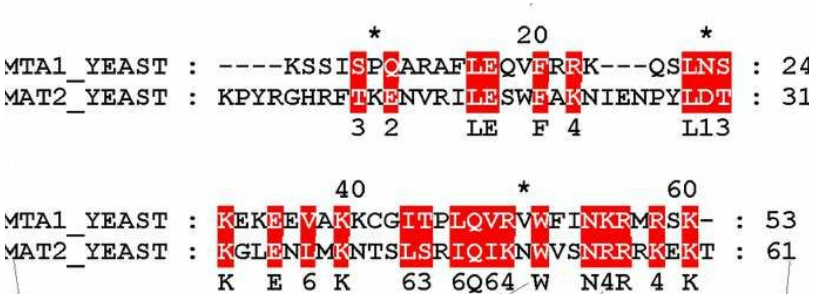
Поиск сходных последовательностей. Выравнивание



Цивов Алексей Владимирович

старший преподаватель, к.х.н.

кафедра органической и биологической химии ЯрГУ

Содержание лекции

- Гомологичные последовательности, типы гомологов
- Способы выравнивания последовательностей
- Локальные и глобальные выравнивания
- Критерии качества выравнивания
- BLAST – поиск сходных последовательностей
- Программы BLAST
- Матрицы замен
- Параметры оценки сходства в BLAST

Сходство последовательностей

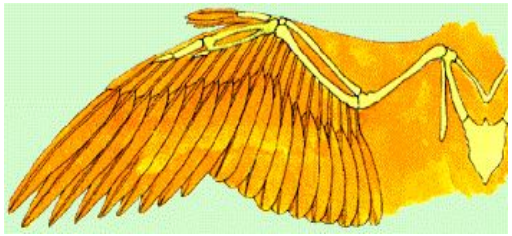
Известно, что:

1. функцию, структуру и многие другие свойства белка/ДНК **определяет последовательность**;
2. родственные белки имеют **похожие свойства**

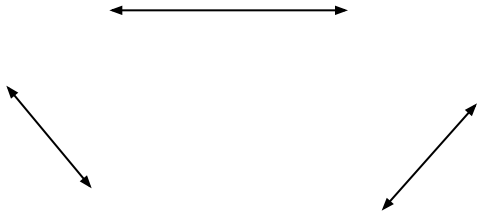
⇒ молекулы, **похожие по последовательности**, **похожи и по свойствам**

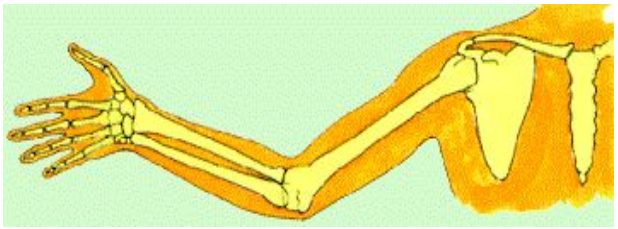
Т.о. свойства можно предсказать, анализируя изученные последовательности, **похожие на данную**

Гомология

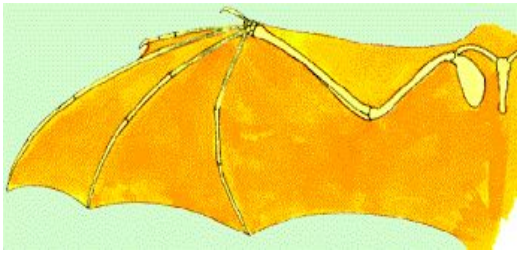


Крыло птицы





Рука человека



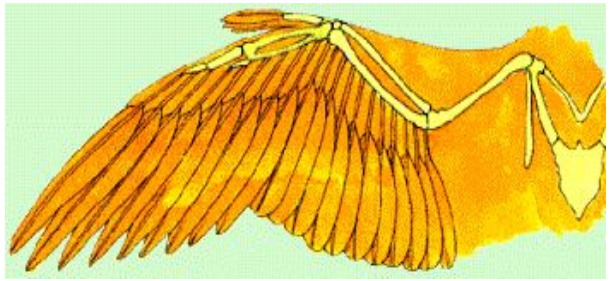
Крыло летучей мыши

- **Гомологичными** в биологии называют сопоставимые части сравниваемых биологических объектов.
- Предполагается, что гомологичные объекты **имеют общего предка**

