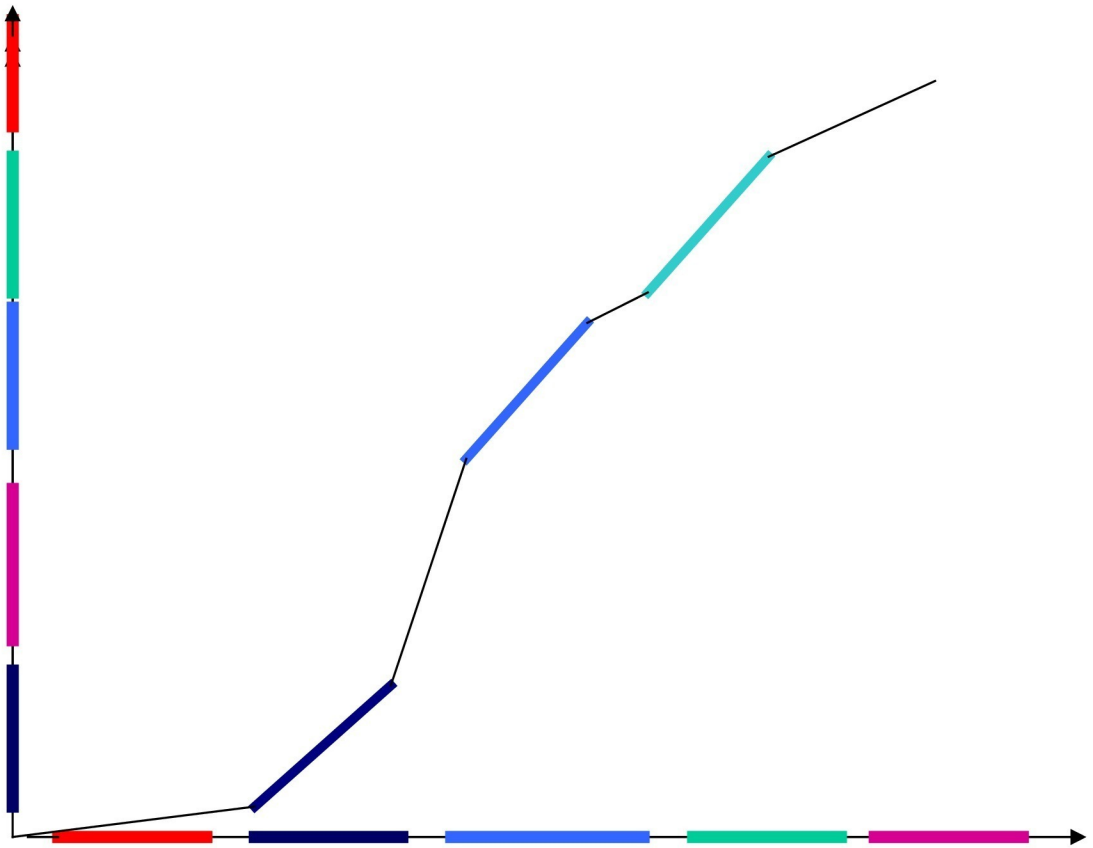
Цепь «не конфликтующих» локальных сходств

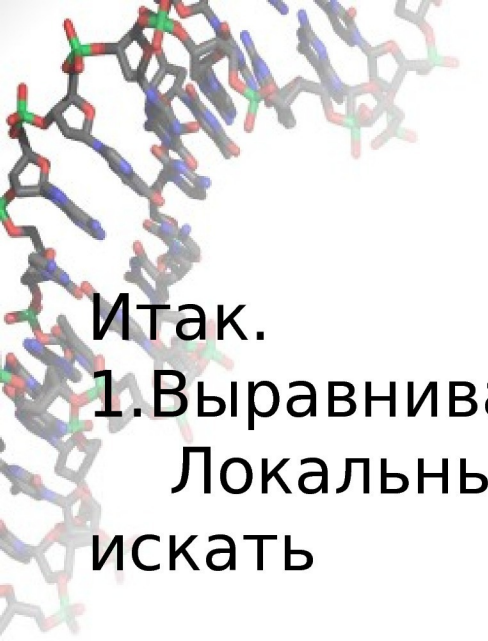




Глобальное выравнивание



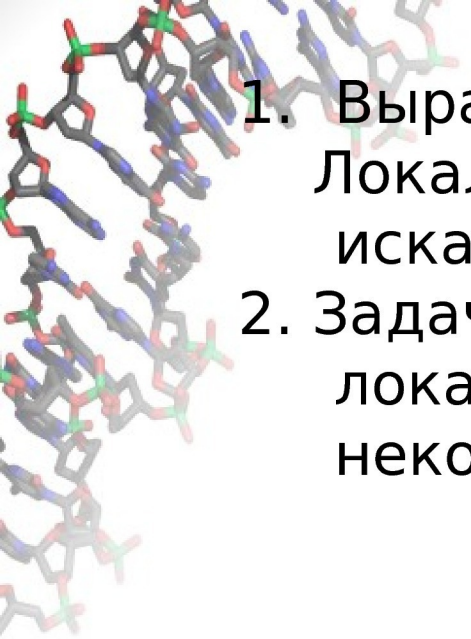




Итак.

1. Выравнивание – это цепочка локальных сходств.
   Локальные (неточные) сходства можно легко искать

2. Задача выравнивания – выделить в множестве локальных сходств «разумную» цепочку неконфликтующих локальных сходств



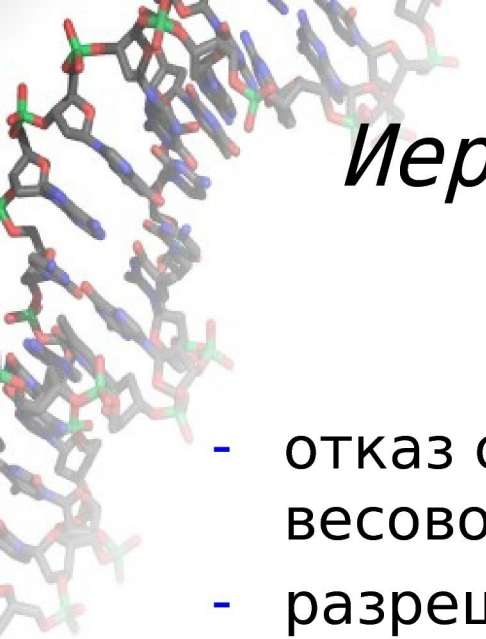
1. Выравнивание – это цепочка локальных сходств. Локальные (неточные) сходства можно легко искать
2. Задача выравнивания – выделить в множестве локальных сходств «разумную» цепочку неконфликтующих локальных сходств

ПРОБЛЕМА:

Чтобы получить детальное выравнивание, исходных локальных сходств должно быть много

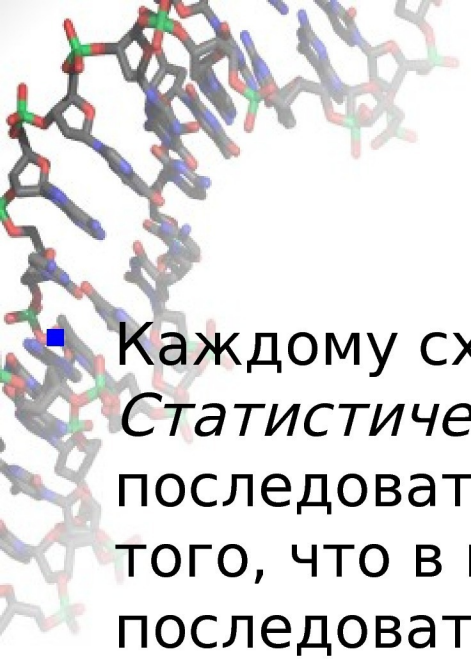
РЕШЕНИЕ:

Иерархическое выравнивание



Иерархическое выравнивание

- отказ от глобальной оптимизации какой-либо весовой весовой функции;
- разрешение конфликта между сходствами производится локально – путем сравнения их статистической значимости
- иерархическая процедура



Статистическая значимость локального сходства

- Каждому сходству *S* приписывается вес *W(S).* *Статистическая значимость локального* сходства *S* в последовательностях длин *m* и *n* – это вероятность того, что в независимых случайных последовательностях длин *m* и *n* есть сходство того же или большего веса.

