

Поиск сходных последовательностей. Выравнивание

```
          *           20           *
MTA1_YEAST : ----KSSISPCARAFLEQVFRRK---QSLNS : 24
MAT2_YEAST : KPYRGHRFTKENVRILESWFAKNIENPYLDT : 31
          3 2      L E   F 4      L13

          40           *           60
MTA1_YEAST : KEKEVAKKCGITPLQVRVWFINKRMRSK- : 53
MAT2_YEAST : KGLENIMNTSLSRIQIKNWVSNRRRKEKT : 61
          K  E 6  K      63 6Q64 W  N4R 4  K
```

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

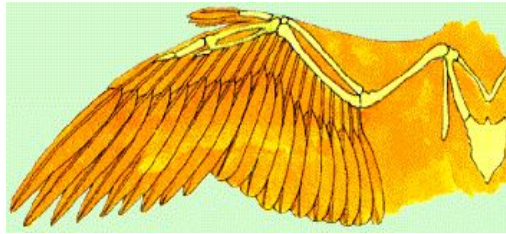
Известно, что:

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**

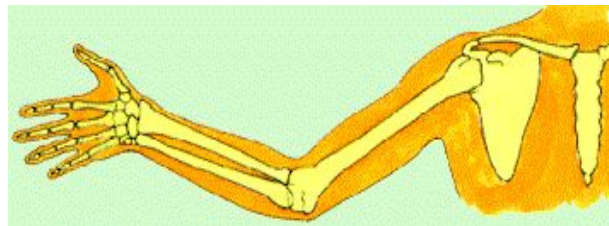
Гомология



Крыло птицы



Крыло летучей
мыши



Рука человека

- ✓ **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)

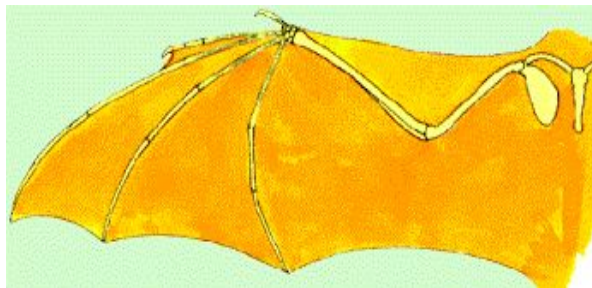
КЭ – развитие **сходных признаков** у **различных организмов**, живущих в **сходных условиях обитания**



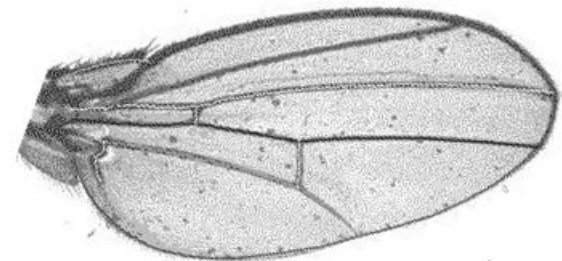
крыло птицы



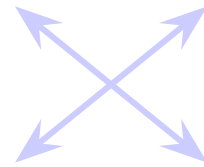
крыло бабочки



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

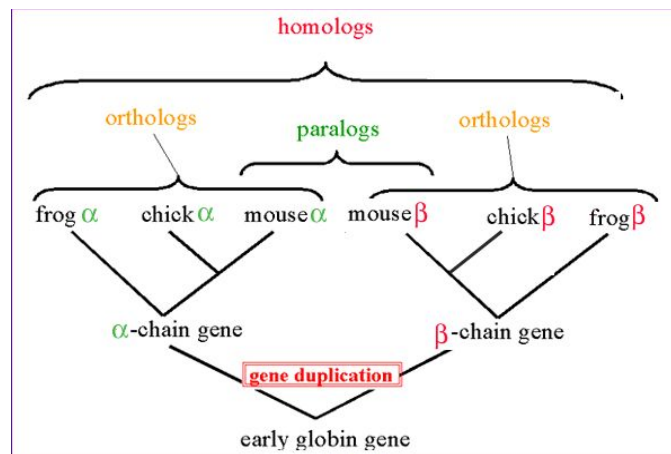
Гомология: некоторые соображения

- 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
СКУПИ**ДОМ**
ВО**ДОМ**ерка