Гомология и аналогия

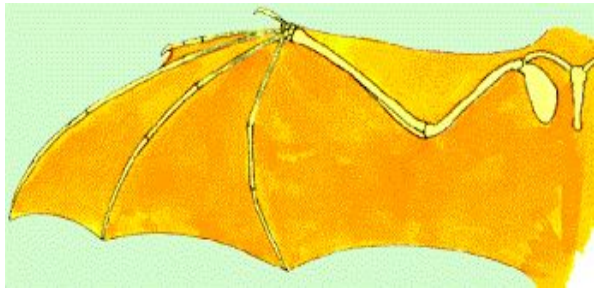
Гомология (**общий предок**) против **аналогии** (**конвергентная эволюция**)

КЭ – развитие **сходных признаков** у **различных организмов**, живущих в **сходных условиях обитания**

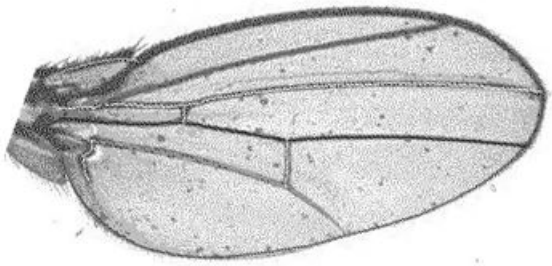


крыло птицы

крыло бабочки



крыло летучей мыши



крыло мухи

Гомологичные последовательности

Гомологичные последовательности – последовательности, имеющие **общее происхождение** (общего предка)

Признаки **гомологичности белков**:

- ✔ сходная 3D-структура
- ✔ в той или иной степени похожая аминокислотная последовательность
- ✔ выполнение одинаковых функций

Схожесть последовательностей и гомология

Следующее утверждение основано на наблюдении и **не является истинным *a priori*:**

- ✔ Если **существенные части** (фрагменты) двух последовательностей обладают **значительной схожестью** между собой, у них, **возможно,** общий предок и одинаковые функции

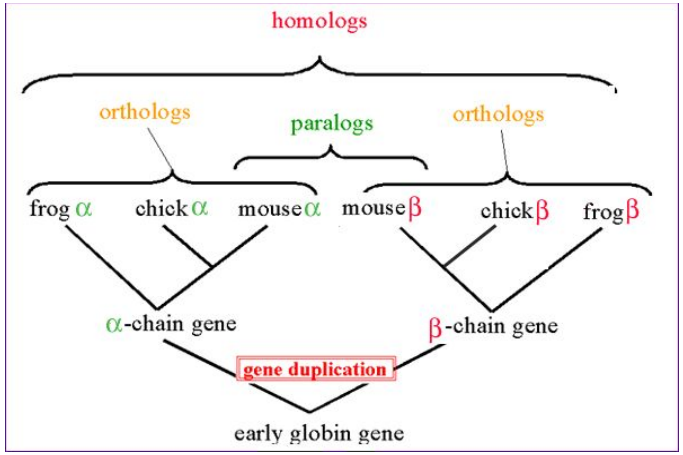
Гомология: некоторые соображения

- Вообще говоря, если две последовательности имеют высокую степень схожести **по всей длине**, то, **вероятно, они гомологичны**
- Схожесть **не обязательно** является индикатором **гомологии**
- Простые участки могут иметь высокую степень схожести, но **не быть гомологами**
- Гомологичные последовательности **не всегда схожи** с высокой степенью

Типы гомологов: ортологи и паралоги

Ортологи — последовательности, возникшие из одного общего предшественника **в процессе видообразования**. Ортологи, как правило, имеют **одну и ту же функцию**

Паралоги — последовательности, возникшие из одного общего предшественника **в результате дупликации** генов в одном организме. Паралоги, как правило, имеют **разные функции**.