Нужно уточнить распределение вероятностей на последовательностях

Как вычислить статистическую значимость *p(W)* локального сходства по его весу W

- Обозначения:
- $Z(v_1, v_2)$ – вес оптимального безделеционного локального сходства последовательностей v_1 и v_2. Весовая матрица *M[a, b]* считается заданной
- *P(w)* – вероятность того, что для независимых случайных бернуллиевских последовательностей v_1 и v_2
- длины *m* и *n* соответственно выполнено: $Z(v_1, v_2) > w$.
- Теорема (Karlin, Altschul, 1999).

$$P(w) \approx \exp(-Kmn\, e^{-\lambda w})$$

здесь ***K, λ*** зависят только от весовой матрицы M.

Термин: Распределение экстремальных значений.

Условия применимости распределения экстремальных значений

- Теорема (Karlin, Altschul, 1999).

$$P(w) \approx \exp(-Kmn\, e^{-\lambda w}) \qquad (*)$$

Теория: Необходимо выполнение следующего:

1) $Exp(M[x, y]) < 0$

2) $Max(M[x, y]) > 0$

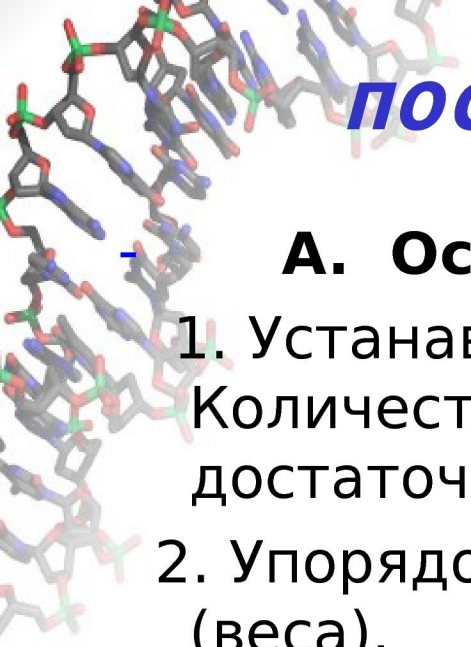
3) распределение вероятностей – бернуллиево

4) рассматриваются только безделеционные сходства

Практика:

Условия 1) и 2) в реальных случаях всегда выполняются

Если условия 3) и 4) не выполняются, то (*) можно использовать приближенно, а коэффициенты ***K, λ*** определить методом Монте-Карло



Иерархическая процедура построения цепочки локальных сходств

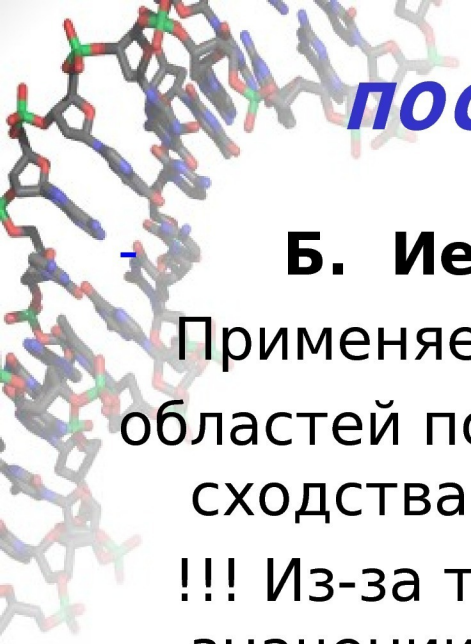
- **А. Основной шаг** (жадный алгоритм)

1. Устанавливаем порог статистической значимости. Количество локальных сходств –кандидатов достаточно мало.
2. Упорядочиваем сходства по убыванию значимости (веса).
3. Берем очередное сходство. Если оно ни с кем не конфликтует – включаем в цепочку.

   Иначе – отбрасываем [на самом деле – откладываем для дальнейшего изучения]

** Техническая проблема: повторы (“low complexity”).

Нужно отдельно фильтровать.



Иерархическая процедура построения цепочки локальных сходств

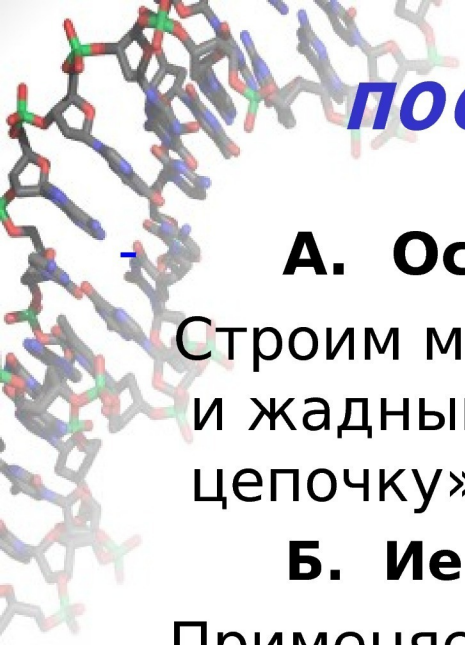
- **Б. Иерархический переход**

Применяем основной шаг отдельно к каждой из областей поиска между отобранными локальными сходствами.

!!! Из-за того, что области стали меньше, при том же значении порога

сходства становятся значимыми при меньшем весе

Поэтому промежутки между сходствами в цепочке будут постепенно заполняться



Иерархическая процедура построения цепочки локальных сходств

А. Основной шаг

Строим множество статистически значимых сходств и жадным алгоритмом выбираем «остовную цепочку».

Б. Иерархический переход

Применяем основной шаг отдельно к каждой из областей поиска между отобранными локальными сходствами.

В. Итерирование

Повторяем шаг **Б** до тех пор, пока появляются новые значимые сходства.

Иерархическая процедура построения цепочки локальных сходств

- **А. Основной шаг**

Строим множество статистически значимых сходств и жадным алгоритмом выбираем «остовную цепочку».

