

# С прошлой лекции

- HGNC - [HUGO Gene Nomenclature Committee](#)
- *Ensembl*:
  - *EnsemblProtists*  
(<http://protists.ensembl.org/index.html>)
  - *EnsemblBacteria*  
<http://bacteria.ensembl.org/index.html>
  - *EnsemblMetazoa* (Инсекты и нематоды)  
<http://metazoa.ensembl.org/index.html>)
  - *EnsemblPlants* (<http://plants.ensembl.org/index.html>)
  - *EnsemblFungi* (<http://fungi.ensembl.org/index.html>)

# DUT ген человека

## 1: DUT dUTP pyrophosphatase [ *Homo sapiens* ]

GeneID: 1854

updated 05-Nov-2007

[Entrez Gene Home](#)

### Summary



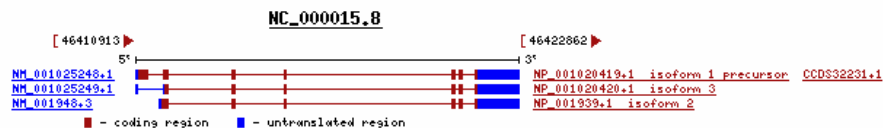
<b>Official Symbol</b>	DUT	provided by <a href="#">HGNC</a>
<b>Official Full Name</b>	dUTP pyrophosphatase	provided by <a href="#">HGNC</a>
<b>Primary source</b>	<a href="#">HGNC:3078</a>	
<b>See related</b>	<a href="#">Ensembl:ENSG00000128951</a> ; <a href="#">HPRD:03165</a> ; <a href="#">MIM:601266</a>	
<b>Gene type</b>	protein coding	
<b>RefSeq status</b>	Reviewed	
<b>Organism</b>	<a href="#">Homo sapiens</a>	
<b>Lineage</b>	<i>Eukaryota</i> ; <i>Metazoa</i> ; <i>Chordata</i> ; <i>Craniata</i> ; <i>Vertebrata</i> ; <i>Euteleostomi</i> ; <i>Mammalia</i> ; <i>Eutheria</i> ; <i>Euarchontoglires</i> ; <i>Primates</i> ; <i>Haplorrhini</i> ; <i>Catarrhini</i> ; <i>Hominidae</i> ; <i>Homo</i>	
<b>Also known as</b>	dUTPase; FLJ20622	

**Summary** This gene encodes an essential enzyme of nucleotide metabolism. The encoded protein forms a ubiquitous, homotetrameric enzyme that hydrolyzes dUTP to dUMP and pyrophosphate. This reaction serves two cellular purposes: providing a precursor (dUMP) for the synthesis of thymine nucleotides needed for DNA replication, and limiting intracellular pools of dUTP. Elevated levels of dUTP lead to increased incorporation of uracil into DNA, which induces extensive excision repair mediated by uracil glycosylase. This repair process, resulting in the removal and reincorporation of dUTP, is self-defeating and leads to DNA fragmentation and cell death. Alternative splicing of this gene leads to different isoforms that localize to either the mitochondrion or nucleus. A related pseudogene is located on chromosome 19.

### Genomic regions, transcripts, and products



Go to [reference sequence details](#)

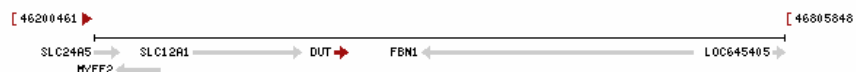


### Genomic context



chromosome: 15; Location: 15q15-q21.1

[See DUT in MapViewer](#)



### Table Of Contents

- Summary
- Genomic regions, transcripts...
- Genomic context
- Bibliography
- Interactions
- General gene information
- General protein information
- Reference Sequences
- Related Sequences
- Additional Links

### Links

- Order cDNA clone
- Conserved Domains
- Genome
- GEO Profiles
- HomoloGene
- Map Viewer
- CoreNucleotide
- EST
- Nucleotide
- OMIM
- Full text in PMC
- Probe
- Protein
- PubMed
- PubMed (GeneRIF)
- SNP
- SNP: Genotype
- SNP: GeneView
- Taxonomy
- UniSTS
- AceView
- CCDS
- Ensembl
- Evidence Viewer
- GDB
- HGNC
- HPRD
- KEGG
- MGC
- ModelMaker
- PharmGKB
- Reactome
- UniGene
- LinkOut

### Entrez Gene Info

### Feedback

### Subscriptions

# BLAST

- Что такое выравнивание
- Выравнивание 2х последовательностей
- BLAST на NCBI:
  - Что это такое
  - Как выбрать правильную программу
  - Как выбрать правильную базу данных
  - Как запустить
  - Как интерпретировать результаты

# Почему нас интересует локальное сходство последовательностей?

Мы верим, 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

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

## **Признаки гомологичности белков**

- **сходная 3D-структура**
- **в той или иной степени похожая  
аминокислотная последовательность**
- **аналогичная функция**
- **разные другие соображения...**

# ГОМОЛОГИ

Ортологи

Паралоги

Ксенологи ?

homologs

orthologs

paralogs

orthologs

frog  $\alpha$  chick  $\alpha$  mouse  $\alpha$  mouse  $\beta$  chick  $\beta$  frog  $\beta$