Выравнивание

- **Выравнивание** - это **поиск сходства** между последовательностями и их фрагментами
- **Простейшее выравнивание** – запись последовательностей одна под другой так, чтобы гомологичные фрагменты оказались друг под другом.

```
      ДОМовой
скупиДОМ
    воДОМерка
```

Способы выравнивания двух последовательностей

Цель - максимальное количество совпадений!

- ✔ Запись последовательностей друг под другом
- ✔ Движение друг относительно друга
- ✔ Вставка пробелов (пропуски, **gap**)
- ✔ Удаление/вставка символов или фрагментов (**делеция и инсерция**)
- ✔ **Замена символов** (нуклеотиды или а/к)

Типы выравнивания

- **Локальное** – поиск **фрагментов**, наиболее похожих друг на друга

домовой **дом**овой
скупи**дом** во**дом**ерка

- **Глобальное** – сравнение **последовательностей целиком**: каждый нуклеотид (аминокислота) находит себе пару

Критерии качества выравнивания

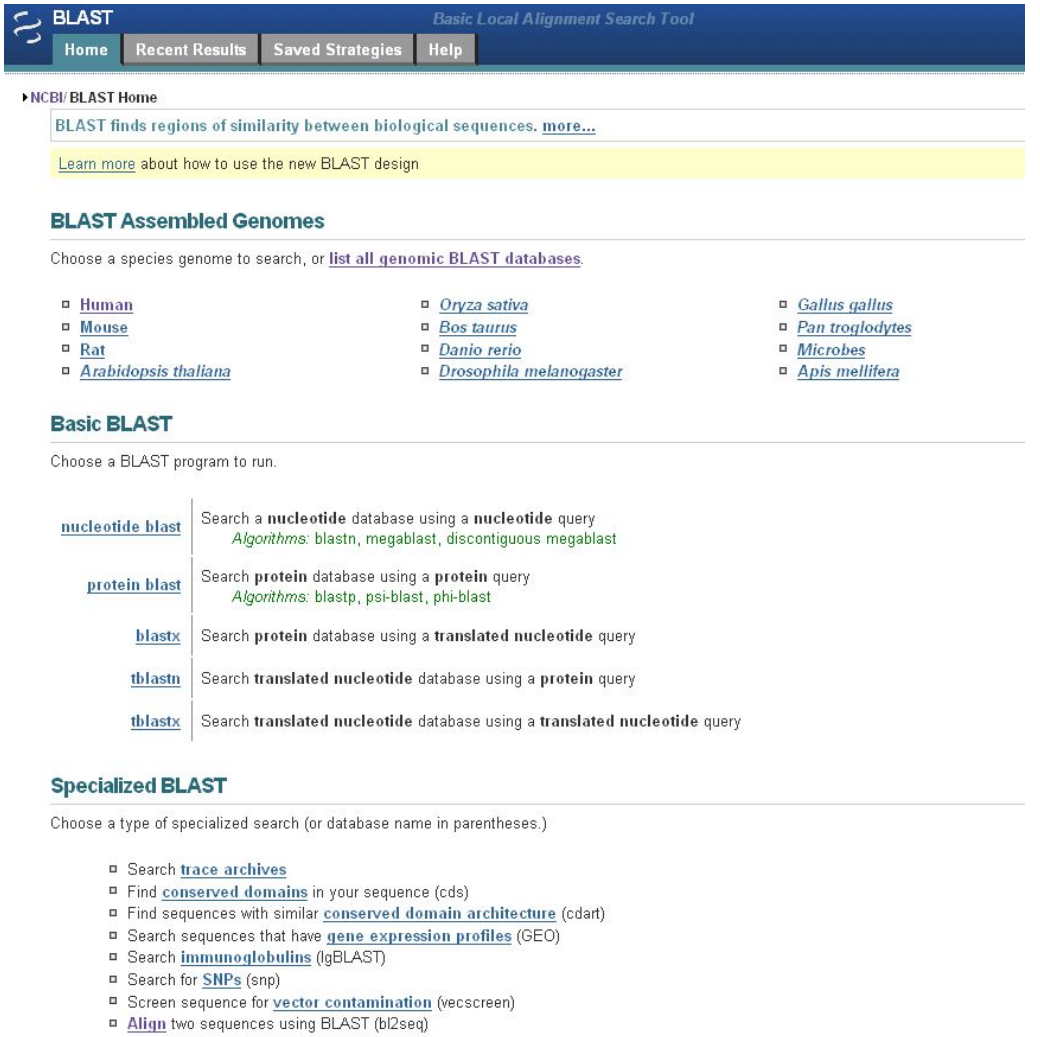
- Количество идентичных аминокислот/нуклеотидов (**Ident, %**)
- Протяженность выравнивания (**Query cover**)
- **Общая мера сходства**, или **вес выравнивания (Score)**
- **Вероятность случайного сходства** между последовательностями **(E-value)**