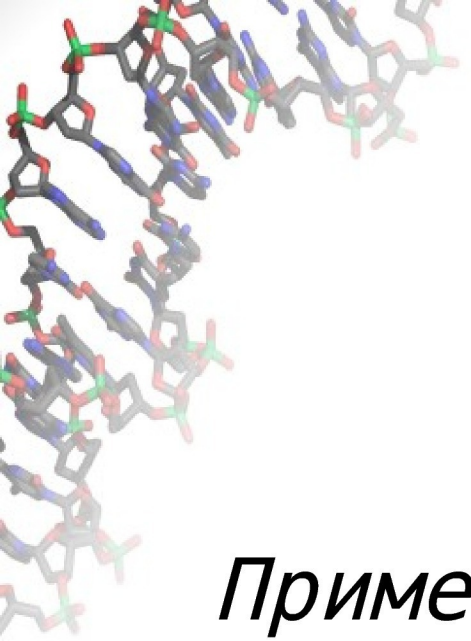
Б. Иерархический переход

Применяем основной шаг отдельно к каждой из областей поиска между отобранными локальными сходствами.

В. Итерирование

Повторяем шаг **Б** до тех пор, пока появляются новые значимые сходства.

Время определяется принятым уровнем значимости (а не минимальным допустимым весом сходства)



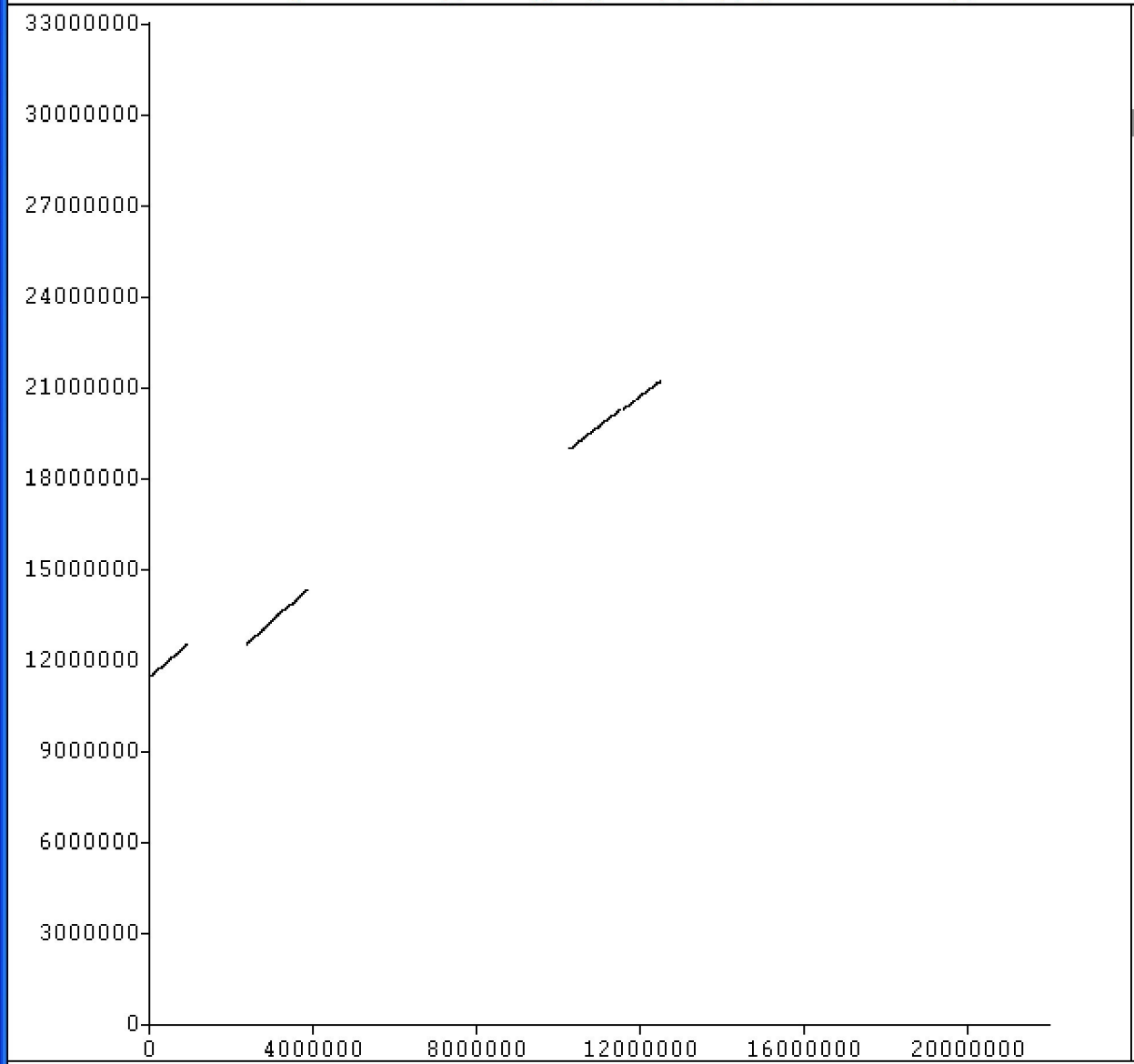
Пример: иерархическое геномное выравнивание (длина ~ 10^7)

Owen v1.4

File Construct Conflict Filter Select Edit Window Help

1 .. 20206255 D.Viril scaffold_12963 type=golden_path_region; loc=scaffold_12963:1..20206255; ID=scaffold_12963; dbxref=GB:CH94064

1 .. 32352404 D.Mojaven scaffold_6500 type=golden_path_region; loc=scaffold_6500:1..32352404; ID=scaffold_6500; dbxref=GB:CH9338



88% 136722 .. 136877 (156) 11585441 .. 11585593 (153

86% 136946 .. 137141 (196) 11585773 .. 11585979 (207

78% 137166 .. 137314 (149) 11586093 .. 11586230 (138

85% 137789 .. 137898 (110) 11587018 .. 11587135 (118

84% 137988 .. 138183 (196) 11587256 .. 11587452 (197

86% 138242 .. 138344 (103) 11587538 .. 11587645 (108

83% 138410 .. 138557 (148) 11587753 .. 11587903 (151

84% 138825 .. 139049 (225) 11588378 .. 11588594 (217

89% 139775 .. 140039 (265) 11589342 .. 11589603 (262

81% 140195 .. 140463 (269) 11589698 .. 11589984 (287

85% 140553 .. 140640 (88) 11590135 .. 11590222 (88)

85% 141082 .. 141235 (154) 11590839 .. 11590995 (157

95% 141304 .. 141365 (62) 11591047 .. 11591109 (63)

77% 141390 .. 141484 (95) 11591139 .. 11591239 (101)

