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*]

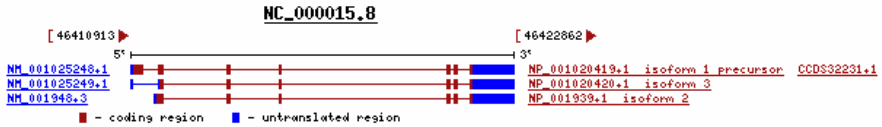
GeneID: 1854 updated 05-Nov-2007

Summary

Official Symbol	DUT provided by HGNC
Official Full Name	dUTP pyrophosphatase provided by HGNC
Primary source	HGNC:3078
See related	Ensembl:ENSG00000128951; HPRD:03165; MIM:601266
Gene type	protein coding
RefSeq status	Reviewed
Organism	*Homo sapiens*
Lineage	*Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo*
Also known as	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.

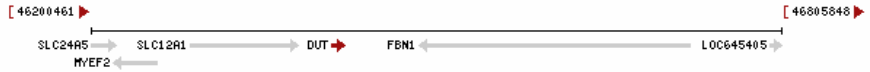
Genomic regions, transcripts, and products

Go to reference sequence details



Genomic context

chromosome: 15; Location: 15q15-q21.1 See DUT in MapViewer



Entrez Gene Home

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

$\Rightarrow$ молекулы, похожие по последовательности, похожи и по свойствам

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

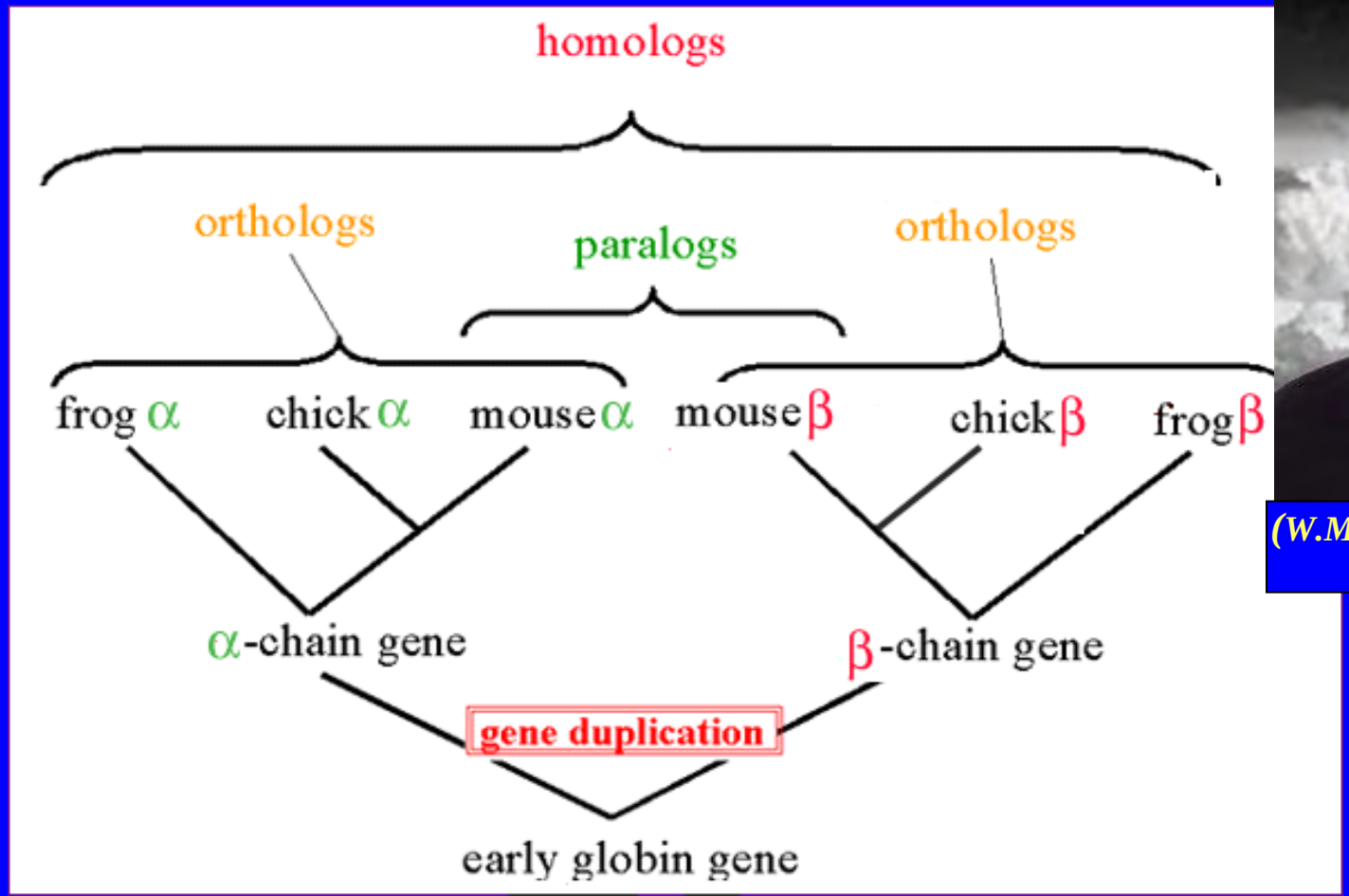
Гомологичные последовательности – последовательности, имеющие общее происхождение (общего предка)