BLAST – Basic Local Alignment and Search Tool

- **Набор алгоритмов** для выравнивания
- **Локальное выравнивание**
- Главная задача – поиск **похожих последовательностей в базах данных** (главное достоинство – скорость)
- **Основная программа поиска** по БД
- Работа с BLAST предполагает **выбор программы** (зависит от поставленной задачи) и **алгоритма поиска** последовательностей

Родной BLAST – NCBI

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi



Программы BLAST

Программа	Описание
blastp	Сравнивает исходную аминокислотную последовательность с последовательностями из базы данных белков
blastn	Сравнивает исходную нуклеотидную последовательность с последовательностями из базы данных нуклеотидных последовательностей
blastx	Сравнивает исходную нуклеотидную последовательность, оттранслированную в аминокислотную по всем шести рамкам считывания, с последовательностями из базы данных белков..
tblastn	Сравнивает исходную аминокислотную последовательность с базой данных нуклеотидных последовательностей, динамически транслируемых по всем шести рамкам считывания
tblastx	Сравнивает все шесть трансляций исходной нуклеотидной последовательности со всеми шестью трансляциями из базы данных нуклеотидных последовательностей.

Алгоритмы поиска

- ✔ Нуклеотидные последовательности:
  - **megaBLAST** – алгоритм для сравнения ДНК. Оптимизирован для длинных похожих последовательностей. Оптимален для поиска совпадений в очень близких видах
  - **Discontiguous megaBLAST** – аналогично, параметры подобраны для более далеких видов
- ✔ Аминокислотные последовательности:
  - **PSI-BLAST (Position-Specific Iterated -BLAST)** поиск удаленных белковых гомологов
  - **PHI-BLAST (Pattern-Hit Initiated -BLAST)** ищет гомологичные белки, удовлетворяющие заданному шаблону (паттерну)