$\alpha$ -chain gene

$\beta$ -chain gene

gene duplication

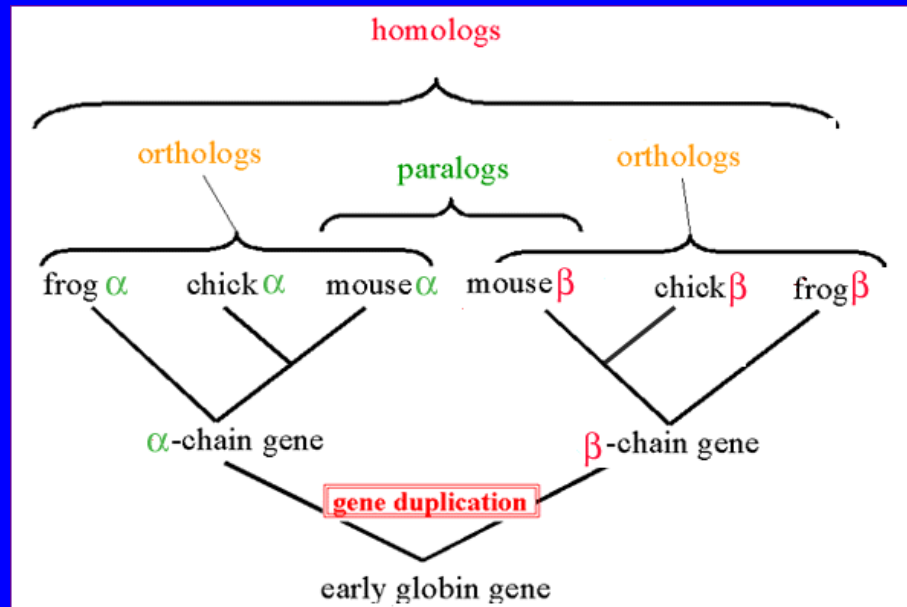
early globin gene



(W.M.Fitch, *Syst.Zool.*19,99(1970))

**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 одного гена в одном организме. Паралоги, как правило, имеют разные функции.



# Средство поиска сходства - выравнивание

«Идеальное» выравнивание – запись последовательностей одна под другой так, чтобы гомологичные фрагменты оказались друг под другом.

домовой  
скупидом  
водомерка

лесовоз

ледоход

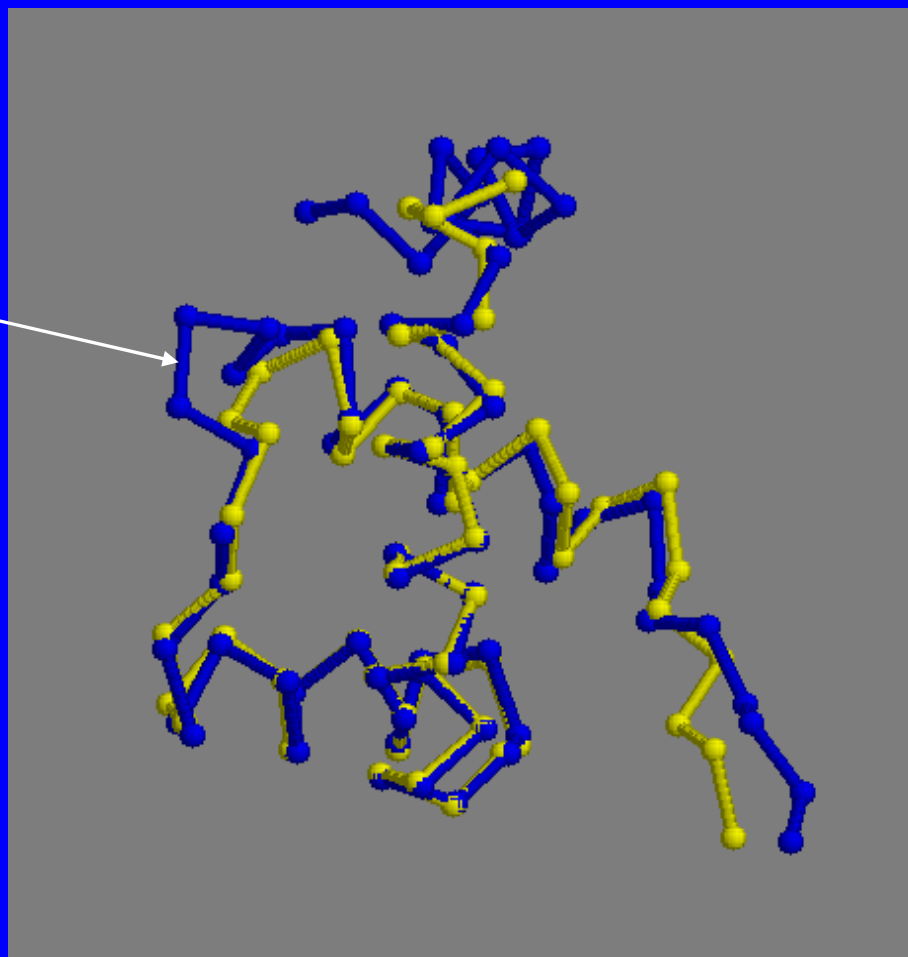
---лесо---воз

лед---оход---



# Схожие 3D структуры

Вставка в «синей»  
последовательности



# Как выровнять 2 последовательности?

Цель - максимальное количество  
совпадений

- Просто написать их друг под другом
- Двигать друг относительно друга
- Вставлять пробелы
- Что лучше?