Признаки гомологичности белков

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

Гомологи

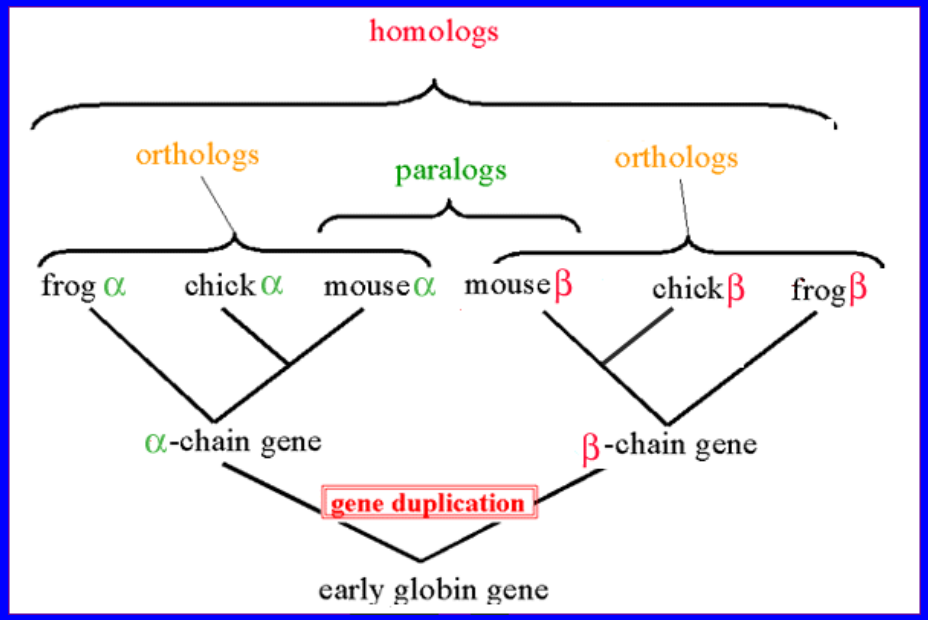
Ортологи Паралоги Ксенологи ?