79% 141529 .. 141733 (205) 11591256 .. 11591474 (219

82% 142350 .. 142536 (187) 11592639 .. 11592816 (178

87% 142592 .. 142712 (121) 11592860 .. 11592986 (127

87% 142728 .. 142957 (230) 11593038 .. 11593267 (230

136946 AGGATAACGATGACGTTGCCTGAGCGAACACAATGAAGCTTATAAACGAAATTCAATAGCTGAAAATCGGCGTCCCATCTGCGCTC---------- 137032

>>>

1585773 ATGATGACGATGACGTTGCCTGAACGCACACAATGAAGCTTATAAACGAAATTCAATAGCTGAAAATCGGCGTCCCATCTGTGCTCTTTGGGATTC 1585868

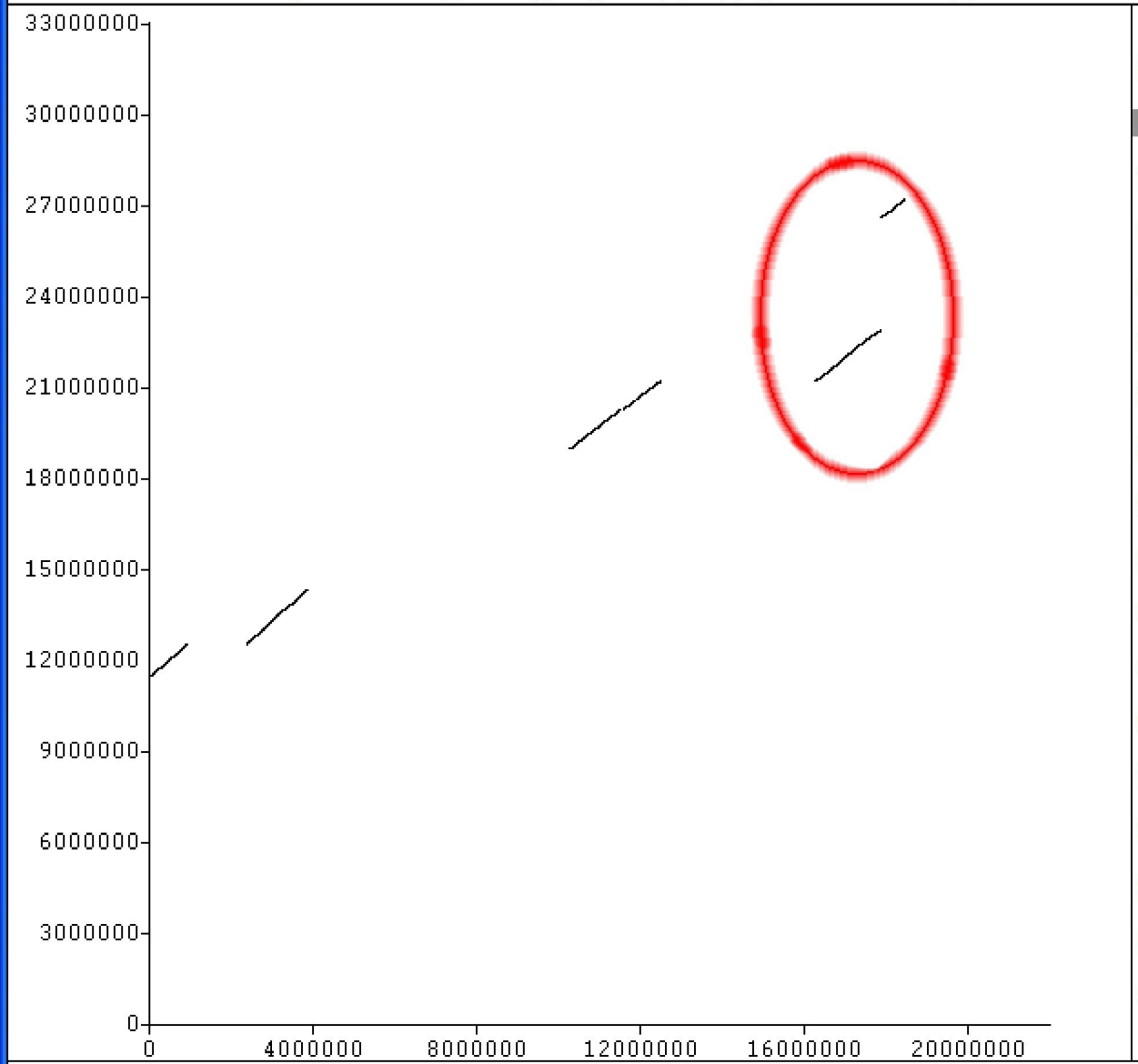
Action: Greedy Deleted: 0 Total: 11707 Conflicting: 0

Owen v1.4

File Construct Conflict Filter Select Edit Window Help

1 .. 20206255 D.Viril scaffold_12963 type=golden_path_region; loc=scaffold_12963:1..20206255; ID=scaffold_12963; dbxref=GB:CH94064

1 .. 32352404 D.Mojaven scaffold_6500 type=golden_path_region; loc=scaffold_6500:1..32352404; ID=scaffold_6500; dbxref=GB:CH9338



88% 136722 .. 136877 (156) 11585441 .. 11585593 (153

86% 136946 .. 137141 (196) 11585773 .. 11585979 (207

78% 137166 .. 137314 (149) 11586093 .. 11586230 (138

85% 137789 .. 137898 (110) 11587018 .. 11587135 (118

84% 137988 .. 138183 (196) 11587256 .. 11587452 (197

86% 138242 .. 138344 (103) 11587538 .. 11587645 (108

83% 138410 .. 138557 (148) 11587753 .. 11587903 (151

84% 138825 .. 139049 (225) 11588378 .. 11588594 (217

89% 139775 .. 140039 (265) 11589342 .. 11589603 (262

81% 140195 .. 140463 (269) 11589698 .. 11589984 (287

85% 140553 .. 140640 (88) 11590135 .. 11590222 (88)

85% 141082 .. 141235 (154) 11590839 .. 11590995 (157

95% 141304 .. 141365 (62) 11591047 .. 11591109 (63)

77% 141390 .. 141484 (95) 11591139 .. 11591239 (101)

79% 141529 .. 141733 (205) 11591256 .. 11591474 (219

82% 142350 .. 142536 (187) 11592639 .. 11592816 (178

87% 142592 .. 142712 (121) 11592860 .. 11592986 (127

87% 142728 .. 142957 (230) 11593038 .. 11593267 (230

136946 AGGATAACGATGACGTTGCCTGAGCGAACACAATGAAGCTTATAAACGAAATTCAATAGCTGAAAATCGGCGTCCCATCTGCGCTC---------- 137032

| ||| |||||||||||||||| || |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| |||| >>>