лесовоз  
ледоход

---лесо---воз  
лед---оход---

*Гэп – пропуск в  
последовательности*

# Матрицы замен

Матрица 20\*20 на пересечении 2х аа их уровень сходства (?):

- Похожесть по свойствам (объем, гидрофильность, заряд и т.д.)
- Эволюционное родство – частота замен 1ой аа на другую в изученных белках

2 сорта последних:

PAM (Point Accepted Mutations) – на выравниваниях очень близких белков (PAM20 = PAM<sup>20</sup>)

BLOSUM (BLOck Scoring Matrix) – на блоках выравниваний далеких белков (без делеций) (BLOSUM62 – на белках со средним уровнем сходства 62% попарно)

# Делеции/инсерции

- ✓ Общий штраф
- ✓ Значительно чаще 1 длинная делеция, чем много коротких => штраф за внесение делеции + штраф за удлинение делеции

# Типы выравнивания

- ✓ Локальное – поиск фрагментов наиболее похожих друг на друга

домовой скупидом      домовой водомерка      домовой водомерка

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

лесовоз  
ледоход

?

---лесо---воз  
лед---оход---

# Критерии качества выравнивания

- ✓ Количество идентичных (похожих) аминокислот/нуклеотидов
  - Для белков – более 25% id при длине  $> 100$  aa
  - Для ДНК – более 70% id при длине  $> 100$  nt
- ✓ Длина выравнивания
- ✓ Вероятность наблюдать такое сходство случайным образом
  - Зависит от базы данных
- ✓ Score – общая мера сходства:
  - Зависит от программы

# BLAST – Basic Local Alignment and Search Tool

- ✓ Локальное выравнивание
- ✓ Главная задача – поиск похожих последовательностей в базах данных (=> главное достоинство – скорость)
- ✓ Очень неточно восстанавливает сходство
- ✓ Основная программа поиска по БД
- ✓ Для специализированных БД часто предлагается на сайте БД
- ✓ Для поиска среди известных последовательностей есть специальные сервера

# 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 navigation bar with the BLAST logo and the text "Basic Local Alignment Search Tool". Below the navigation bar, there are tabs for "Home", "Recent Results", "Saved Strategies", and "Help". The main content area is titled "NCBI/BLAST Home" and contains a description of BLAST: "BLAST finds regions of similarity between biological sequences. [more...](#)". There is a link to "Learn more about how to use the new BLAST design". Below this, there is a section titled "BLAST Assembled Genomes" with the instruction "Choose a species genome to search, or [list all genomic BLAST databases](#)". A list of species is provided, including Human, Mouse, Rat, Arabidopsis thaliana, Oryza sativa, Bos taurus, Danio rerio, Drosophila melanogaster, Gallus gallus, Pan troglodytes, Microbes, and Apis mellifera. The next section is "Basic BLAST" with the instruction "Choose a BLAST program to run." and a list of programs: nucleotide blast, protein blast, blastx, tblastn, and tblastx, each with a brief description and the algorithms used. The final section is "Specialized BLAST" with the instruction "Choose a type of specialized search (or database name in parentheses.)" and a list of specialized search options: trace archives, conserved domains, conserved domain architecture, gene expression profiles, immunoglobulins, SNPs, vector contamination, and Align two sequences using BLAST (bl2seq).

**BLAST** Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

► NCBI/BLAST Home

BLAST finds regions of similarity between biological sequences. [more...](#)

[Learn more](#) about how to use the new BLAST design

**BLAST Assembled Genomes**

Choose a species genome to search, or [list all genomic BLAST databases](#).

- Human
- Mouse
- Rat
- Arabidopsis thaliana*
- Oryza sativa*
- Bos taurus*
- Danio rerio*
- Drosophila melanogaster*
- Gallus gallus*
- Pan troglodytes*
- Microbes
- Apis mellifera*

**Basic BLAST**

Choose a BLAST program to run.

<a href="#">nucleotide blast</a>	Search a <b>nucleotide</b> database using a <b>nucleotide</b> query <i>Algorithms:</i> blastn, megablast, discontinuous megablast
<a href="#">protein blast</a>	Search <b>protein</b> database using a <b>protein</b> query <i>Algorithms:</i> blastp, psi-blast, phi-blast
<a href="#">blastx</a>	Search <b>protein</b> database using a <b>translated nucleotide</b> query
<a href="#">tblastn</a>	Search <b>translated nucleotide</b> database using a <b>protein</b> query
<a href="#">tblastx</a>	Search <b>translated nucleotide</b> database using a <b>translated nucleotide</b> query

**Specialized BLAST**

Choose a type of specialized search (or database name in parentheses.)

- 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)



# Какую программу BLAST выбрать?

Программа	Query	Тип БД	Сравнивает
Blastn	ДНК	ДНК	ДНК
Blastp	белок	белок	белки
Blastx	ДНК	белок	белки
Tblastn	белок	ДНК	белки
Tblastx	ДНК	ДНК	белки