(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

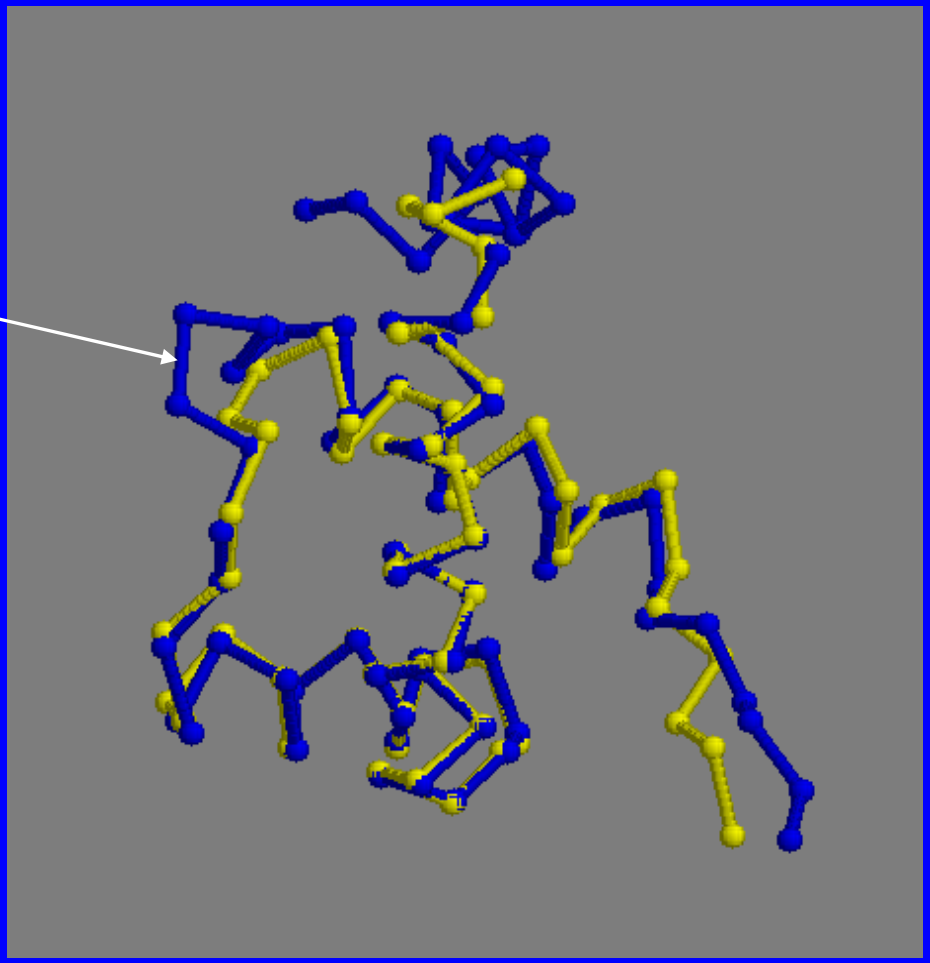
«Идеальное» выравнивание – запись последовательностей одна под другой так, чтобы гомологичные фрагменты оказались друг под другом.

```
     домовой
скупидом
    водомерка
```

```
лесовоз          ---лесо---воз
ледоход          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?

Гэп – пропуск в последовательности

```
лесовоз        ---лесо---воз
ледоход        лед---оход---
```

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^20)

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     домовой      домовой
скупидом       водомерка          водомерка
```

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

```
лесовоз          ---лесо---воз
ледоход     ?    лед---оход---
```