1585773 ATGATGACGATGACGTTGCCTGAACGCACACAATGAAGCTTATAAACGAAATTCAATAGCTGAAAATCGGCGTCCCATCTGTGCTCTTTGGGATTC 1585868

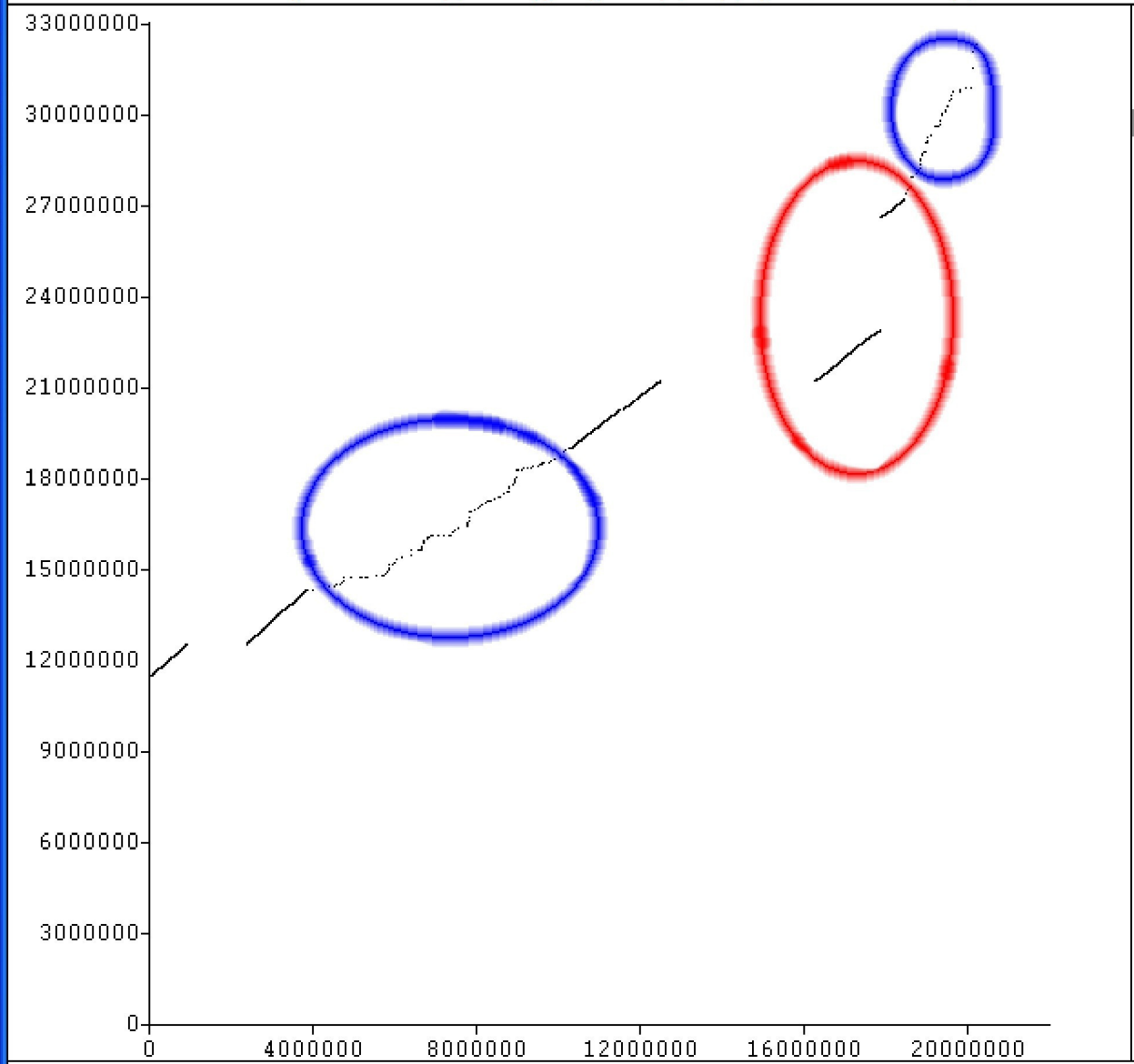
Action: Greedy Deleted: 0 Total: 11707 Conflicting: 0

Owen v1.4

File Construct Conflict Filter Select Edit Window Help

1 .. 20206255 D.Viril scaffold_12963 type=golden_path_region; loc=scaffold_12963:1..20206255; ID=scaffold_12963; dbxref=GB:CH94064

1 .. 32352404 D.Mojaven scaffold_6500 type=golden_path_region; loc=scaffold_6500:1..32352404; ID=scaffold_6500; dbxref=GB:CH9338



88% 136722 .. 136877 (156) 11585441 .. 11585593 (153

86% 136946 .. 137141 (196) 11585773 .. 11585979 (207

78% 137166 .. 137314 (149) 11586093 .. 11586230 (138

85% 137789 .. 137898 (110) 11587018 .. 11587135 (118

84% 137988 .. 138183 (196) 11587256 .. 11587452 (197

86% 138242 .. 138344 (103) 11587538 .. 11587645 (108

83% 138410 .. 138557 (148) 11587753 .. 11587903 (151

84% 138825 .. 139049 (225) 11588378 .. 11588594 (217

89% 139775 .. 140039 (265) 11589342 .. 11589603 (262

81% 140195 .. 140463 (269) 11589698 .. 11589984 (287

85% 140553 .. 140640 (88) 11590135 .. 11590222 (88)

85% 141082 .. 141235 (154) 11590839 .. 11590995 (157

95% 141304 .. 141365 (62) 11591047 .. 11591109 (63)

77% 141390 .. 141484 (95) 11591139 .. 11591239 (101)

79% 141529 .. 141733 (205) 11591256 .. 11591474 (219

82% 142350 .. 142536 (187) 11592639 .. 11592816 (178

87% 142592 .. 142712 (121) 11592860 .. 11592986 (127

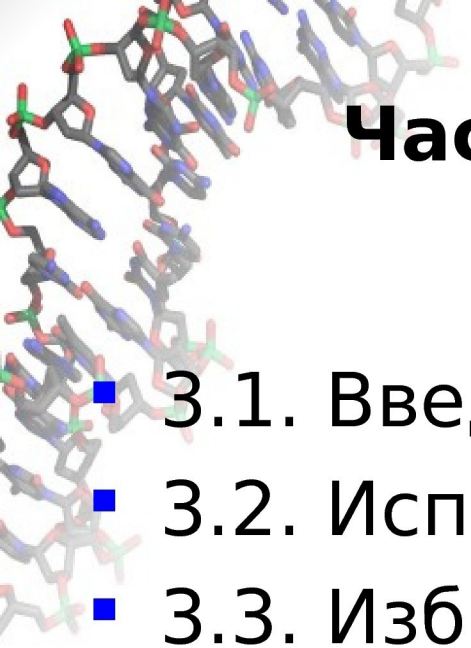
87% 142728 .. 142957 (230) 11593038 .. 11593267 (230

136946 AGGATAACGATGACGTTGCCTGAGCGAACACAATGAAGCTTATAAACGAAATTCAATAGCTGAAAATCGGCGTCCCATCTGCGCTC---------- 137032

| ||| |||||||||||||||| || |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| |||| >>>