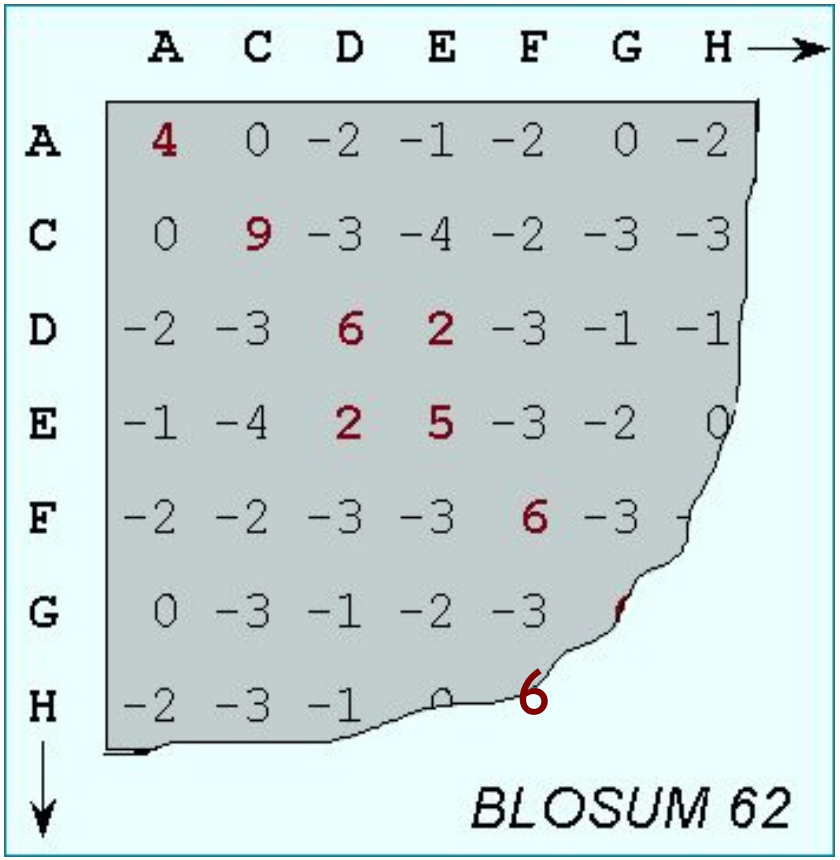
Как считается вес (**score, S**)

- Качество каждого попарного выравнивания представлено **в виде веса**,
- Чем **выше** значение – тем **лучше** результат!
- Вес выравнивания рассчитывается как **сумма баллов** совпадений/замен и пропусков
- Для вычисления **веса замен** используются **матрицы весов** (**PAM, BLOSUM**). Вес считается по каждому выровненному основанию (ДНК) или аминокислоте (белок).
- Вес пропусков назначается в виде **штрафов** за делеции и вставку пробелов

Матрицы весов

- Матрицы весов (замен) **20x20** используются **для аминокислотных выравниваний**

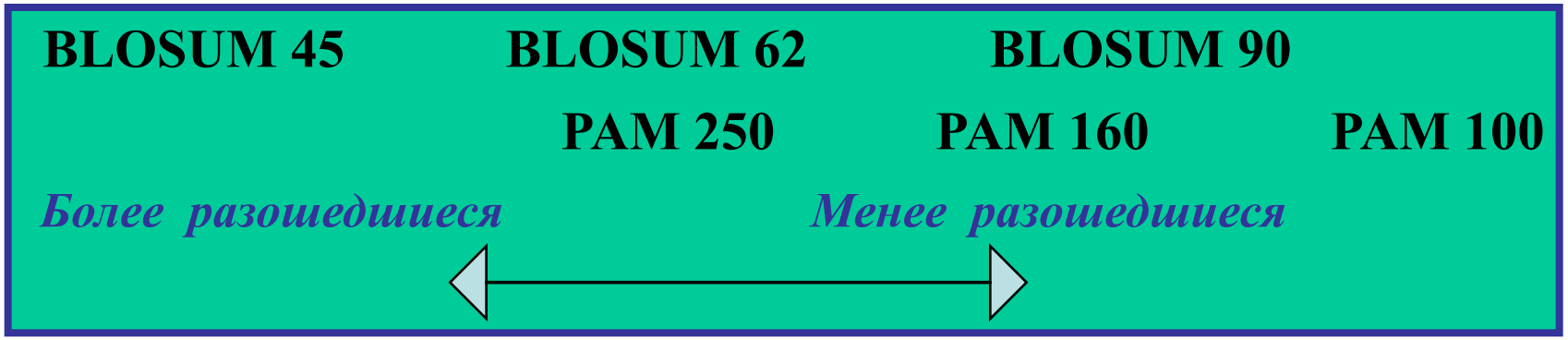
- Более простая матрица **4x4 используется для ДНК-выравнивания** (**+1** для совпадения, **-2** для несовпадения)