Критерии качества выравнивания

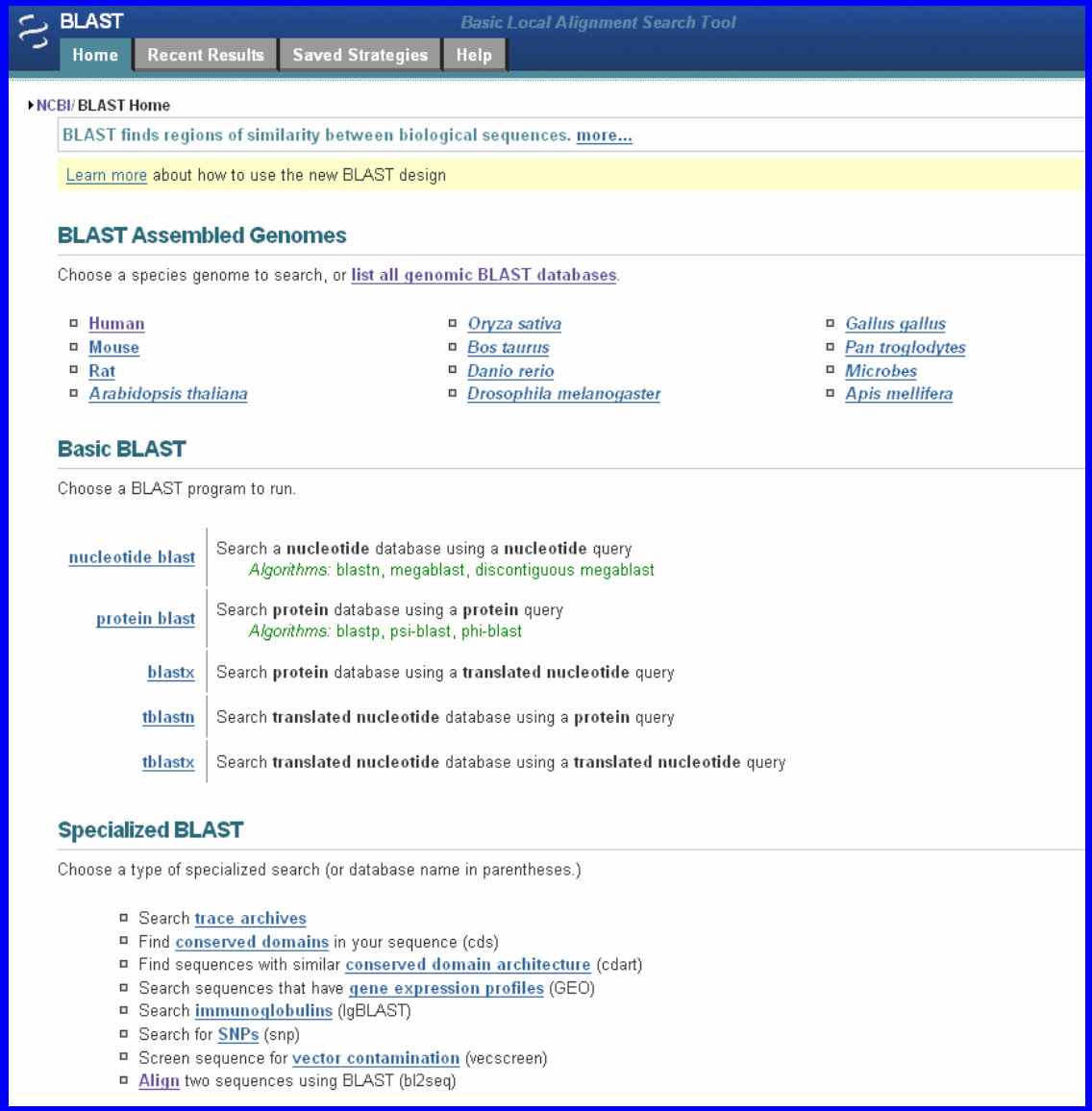
- ✓ Количество идентичных (похожих) аминокислот/нуклеотидов
  - – Для белков – более 25% id при длине > 100 aa
  - – Для ДНК – более 70% id при длине > 100 nt
- ✓ Длина выравнивания