1585773 ATGATGACGATGACGTTGCCTGAACGCACACAATGAAGCTTATAAACGAAATTCAATAGCTGAAAATCGGCGTCCCATCTGTGCTCTTTGGGATTC 1585868

Action: Greedy Deleted: 0 Total: 11707 Conflicting: 0

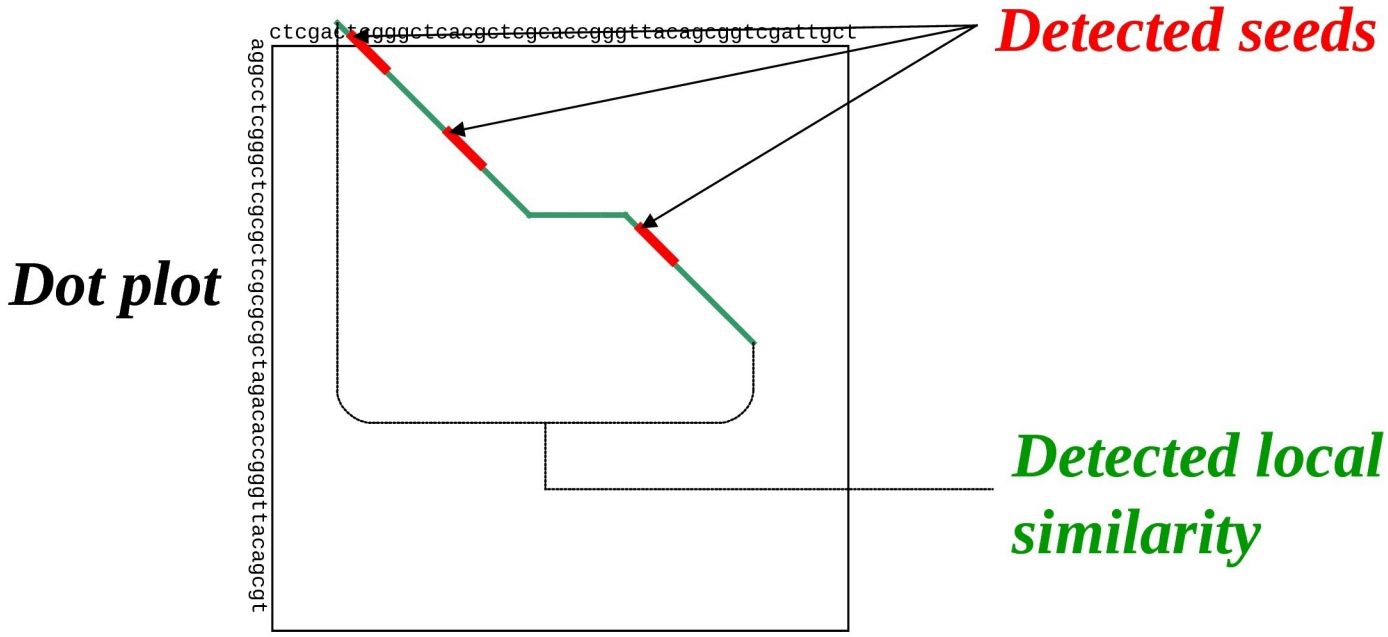


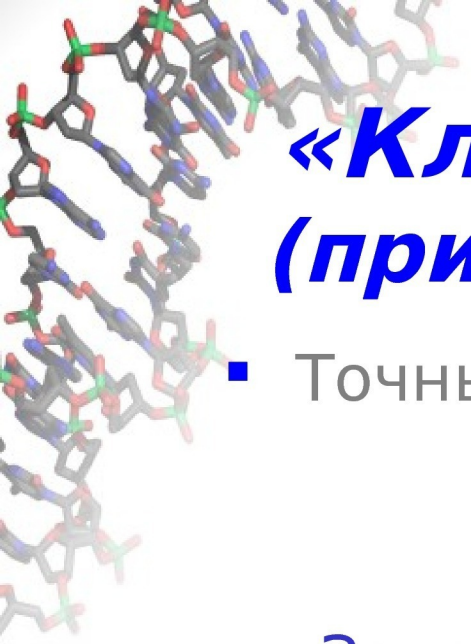
Часть *3. Поиск всех разумных локальных сходств*

- 3.1. Введение
- 3.2. Использование затравок (seed)
- 3.3. Избирательность и чувствительность
- 3.4. Типы затравок (seed model)

Затравки: фильтрация пространства поиска

- Сначала ищем небольшие и легко диагностируемые участки сходства («затравочные сходства», *seed similarities*).
- Далее ищем сходства только в окрестностях затравочных сходств (одного или нескольких).





«Классическая затравка» (пример: 6 совпадений подряд)

- Точные совпадения :

```
ATCAGT
||||||
ATCAGT
```

Затравка («затравочное слово», описание затравочных сходств) : ######

Вес **: 6** [количество #]

- **Пример : 16 совпадений из 20**

```
       #######
ATCAGTGCAATGCTCATGAA
|||.|.|||||||:||.|||
ATCGGCGCAATGCGCAAGAA
```

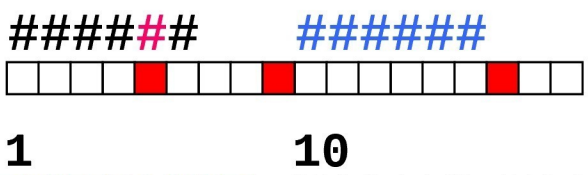
Затравка ловит сходство
(затравка соответствует сходству)

- Затравка ##### ← *seed*

Затравочное сходство (... выравнивание)

ATGCAA

ATGCAA



Затравка соответствует сходству в позиции 10

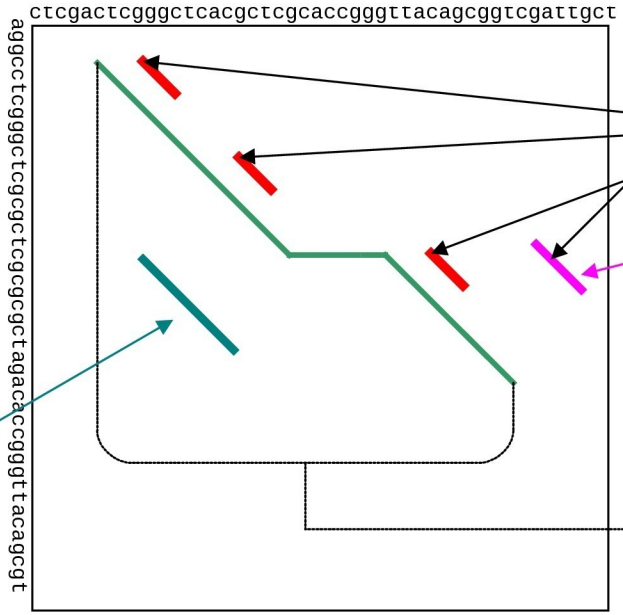
Затравка не соответствует сходству в позиции 1

Затравка ловит сходство

Возможные проблемы

```
##☹### [16 of 20!]
ATCAGTGCGATGCTCATGAA
|||.||||:|||:||.|||
ATCGGTGCGGTGCGCAAGAA
```

```
######
TCAGTGCAATGCTCATGAA
:|::::||||||:::..::
CCGACACAATGCGTGACCC
```

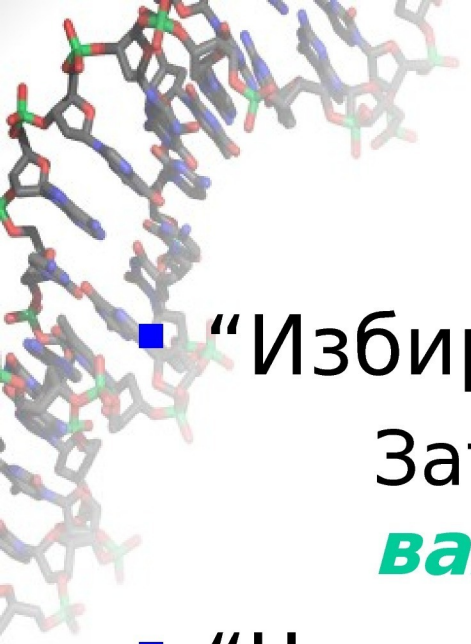


Найденные затравки

Случайное сходство

Detected local similarity

Пропущенное сходство: не содержит затравок



Итак:

- “Избирательность”