# Дополнительные программы

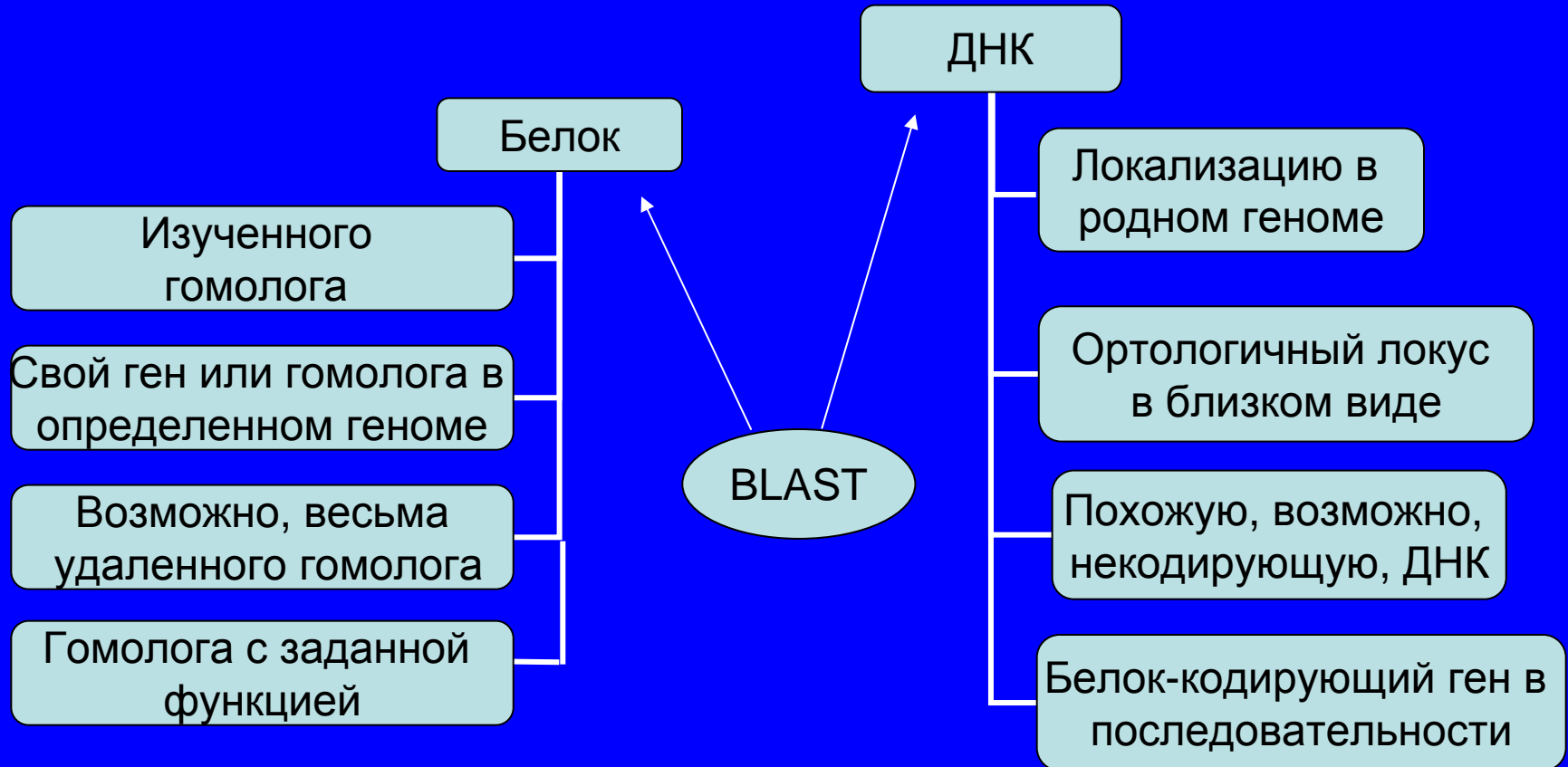
## ✓ ДНК:

- megaBLAST – другой алгоритм для сравнения ДНК. Оптимизирован для длинных похожих последовательностей. Оптимален для поиска хитов в родном геноме или очень близких видах
- Discontiguous megaBLAST – аналогично, параметры подобраны для более далеких видов

## ✓ Белок:

- PSI-BLAST (Position-Specific Iterated -BLAST) поиск удаленных белковых гомологов с использованием PSSM (position-specific scoring matrix)
- PHI-BLAST (Pattern-Hit Initiated -BLAST) ищет гомологичные белки, удовлетворяющие заданному паттерну

# Какую программу выбрать?



# Стандартный input

**BLAST** Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI Sign In | Re

NCBI/BLAST/blastp suite

blastn blastp blastx tblastn tblastx

BLASTP programs search protein databases using a protein query. [more...](#) [Reset page](#) [Bookmark](#)

Enter Query Sequence

Enter accession number, gi, or FASTA sequence  [Clear](#) Query subrange  From  To

Or, upload file  [Обзор...](#)

Job Title   
Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

**Database** Non-redundant protein sequences (nr)

**Organism**   Exclude  Exclude  Exclude  
Enter organism name or id--completions will be suggested  
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

**Exclude**  Models (XM/XP)  Environmental sample sequences

**Entrez Query**   
Enter an Entrez query to limit search

Program Selection

**Algorithm**  blastp (protein-protein BLAST)  PSI-BLAST (Position-Specific Iterated BLAST)  PHI-BLAST (Pattern Hit Initiated BLAST)  
Choose a BLAST algorithm

**BLAST** Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)  Show results in a new window

# Промежуточная страница - CD

23

BLAST

Basic Local Alignment Search Tool

Home

Recent Results

Saved Strategies

Help

► [NCBI/BLAST/blastp/Formatting Results - M381T1VP01R](#) [\[Formatting options\]](#)

Job Title: **destabilase**

Putative conserved domains have been detected, click on the image below for detailed results.