- ✓ Вероятность наблюдать такое сходство случайным образом
  - – Зависит от базы данных
- ✓ Score – общая мера сходства: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



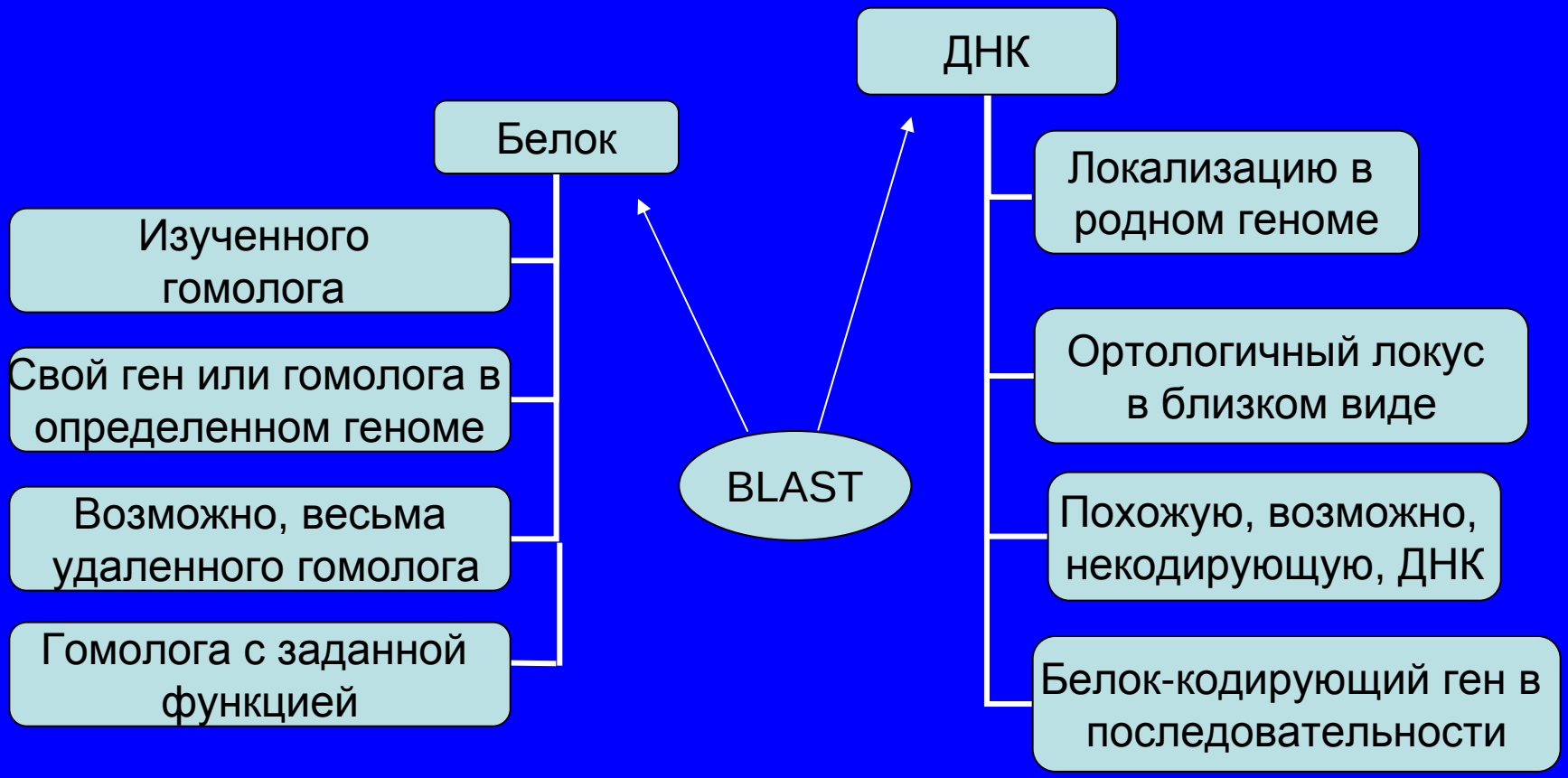
Какую программу BLAST выбрать?