BLOSUM vs PAM

- **PAM (Point Accepted Mutations)** – выравнивание **очень близких** (родственных) белков

- **BLOSUM (BLOck Scoring Matrix)** – выравнивание **далеких** белков (BLOSUM62 – для белков со средним уровнем сходства, используется по умолчанию)



E-values

- ✔ Показывает вероятность **случайного** сходства, т.е. **отсутствия родственной связи** (чем **выше** значение, тем ***хуже*** результат!)
- ✔ **Низкие значения E-values означают, что последовательности гомологичны**
  - Однако, высокие значения необязательно означают негомологичность!
- ✔ **Значение зависит как от размера выровненного участка, так и от размера базы данных**
  - E-value увеличивается с увеличением размера базы данных
  - E-value уменьшается с увеличением размера участка выравнивания

Применимость критериев BLAST

- Для поиска в базах данных **нуклеотидных** последовательностей надо рассматривать результаты со значениями вероятностей **(E-values) меньше 10^{-6}** и процентом идентичности последовательностей **Ident = 70% или более**

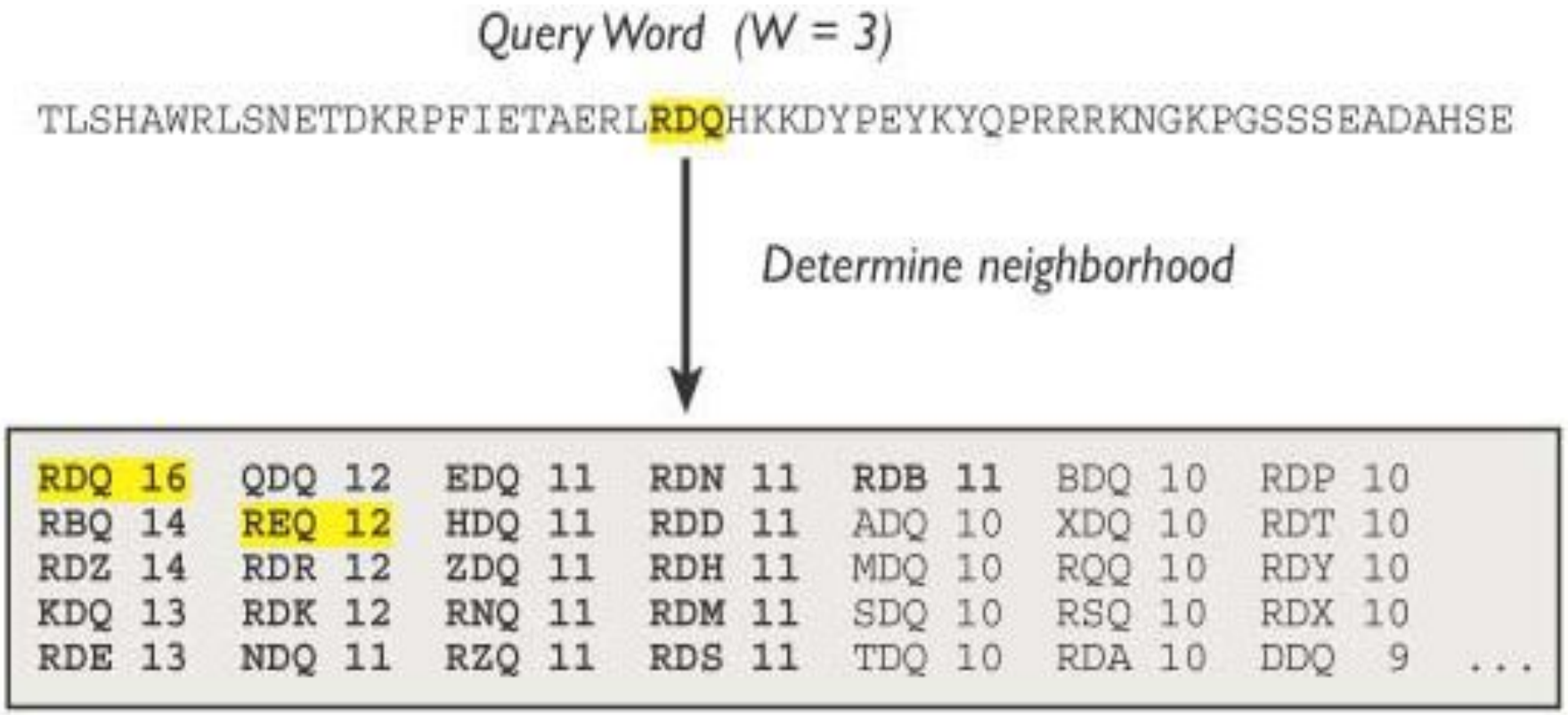
- Для поиска в базах данных **аминокислотных последовательностей** надо рассматривать результаты со значениями вероятностей **(E-values) меньше 10^{-3}** и процентом идентичности последовательностей **Ident = 25% или более**