Request ID	M381T1VP01R
Status	Searching
Submitted at	Mon Nov 19 09:39:53 2007
Current time	Mon Nov 19 09:40:00 2007
Time since submission	

This page will be automatically updated in 15 seconds

# Output – I (parameters and options)

**BLAST** Basic Local Alignment Search Tool My NCBI  
[Sign In]

[Home](#) [Recent Results](#) [Saved Strategies](#) [Help](#)

NCBI/BLAST/blastp suite/ Formatting Results - G31XSVVW013

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#)

**gi|1255718|gb|AAA96144.1| destabilase I [Hirudo...**

<b>Query ID</b>	lc 22547	<b>Database Name</b>	nr
<b>Description</b>	gi 1255718 gb AAA96144.1  destabilase I [Hirudo medicinalis]	<b>Description</b>	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+excluding environmental samples from WGS projects
<b>Molecule type</b>	amino acid	<b>Program</b>	BLASTP 2.2.22+ <a href="#">Citation</a>
<b>Query Length</b>	136		

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#) [Related Structures](#) [Multiple alignment](#) **NEW**

### Graphic Summary

[Show Conserved Domains](#)

Putative conserved domains have been detected, click on the image below for detailed results.

Query seq. Superfamilies

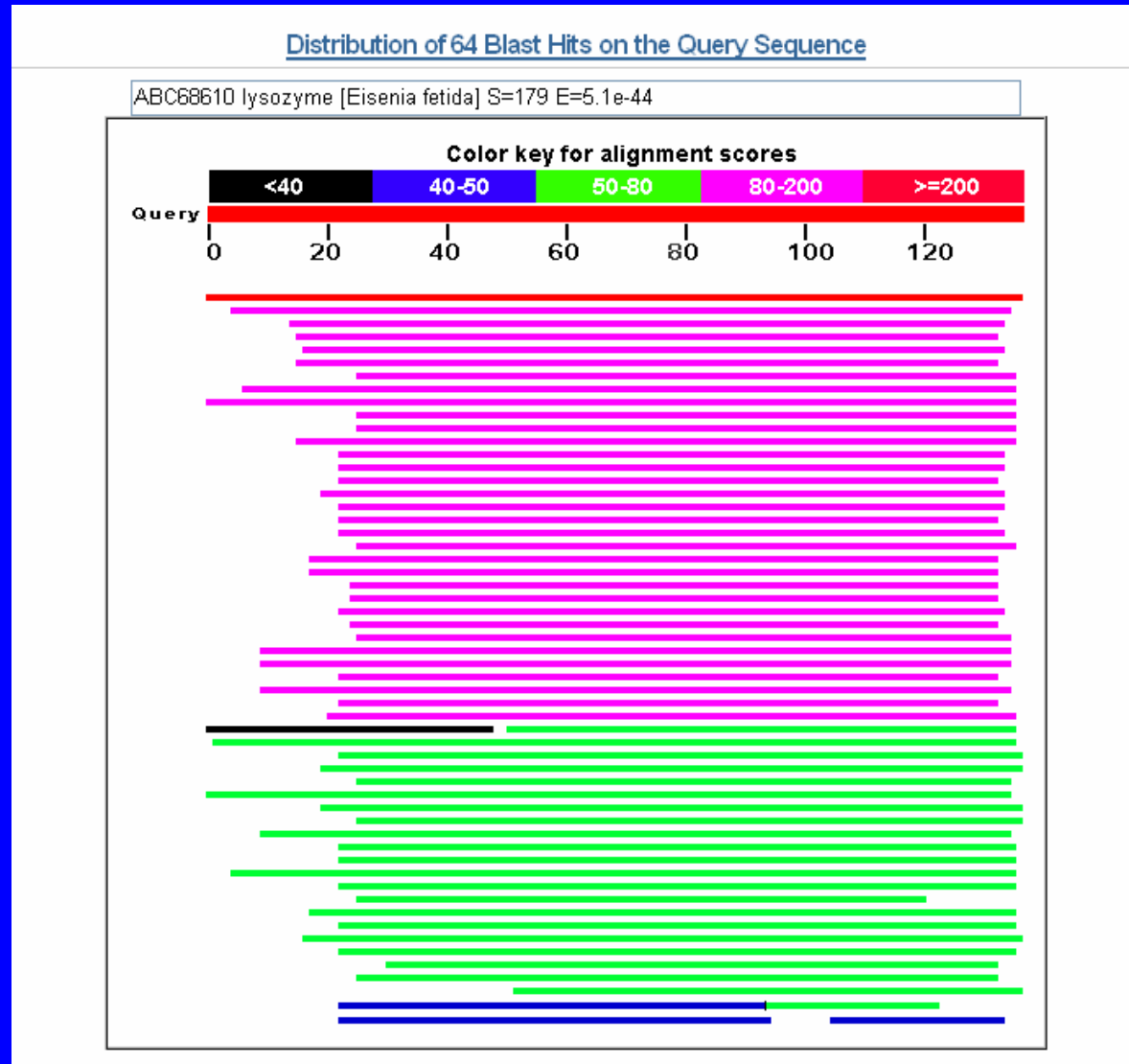
Distribution of 100 Blast Hits on the Query Sequence

Mouse-over to show define and scores, click to show alignments

**Color key for alignment scores**

<40	40-50	50-80	80-200	>=200
-----	-------	-------	--------	-------

# Output – II (color chart)



# Output – III (descriptions)

▼ Descriptions

Sequences producing significant alignments:

		Score (Bits)	E Value	
<a href="#">gb AAA96144.1 </a>	destabilase I [Hirudo medicinalis]	<a href="#">278</a>	9e-74	
<a href="#">gb ABC68610.1 </a>	lysozyme [Eisenia andrei]	<a href="#">197</a>	3e-49	
<a href="#">gb AAN16207.1 </a>	lysozyme [Mytilus edulis]	<a href="#">135</a>	1e-30	
<a href="#">gb AAN16210.1 </a>	lysozyme [Mytilus galloprovincialis]	<a href="#">134</a>	4e-30	
<a href="#">gb AAN16209.1 </a>	lysozyme [Bathymodiolus thermophilus]	<a href="#">134</a>	4e-30	
<a href="#">gb AAN16208.1 </a>	lysozyme [Bathymodiolus azoricus]	<a href="#">131</a>	3e-29	
<a href="#">ref XP_792095.1 </a>	PREDICTED: similar to lysozyme [Strongylocen...	<a href="#">127</a>	5e-28	<b>G</b>
<a href="#">gb AAR29291.1 </a>	lysozyme [Asterias rubens]	<a href="#">126</a>	6e-28	
<a href="#">gb ABK34500.1 </a>	lysozyme [Apostichopus japonicus]	<a href="#">121</a>	2e-26	
<a href="#">sp P83673.3 LYS1 CRAVI</a>	RecName: Full=Lysozyme 1; AltName: Ful...	<a href="#">120</a>	6e-26	
<a href="#">ref XP_002602594.1 </a>	hypothetical protein BRAFLDRAFT_81868 [Br...	<a href="#">117</a>	3e-25	<b>G</b>
<a href="#">gb ACU83237.1 </a>	lysozyme [Venerupis philippinarum]	<a href="#">117</a>	4e-25	
<a href="#">dbj BAC15553.1 </a>	lysozyme [Venerupis philippinarum]	<a href="#">116</a>	6e-25	
<a href="#">sp Q6L6Q6.1 LYS CRAGI</a>	RecName: Full=Lysozyme; AltName: Full=1...	<a href="#">116</a>	7e-25	
<a href="#">ref XP_788343.1 </a>	PREDICTED: similar to lysozyme [Strongylocen...	<a href="#">116</a>	9e-25	<b>UG</b>
<a href="#">ref XP_001175841.1 </a>	PREDICTED: similar to lysozyme [Strongylo...	<a href="#">116</a>	9e-25	<b>UG</b>
<a href="#">dbj BAF48045.1 </a>	lysozyme 1M [Crassostrea gigas]	<a href="#">116</a>	9e-25	
<a href="#">pdb 2DQA A</a>	Chain A, Crystal Structure Of Tapes Japonica Lysoz...	<a href="#">115</a>	1e-24	<b>S</b>
<a href="#">sp Q6L6Q5.1 LYS OSTED</a>	RecName: Full=Lysozyme; AltName: Full=1...	<a href="#">115</a>	2e-24	
<a href="#">ref XP_788380.2 </a>	PREDICTED: similar to lysozyme [Strongylocen...	<a href="#">114</a>	3e-24	<b>G</b>
<a href="#">sp Q1XG90.1 LYS2 CRAVI</a>	RecName: Full=Lysozyme 2; AltName: Ful...	<a href="#">114</a>	4e-24	
<a href="#">sp B3A003.1 LYS3 CRAVI</a>	RecName: Full=Lysozyme 3; AltName: Ful...	<a href="#">113</a>	8e-24	
<a href="#">ref XP_002593923.1 </a>	hypothetical protein BRAFLDRAFT_234809 [B...	<a href="#">112</a>	1e-23	<b>G</b>
<a href="#">dbj BAE33389.1 </a>	lysozyme [Venerupis philippinarum]	<a href="#">112</a>	1e-23	
<a href="#">ref XP_002593924.1 </a>	hypothetical protein BRAFLDRAFT_98223 [Br...	<a href="#">112</a>	1e-23	<b>G</b>
<a href="#">dbj BAF63423.1 </a>	lysozyme 2 [Mytilus galloprovincialis]	<a href="#">111</a>	3e-23	
<a href="#">emb CAB63451.1 </a>	chlamysin [Chlamys islandica]	<a href="#">109</a>	1e-22	
<a href="#">gb AAN16211.1 </a>	lysozyme [Calyptogena sp. SB2001_1]	<a href="#">108</a>	2e-22	
<a href="#">emb CAC34834.1 </a>	lysozyme [Chlamys islandica]	<a href="#">107</a>	3e-22	
<a href="#">ref NP_500206.1 </a>	Invertebrate LYSozyme family member (ily-3)...	<a href="#">106</a>	7e-22	<b>UG</b>
<a href="#">ref NP_500207.1 </a>	Invertebrate LYSozyme family member (ily-2)...	<a href="#">105</a>	2e-21	<b>UG</b>
<a href="#">gb AAN16212.1 </a>	lysozyme [Calyptogena sp. SB2001_2]	<a href="#">104</a>	3e-21	
<a href="#">ref XP_788357.1 </a>	PREDICTED: similar to lysozyme [Strongylocen...	<a href="#">103</a>	4e-21	<b>G</b>
<a href="#">ref XP_001676933.1 </a>	Hypothetical protein CBG10836 [Caenorhabd...	<a href="#">103</a>	4e-21	<b>G</b>
<a href="#">dbj BAF94156.1 </a>	lysozyme 3 [Crassostrea gigas]	<a href="#">103</a>	5e-21	
<a href="#">ref NP_001024594.1 </a>	Invertebrate LYSozyme family member (ily...	<a href="#">100</a>	5e-20	<b>UG</b>
<a href="#">dbj BAF48044.2 </a>	lysozyme 2 [Crassostrea gigas]	<a href="#">99.4</a>	1e-19	
<a href="#">gb ABB76765.1 </a>	lysozyme [Mytilus edulis]	<a href="#">93.2</a>	7e-18	
<a href="#">ref XP_002410814.1 </a>	lysozyme, putative [Ixodes scapularis] >g...	<a href="#">92.4</a>	1e-17	<b>UG</b>
<a href="#">gb ABD65298.1 </a>	destabilase I [Litopenaeus vannamei]	<a href="#">87.4</a>	4e-16	
<a href="#">ref XP_001869584.1 </a>	conserved hypothetical protein [Culex qui...	<a href="#">76.3</a>	9e-13	<b>UG</b>
<a href="#">ref XP_001653399.1 </a>	hypothetical protein AaeL_AAEL001485 [Aed...	<a href="#">74.7</a>	2e-12	<b>UG</b>