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

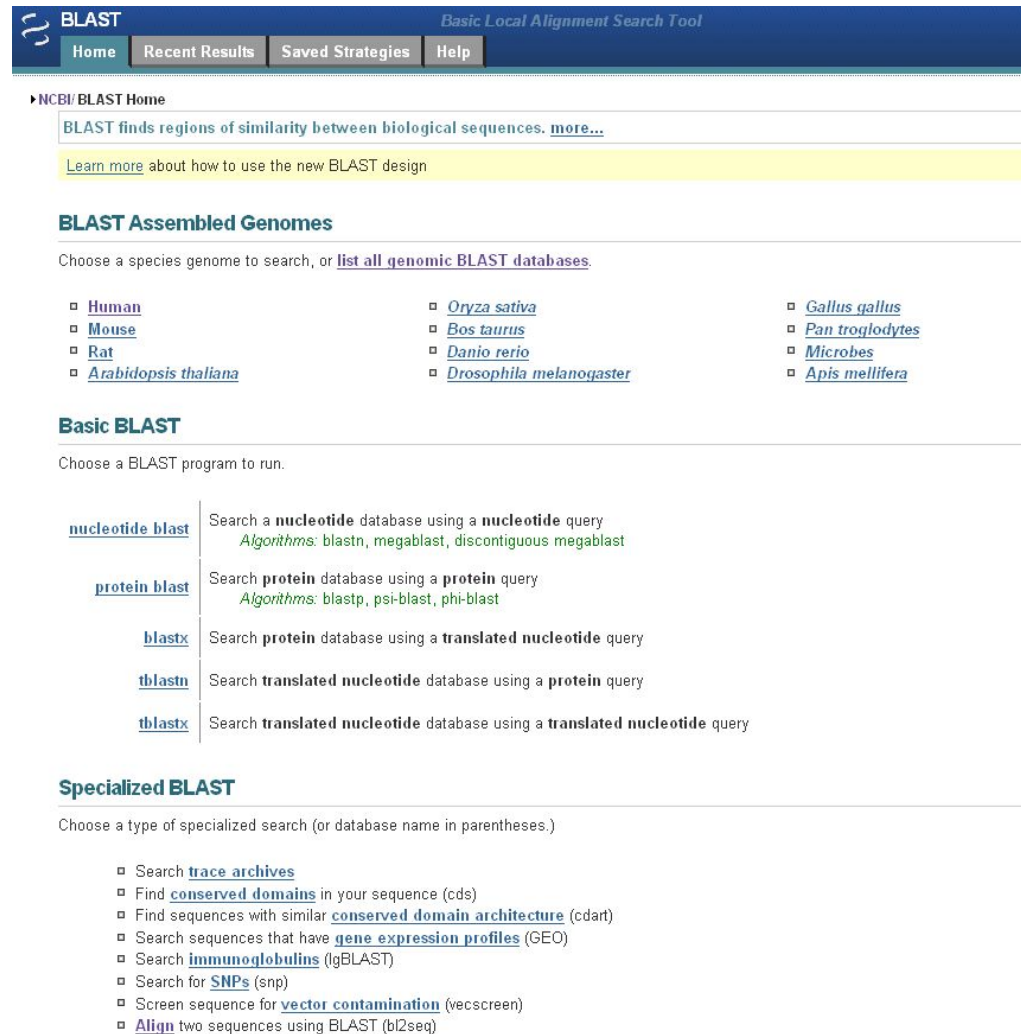
- ✓ Количество идентичных аминокислот/нуклеотидов (**Ident, %**)
- ✓ Протяженность выравнивания (**Query cover**)
- ✓ **Общая мера сходства, или вес выравнивания (Score)**
- ✓ **Вероятность случайного сходства** между последовательностями (**E-value**)

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>



The screenshot shows the NCBI BLAST website interface. At the top, there is a dark blue header with the NCBI logo on the left, the text "BLAST Basic Local Alignment Search Tool" in the center, and navigation buttons for "Home", "Recent Results", "Saved Strategies", and "Help". Below the header, the page is titled "NCBI/BLAST Home" and contains a brief description: "BLAST finds regions of similarity between biological sequences. [more...](#)". A yellow banner below this says "Learn more about how to use the new BLAST design".

The next section is "BLAST Assembled Genomes", which prompts the user to "Choose a species genome to search, or [list all genomic BLAST databases.](#)". It features a grid of checkboxes for various species: Human, Mouse, Rat, Arabidopsis thaliana, Oryza sativa, Bos taurus, Danio rerio, Drosophila melanogaster, Gallus gallus, Pan troglodytes, Microbes, and Apis mellifera.

The "Basic BLAST" section asks the user to "Choose a BLAST program to run." and lists several options with descriptions and algorithms:

- nucleotide blast**: Search a **nucleotide** database using a **nucleotide** query. Algorithms: blastn, megablast, discontinuous megablast.
- protein blast**: Search **protein** database using a **protein** query. Algorithms: blastp, psi-blast, phi-blast.
- blastx**: Search **protein** database using a **translated nucleotide** query.
- tblastn**: Search **translated nucleotide** database using a **protein** query.
- tblastx**: Search **translated nucleotide** database using a **translated nucleotide** query.