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

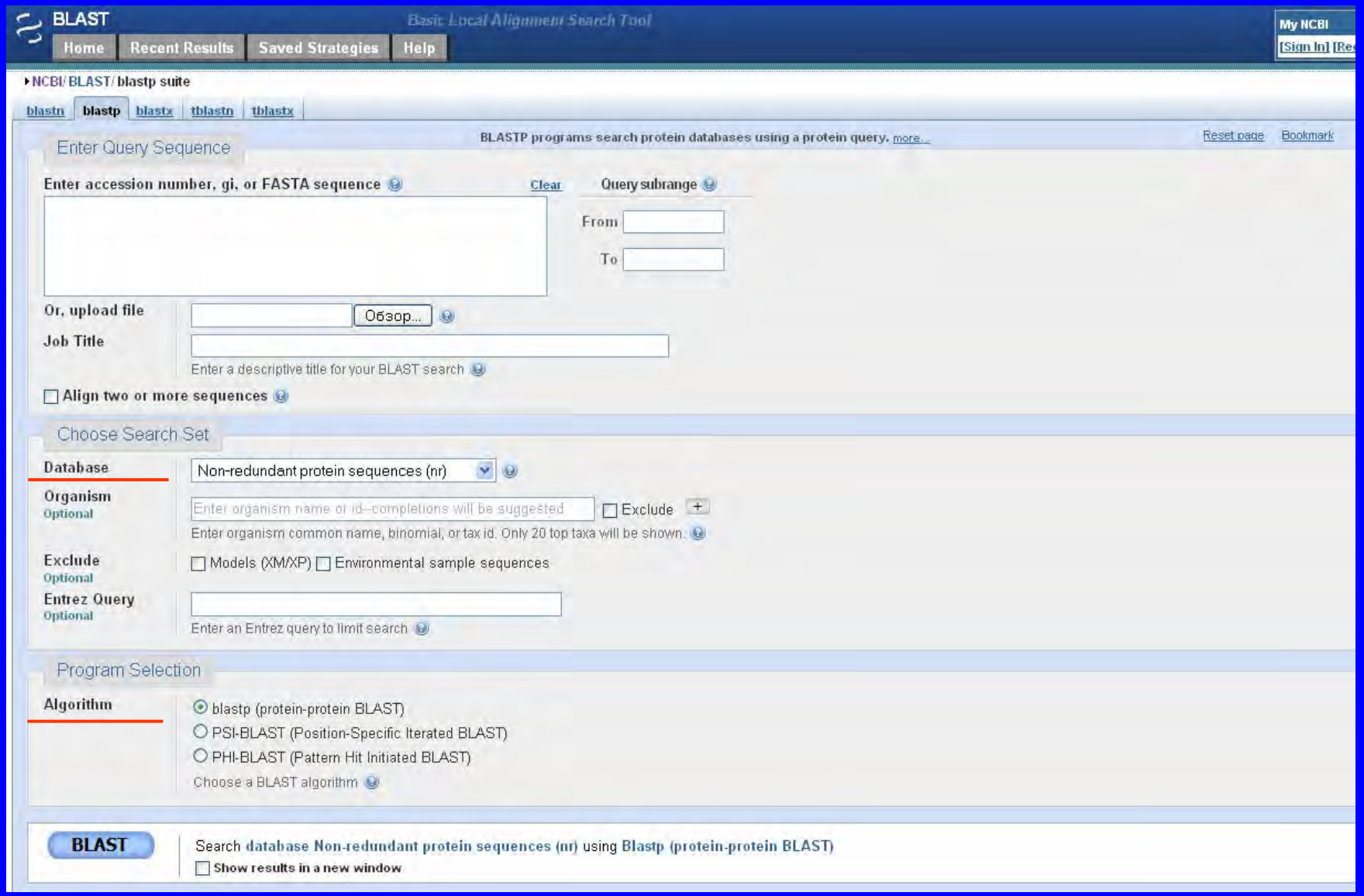
Дополнительные программы

- ✓ 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

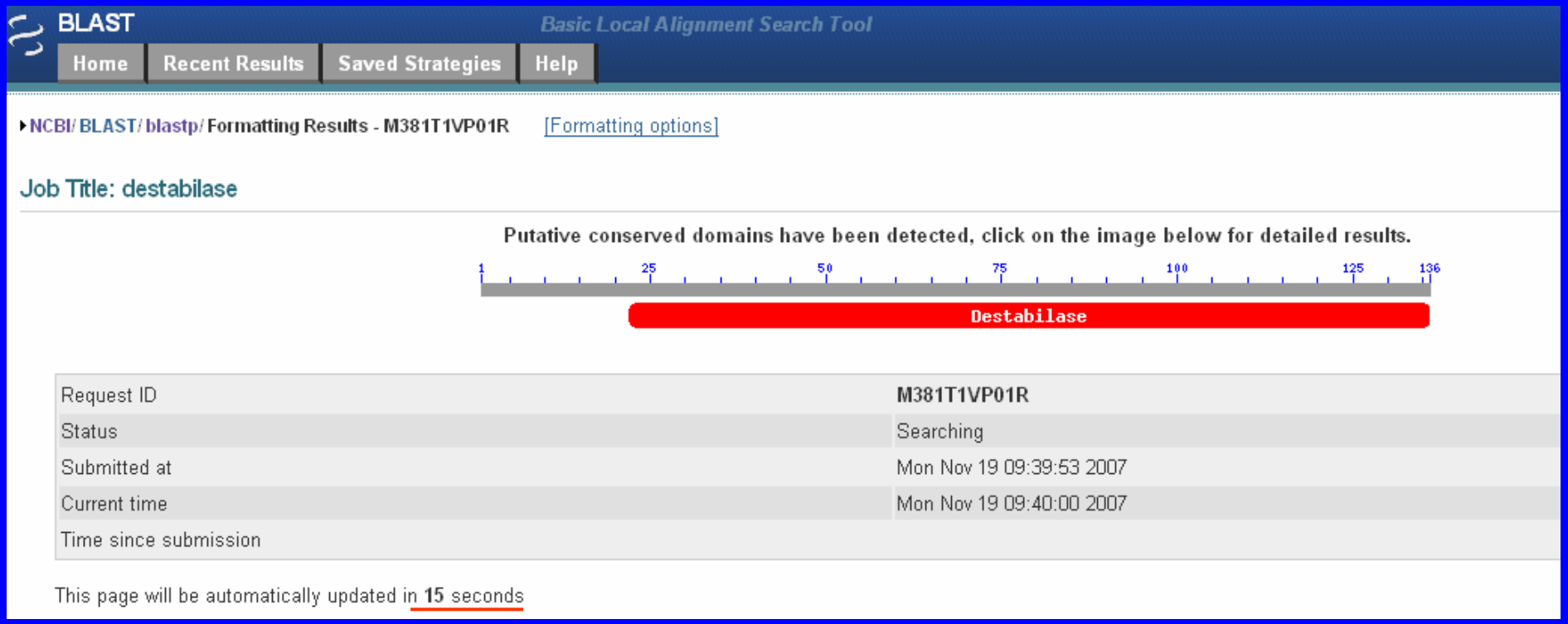
Какую программу выбрать?