# Output - IV (alignments)

▼ **Alignments**  Select All [Get selected sequences](#) [Distance tree of results](#) [Multiple alignment](#) **NEW**

> [gb|AAA96144.1](#) destabilase I [Hirudo medicinalis]  
Length=136

Score = 278 bits (712), Expect = 9e-74, Method: Compositional matrix adjust.  
Identities = 136/136 (100%), Positives = 136/136 (100%), Gaps = 0/136 (0%)

Query	1	MIIAIYVSLALLIASVEVNSQFTDSCLRICKVEGCDSSQIGKCGMDVGSLSGPGYQIKKP	60
		MIIAIYVSLALLIASVEVNSQFTDSCLRICKVEGCDSSQIGKCGMDVGSLSGPGYQIKKP	
Sbjct	1	MIIAIYVSLALLIASVEVNSQFTDSCLRICKVEGCDSSQIGKCGMDVGSLSGPGYQIKKP	60
Query	61	YWIDCGKPGGGYESTKKNKACSETCVRAYMKRYGTFCTGGRTPTCQDYARIHNGGPRGCK	120
		YWIDCGKPGGGYESTKKNKACSETCVRAYMKRYGTFCTGGRTPTCQDYARIHNGGPRGCK	
Sbjct	61	YWIDCGKPGGGYESTKKNKACSETCVRAYMKRYGTFCTGGRTPTCQDYARIHNGGPRGCK	120
Query	121	SSATVGYWNVQKCLR	136
		SSATVGYWNVQKCLR	
Sbjct	121	SSATVGYWNVQKCLR	136

> [gb|ABC68610.1](#) lysozyme [Eisenia andrei]  
Length=160

Score = 197 bits (501), Expect = 3e-49, Method: Compositional matrix adjust.  
Identities = 88/132 (66%), Positives = 114/132 (86%), Gaps = 2/132 (1%)

Query	3	IAIYVSLALLIASVEVNSQFTDSCLRICKVEGCDSSQIGKCGMDVGSLSGPGYQIKKPYW	62
		+ IY +L+ ++A+ +0 +++CL CIC++EGC+SQIGKC MDVGSLSGPG+QIK+PYW	
Sbjct	1	MFIYFALSCILATAA--AQISENCLNCICQIEGCESSQIGKCRMDVGSLSGPGFQIKPEPYW	58
Query	63	IDCGKPGGGYESTKKNKACSETCVRAYMKRYGTFCTGGRTPTCQDYARIHNGGPRGCKSS	122
		IDCG+PGG ++SCT CS TCVR+YMKRYGT+CTGGR PTCQDYARIHNGGPG+GC+ +	
Sbjct	59	IDCGRPGGDWKSCTTQMDCSRTCVRSYMRYGTYCTGGRAPTCQDYARIHNGGPKGCQHA	118
Query	123	ATVGYWNVQK	134
		+TVGYWNV+C	
Sbjct	119	STVGYWNVKQ	130

> [gb|AAN16207.1](#) lysozyme [Mytilus edulis]  
Length=176

Score = 135 bits (340), Expect = 1e-30, Method: Compositional matrix adjust.  
Identities = 64/120 (53%), Positives = 85/120 (70%), Gaps = 4/120 (3%)

Query	15	SVEVNSQFTDSCLRICKVEG-CDSQIGKCGMDVGSLSGPGYQIKKPYWIDCGKPGGGYE	73
		S++ N +D C+RCIC VE C++ IG C MDVGSLSGPG+QIKK YWIDCG+P G Y+	
Sbjct	53	SIDSNGLVSDKCMRCICMVESHCMNNIG-CRMDVGSLSGPGFQIKKAYWIDCGPKGQDYK	111
Query	74	SCTKNKACSETCVRAYMKRYGTFCTGGRTPTCQDYARIHNGGPRGCKSSATVGYWNVQK	133
		+C + +C+ C++ YM RY G C+ YARIHNGGPRGC + T+GYWNVK+++	
Sbjct	112	TCANDYSAYNCIQTYMARY--IGHSGCPKNCESYARIHNGGPRGCTNPNTIGYWNVKIKQ	169

# E-value, bit score

- ✓ E-value (the expectation value) – оценка количества случайных хитов такого же качества при таком размере базы данных (0 -  $e^{-6}$  – хорошо,  $> 0.001 - 0.01$  – плохо)

Как правило, BLAST недооценивает e-value!