Как работает BLAST?

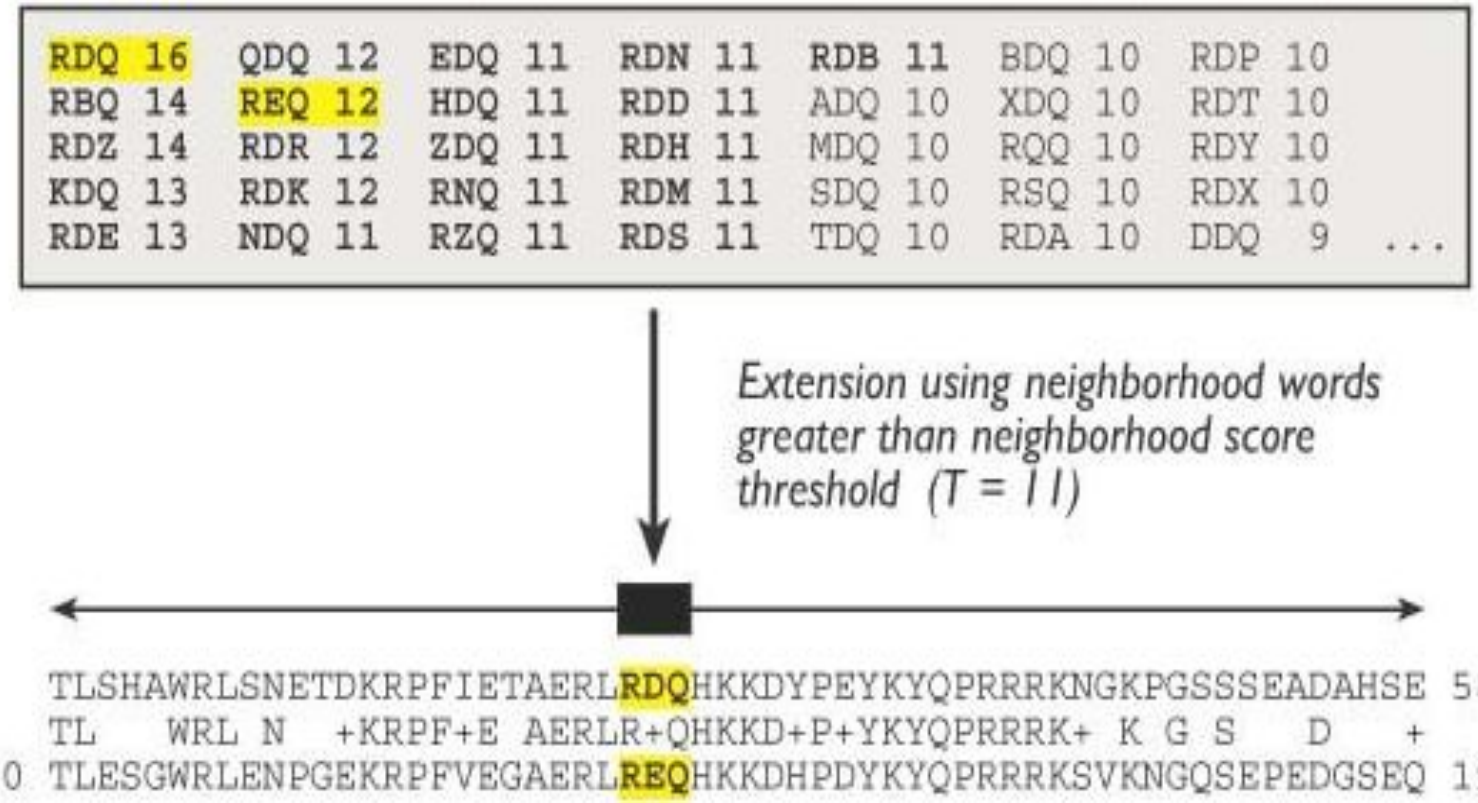
- ✔ Качество и высокая скорость поиска программ BLAST достигается с помощью подхода, при котором исходная последовательность и последовательности базы данных **разбиваются на фрагменты** (**слова, "words"**), и первоначальный поиск совпадений производится **между фрагментами.**