The "Specialized BLAST" section prompts the user to "Choose a type of specialized search (or database name in parentheses.)" and lists several options:

- Search **trace archives**
- Find **conserved domains** in your sequence (cds)
- Find sequences with similar **conserved domain architecture** (cdart)
- Search sequences that have **gene expression profiles** (GEO)
- Search **immunoglobulins** (IgBLAST)
- Search for **SNPs** (snp)
- Screen sequence for **vector contamination** (vecscreen)
- Align** two sequences using BLAST (bl2seq)

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)

	A	C	D	E	F	G	H	→
A	4	0	-2	-1	-2	0	-2	
C	0	9	-3	-4	-2	-3	-3	
D	-2	-3	6	2	-3	-1	-1	
E	-1	-4	2	5	-3	-2	0	
F	-2	-2	-3	-3	6	-3		
G	0	-3	-1	-2	-3			
H	-2	-3	-1	0	6			

↓

BLOSUM 62

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)

BLOSUM 45

BLOSUM 62

BLOSUM 90

PAM 250

PAM 160

PAM 100

Более разошедшиеся

Менее разошедшиеся



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?

Query Word ($W = 3$)

TLSHAWRLSNETDKRPFIEAERL**RDQ**HKKDYPEYKYQPRRRKNGKPGSSSEADAHSE

Determine neighborhood

RDQ 16	QDQ 12	EDQ 11	RDN 11	RDB 11	BDQ 10	RDP 10
RBQ 14	REQ 12	HDQ 11	RDD 11	ADQ 10	XDQ 10	RDT 10
RDZ 14	RDR 12	ZDQ 11	RDH 11	MDQ 10	RQQ 10	RDY 10
KDQ 13	RDK 12	RNQ 11	RDM 11	SDQ 10	RSQ 10	RDX 10
RDE 13	NDQ 11	RZQ 11	RDS 11	TDQ 10	RDA 10	DDQ 9 ...

Как работает BLAST?

RDQ 16	QDQ 12	EDQ 11	RDN 11	RDB 11	BDQ 10	RDP 10	
RBQ 14	REQ 12	HDQ 11	RDD 11	ADQ 10	XDQ 10	RDT 10	
RDZ 14	RDR 12	ZDQ 11	RDH 11	MDQ 10	RQQ 10	RDY 10	
KDQ 13	RDK 12	RNQ 11	RDM 11	SDQ 10	RSQ 10	RDX 10	
RDE 13	NDQ 11	RZQ 11	RDS 11	TDQ 10	RDA 10	DDQ 9	...

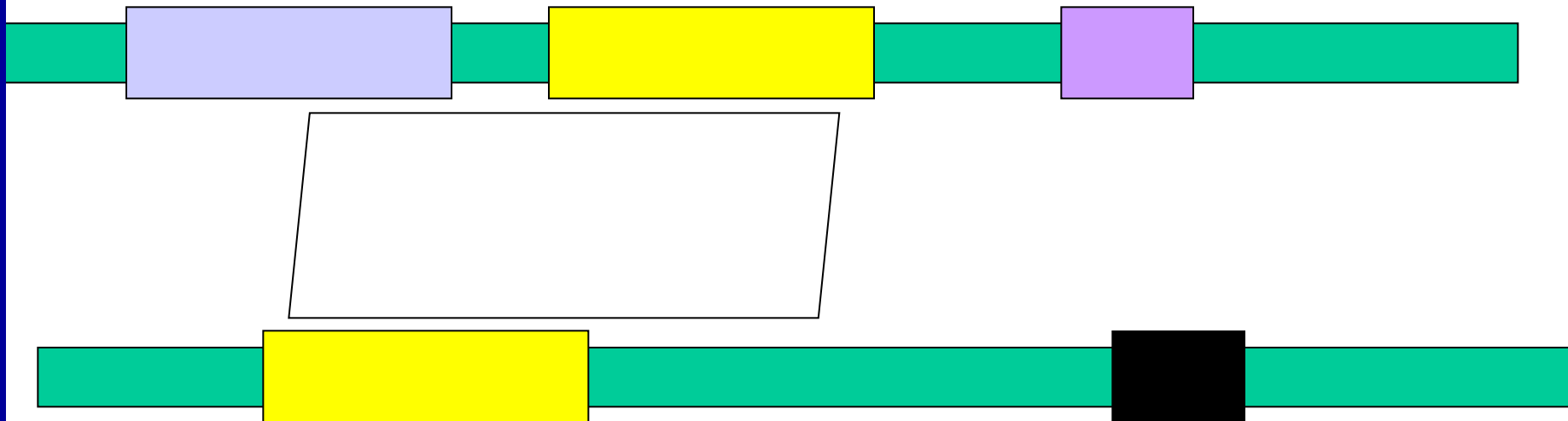
*Extension using neighborhood words
greater than neighborhood score
threshold ($T = 11$)*



Query: 1 TL SHAWRLSNETDKRPFIEAERL**RDQ**HKKDYPEYKYQPRRRKNGKPGSSSEADAHSE 58
TL WRL N +KRPF+E AERLR+QHKKD+P+YKYQPRRRK+ K G S D +
Sbjct: 140 TLESGWRLNPGKRPFVEGAERL**REQ**HKKDHPDYKYQPRRRKSVKNGQSEPEDGSEQ 197

Результат - локальное выравнивание

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



Благодарю за внимание!