- ✓ Bit Score – мера статистической значимости (вес – сумма стоимостей всех точечных замен) выравнивания, (меньше 50 – плохо)

# Как сохранить результаты BLAST?

- ✓ Распечатывать плохо – слишком много
- ✓ Сохранить как Web-страницу в браузере – сохраняются ссылки
- ✓ Можно сохранить в .pdf
- ✓ Графический дисплей можно сохранить как картинку, а остальное – как текст
- ✓ Сохранять выравнивания или условия бласта из раздела “download”

# Выбор параметров

Меняйте параметры только, если по умолчанию не работает (параметры по умолчанию подобраны хорошо для большинства ситуаций)

Для того, чтобы выбрать более подходящие параметры надо очень ТОЧНО сформулировать задачу

# Какие параметры менять?

## Фильтрация

Low-complexity region – другой aa-состав

- ✓ Фильтрация: если Ваш белок содержит большой регион низкой сложности – попробуйте использовать BLAST без соответствующей фильтрации
- ✓ Если Ваш белок содержит очень часто встречающиеся домены, их тоже можно отфильтровать – в ручную
- ✓ ДНК – геном-специфичные повторы!

# Параметры выравнивания

- ✓ Матрица: BLOSUM для локального выравнивания обычно лучше, чем PAM
  - Чем выше номер BLOSUM – тем строже выравнивание (BLOSUM80 вместо BLOSUM45 – более короткие выравнивания)
  - PAM – чем ниже, тем строже
- ✓ Штрафы за делеции:
  - Чем больше штраф за внесение, тем короче выравнивания
  - Меняете матрицу – надо менять и штраф
  - Чем ниже номер BLOSUM (выше PAM), тем меньше штраф за внесение делеции
  - Штраф за удлинение ~10 раз ниже, чем за внесение
- ✓ Если сравниваете удаленных гомологов, то лучше всего довольно высокий штраф за внесение делеции и низкий за удлинение
- ✓ Близкие гомологи – штрафы ближе друг к другу

# Параметры output-формата

- Количество хитов
- Выбор базы данных (организм)
- Выбор порога - Ехрест (если хитов мало, то можно посмотреть на более подозрительные)
- Entrez query – ключевые слова (например, “protease AND human”)

# PSI - BLAST

## Алгоритм:

- Несколько раундов поиска
- Первый раунд – просто blastp (BLOSUM62)
- Построение PSSM на основе полученных хитов (можете выбрать те, что надо)
- Следующий раунд на основе этой PSSM
- Методов итераций, пока множество хитов не перестанет меняться



# PHI - BLAST

Query – белок + паттерн, которому этот белок удовлетворяет

Пример:

>P28332|ADH6\_HUMAN Alcohol dehydrogenase 6 - Homo sapiens  
(Human)

```
MSTTGQVIRCKAAILWKPGAPFSIEEVEVAPPKAKEVRIKVVATGLCGTEMKVLGSKHLD  
LLYPTILGHEGAGIVESIGEGVSTVKPGDKVITLFLPQCGETSCLNSEGNFQFKQSK  
TQLMSDGTSRFTCKGKSIYHFGNTSTFCEYTVIKEISVAKIDAVAPLEKVCLISCGFSTG  
FGAAINTAKVTPGSTCAVFGLGGVGLSVVMGCKAAGAARIIGVDVNKEKFKKAQELGATE  
CLNPQDLKKPIQEVLFDMTDAGIDFCFEAIGNLDVLAALASCNESYGVCVVVGVLPASV  
QLKISGQLFFSGRSLKGSVFGGWKSRQHIPKLVADYMAEKLNLDP LITHTLNLDKINEAV  
ELMKTGKW
```

**G - H - E - x - {EL} - G - {AP} - x(4) - [GA] - x(2) -  
[IVSAC]**

# Пример простого мотива

Алкогольдегидрогеназа 6 (человек)	68 - 82:	GHEgAGIvesiGegV
Алкогольдегидрогеназа класса 3 (рис)	70 - 84:	GHEaAGIvesvGegV
Алкогольдегидрогеназа, специфичная к пропанолу (кишечная палочка)	57 - 71:	GHEgIGVvaevGpgV

Распознающее правило типа «паттерн»:

**G - H - E - x - {EL} - G - {AP} - x(4) - [GA] - x(2) - [IVSAC]**

Паттерн – регулярное выражение UNIX'а:

Например, выражение [AC]-x-V-x(4)-{ED} читается как

Ala или Cys- x-Val- x- x- x - x- (любой остаток, но не Glu и не Asp)

# Align2seq

Выравнивает 2 последовательности  
точно, как BLAST по базе данных  
(быстро, но не аккуратно)

# Другие программы построения выравниваний

## ✓ Поиск по БД:

- FASTA ([www.ebi.ac.uk/fasta33/](http://www.ebi.ac.uk/fasta33/))
- Ssearch (алгоритм Smith-Waterman) ([www.ch.embnet.org](http://www.ch.embnet.org))
- BLAT ([genome.ucsc.edu](http://genome.ucsc.edu))

## ✓ Парное выравнивание:

- Lalign ([www.ch.embnet.org](http://www.ch.embnet.org))
- Любая программа из следующей лекции