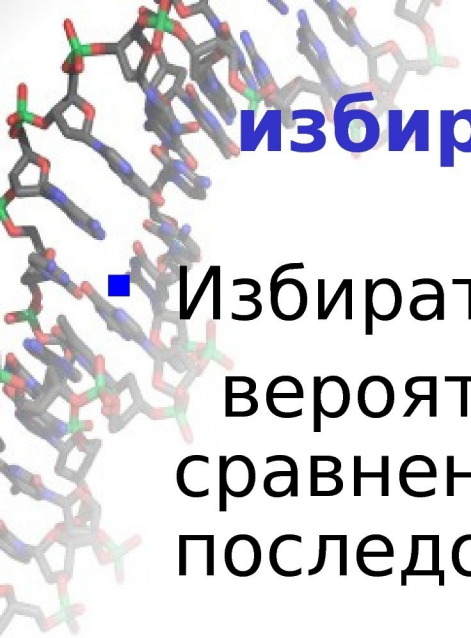
  Затравка может НЕ быть частью ***важного (для нас) сходства***

- “Чувствительность”

  Важное (для нас) сходство *может НЕ содержать ни одной затравки*

Нужно уточнить:

- **Что такое «важное сходство»?**



Что может быть мерой избирательности и чувствительности

- Избирательность затравки: ~ ***4-weight***

  вероятность ее обнаружения при сравнении независимых случайных последовательностей

- Чувствительность затравки:

  вероятность *того, что затравка попадет в* ***важное сходство****.*

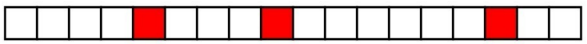
Нужно уточнить:

- **Что такое «важное сходство»?**
- **Каково распределение вероятностей для важных сходств?**

Множество важных [целевых] выравниваний и их вероятности

- Выравнивания фиксированной длины без удалений

GCTACGACTTCGAGCTGC



...CTCAGCTATGACCTCGAGCGGCCTATCTA. ***L=18***

- Вероятностная модель: ***Бернулли*** ;

Случайные выравнивания: *Prob(match) =0.25*

Целевые выравнивания: *Prob(match) >> 0.25*

Обобщения: Марковские модели, скрытые марковские модели (сегодня не рассматриваем)

Разреженные затравки

Ma, Tromp, Li 2002 (PatternHunter)

- Затравка: ###--#-##

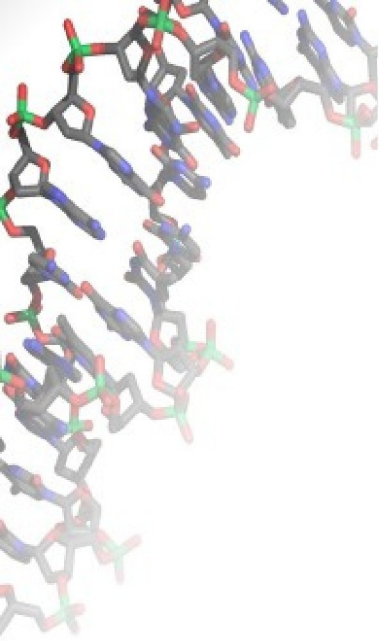
  {‘#’ : должно быть совпадение

  ‘-’ : «джокер» (“все равно, что”)

  Вес **: 6** [количество #]

- **Пример:**

```
###--#-##
ATCAGTGCAATGCTCAAGA
|||||.||.||||:||||||
ATCAGCGCGATGCGCAAGA
```

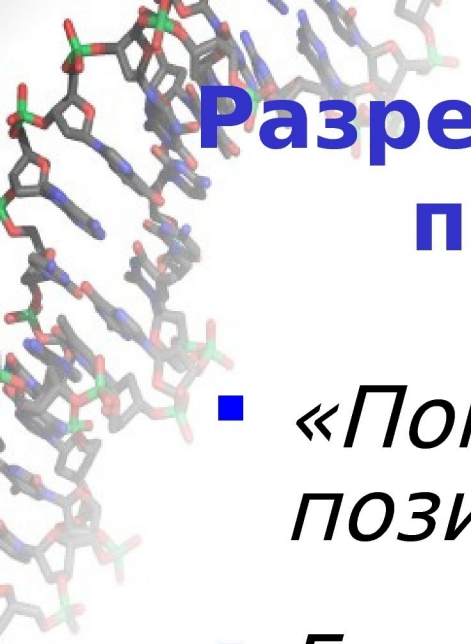


######

```
ATCAGTGCAATGCTCAAGA
|||||.||.||||:||||||
ATCAGCGCGATGCGCAAGA
######
 ######
  ######
   ######
    ######
     ######
      ######
       ######
        ######
         ######
          ######
           ######
            ######
             ######
```

###--#-##

```
ATCAGTGCAATGCTCAAGA
|||||.||.||||:||||||
ATCAGCGCGATGCGCAAGA
###--#-##
 ###--#-##
  ###--#-##
   ###--#-##
    ###--#-##
     ###--#-##
      ###--#-##
       ###--#-##
        ###--#-##
         ###--#-##
          ###--#-##
```