- ✔ После изначального нахождения совпадающих “слов” выравнивание продолжается (**вставки пробелов, инсерции, делеции, замены**) с целью сгенерировать результат с некоторым **весом S и значением E-value**

Как работает BLAST?

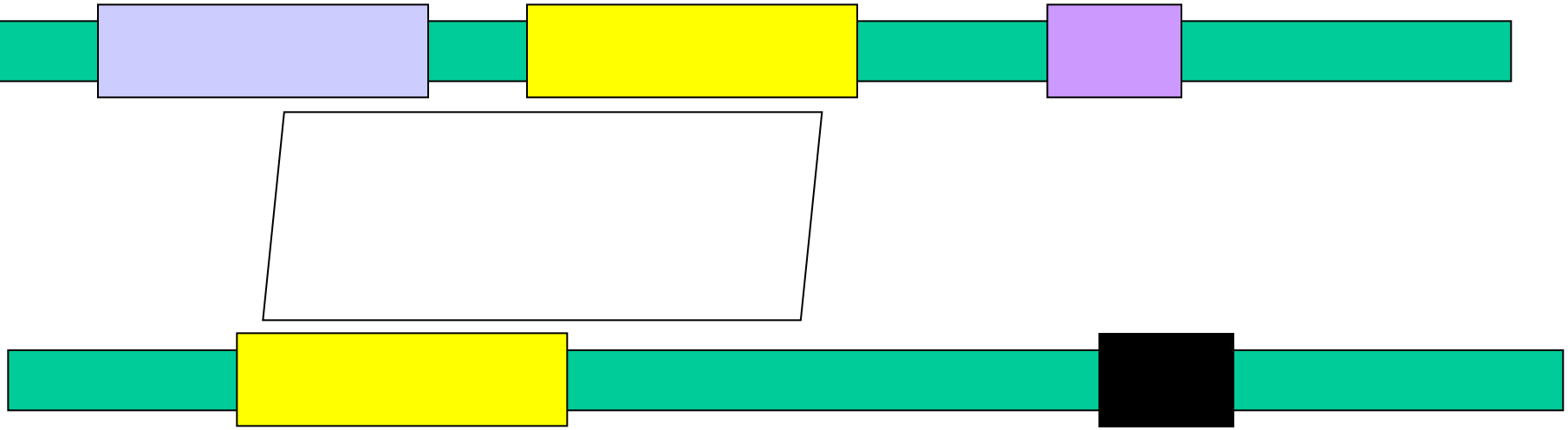


Как работает BLAST?



Результат - локальное выравнивание

- В результате BLAST выдает набор **локальных выравниваний** между исходной последовательностью и различными найденными совпадениями



Благодарю за внимание!