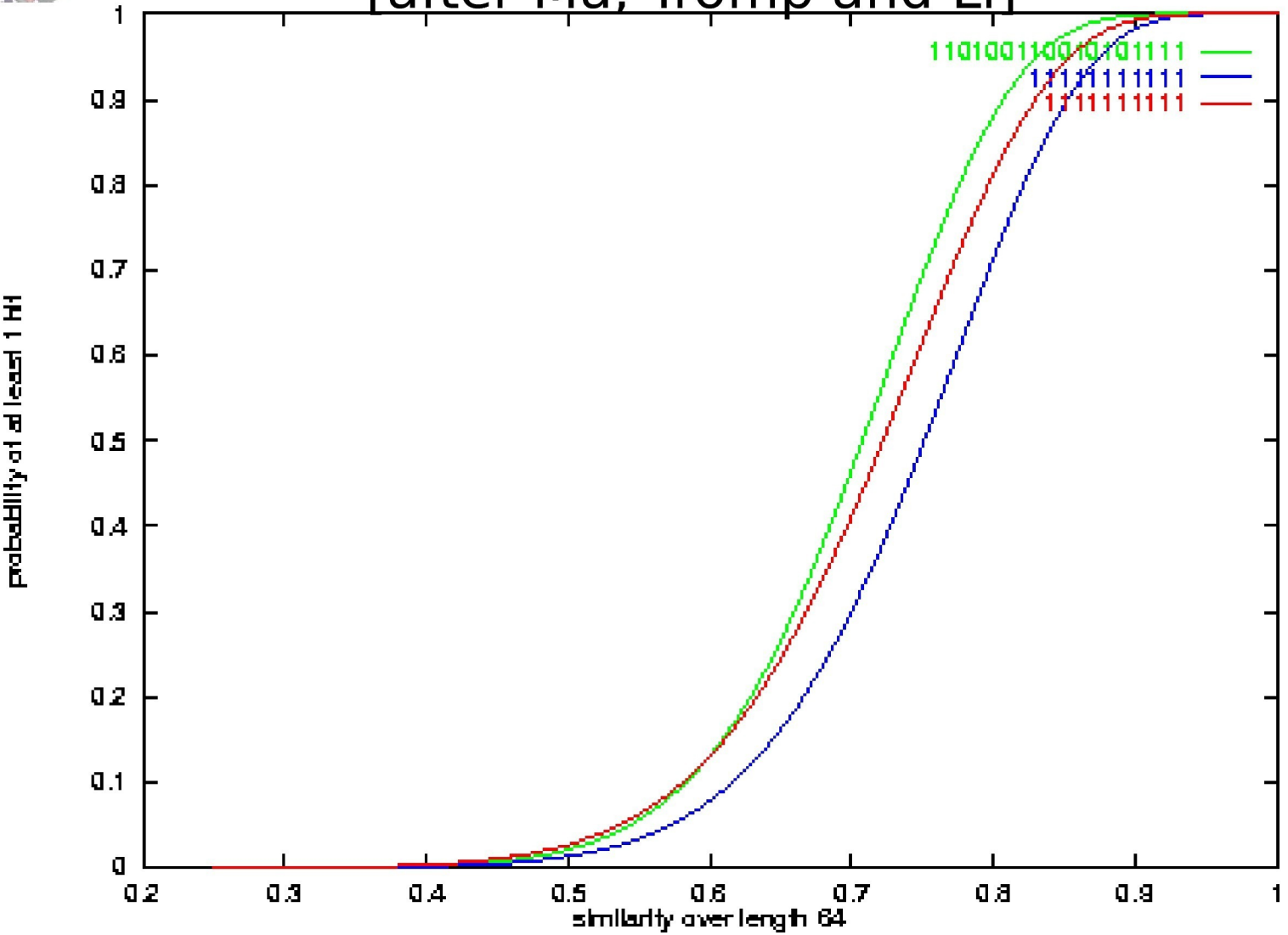


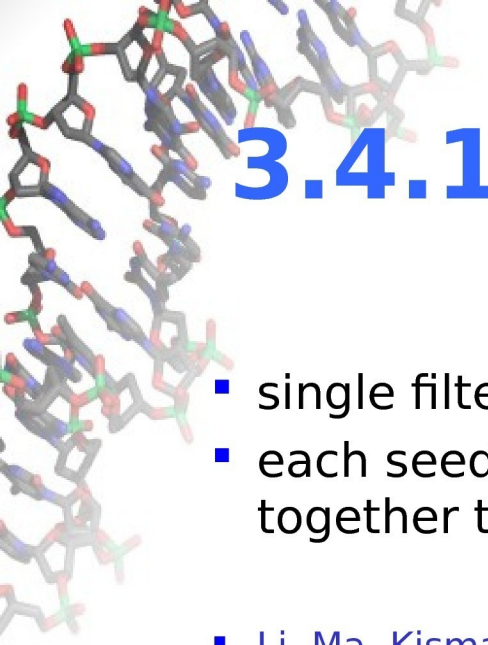
Разреженные затравки: в чем преимущество перед классическими?

- *«Попадания» в соседних позициях менее зависимы*
- *For contiguous vs. spaced seeds of the same weight, the expected number of hits is (basically) the same but the probabilities of having **at least one hit** are very different*

Sensitivity: PH weight 11 seed vs BLAST 11 & 10

[after Ma, Tromp and Li]

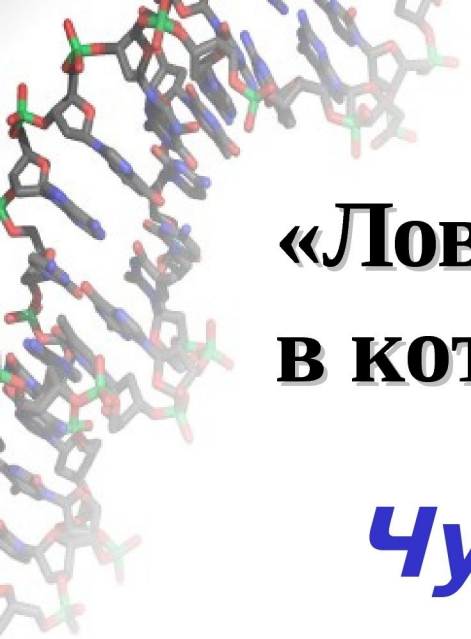




3.4.1. Семейства затравок

- single filter based on several distinct seed patterns
- each seed pattern detects a part of interesting similarities but together they detect [almost] all of them

- Li, Ma, Kisman, Tromp 2004 (PatternHunter II)
- Kucherov, Noe, Roytberg, 2005
- Sun, Buhler, RECOMB 2004



Пример: BCE (18,3)

«Ловить» все сходства длины 18,
в которых не более 3 несовпадений

Чувствительность = 1.0

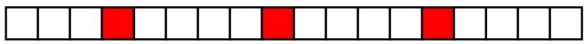
Избирательность

(вероятность случайного появления затравочного сходства) ***-> MIN***

Пример: BCE (18,3)

Обнаружить все сходства длины 18, в которых не более 3 несовпадений

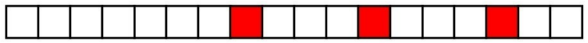
Множественная затравка *F* решает проблему BCE(18, 3)



$F \{ \begin{matrix} \#\#-\#-\#\#\#\# \\ \#\#\#--\,-\#-\,-\#\#-\# \end{matrix}$

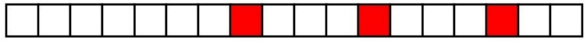
Затравка *F* состоит из двух простых затравок, каждая из них имеет вес 7

Пример: ВСЕ (18.3)



###--#--##-#

###--#--##-#



###-##--#-###

##-#-####

###--#--##-#

w=7

##-##-#####

###-####--##

###-##--#-###

##----####-###

###--#-#-##-##

###-#-#-#----###

w=9

Пример: ВСЕ (18.3). Избирательности

Избирательность семейства затравок – вероятность встретить хотя бы одну из них в случайном месте ($p(match) = 1/4$)

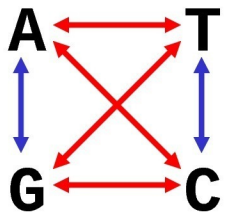
`####`	**w=4**	**~39. 10^{-4}**
`###-##`	**w=5**	**~9.8 10^{-4}**
`##-#-####` `###---#--##-#`	**w=7**	**~1.2 10^{-4}**

`##-##-#####`
`###-####--##`
`###-##---#-###`
`##----####-###`
`###---#-#-##-##`
`###-#-#-#----###`
	w=9	**~0.23 10^{-4}**

3.4.2. ДА, НЕТ и МНОГОЕ ДРУГОЕ

Different mutational events have different probabilities

transversions :



transitions .

```
ATCAGTGCAATGCTCAAGA
|||||.||.||||:||||||
ATCAGCGCGATGCGCAAGA
```

Transitions are usually over-represented.

Extended seed alphabet

- seed: ##@#-#@-###
  - '#' : obligatory match position
  - '-' : joker position ("don't care" position)
  - '@' : ***transition-constrained*** position

position that corresponds to either a match or a transition.

Subset letters and seeds

- Seed ***letter*** is a subset of aligned pairs.
- # = {(A,A), (C,C), (G,G), (T,T)}
- @ = { (A,A), (C,C), (G,G), (T,T),
  (A,G), (G,A), (T,C), (C,T)}
- - = {all pairs}

Затравки. Выводы.

- Классические заправки не оптимальны
- Применяя затравки, разберитесь, насколько они адекватны интересующему Вас классу сходств.

О чем мы не говорили:

- Как строить хорошие затравки?
- Как вычислить чувствительность затравки?
- И еще о многом…