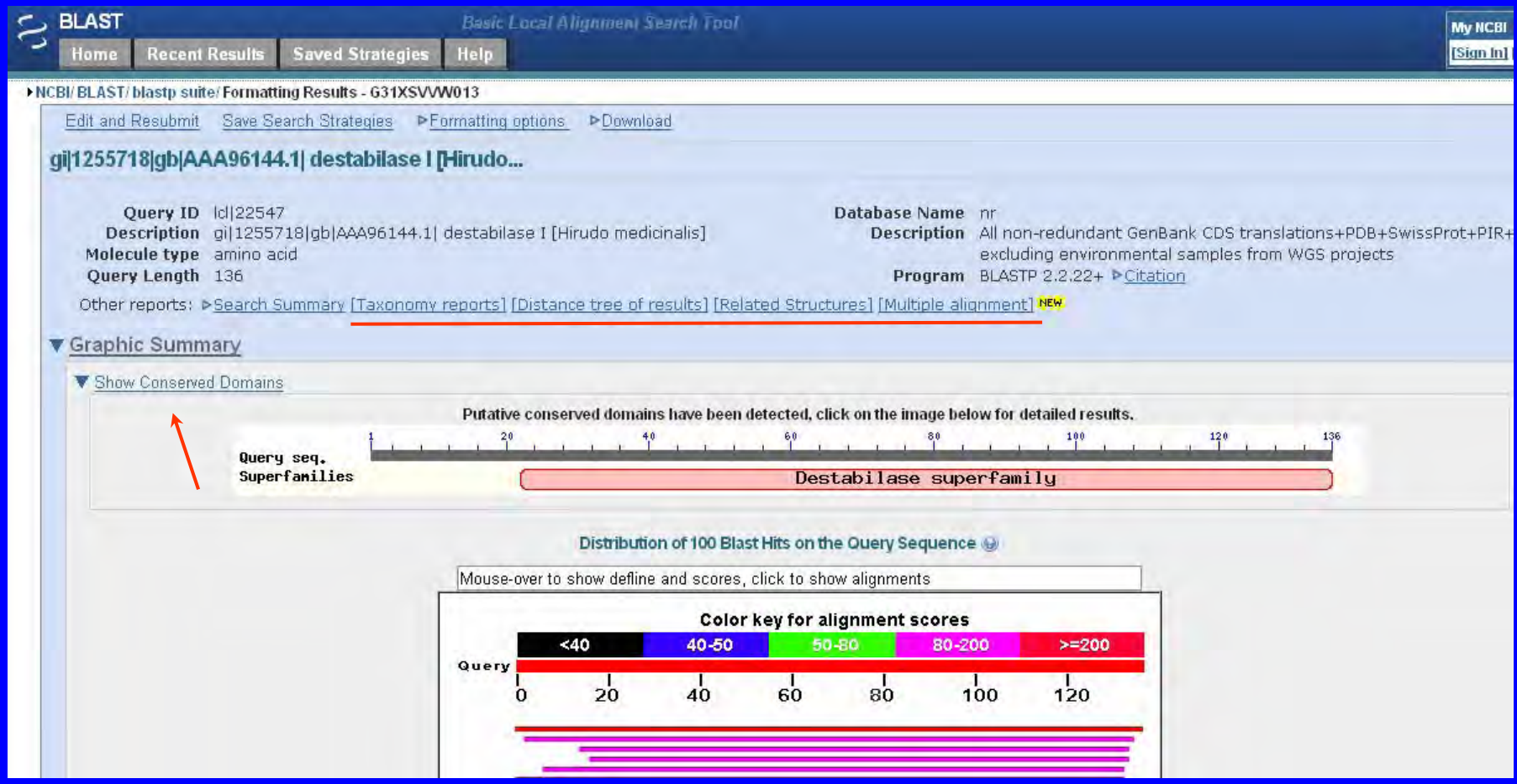
Стандартный input



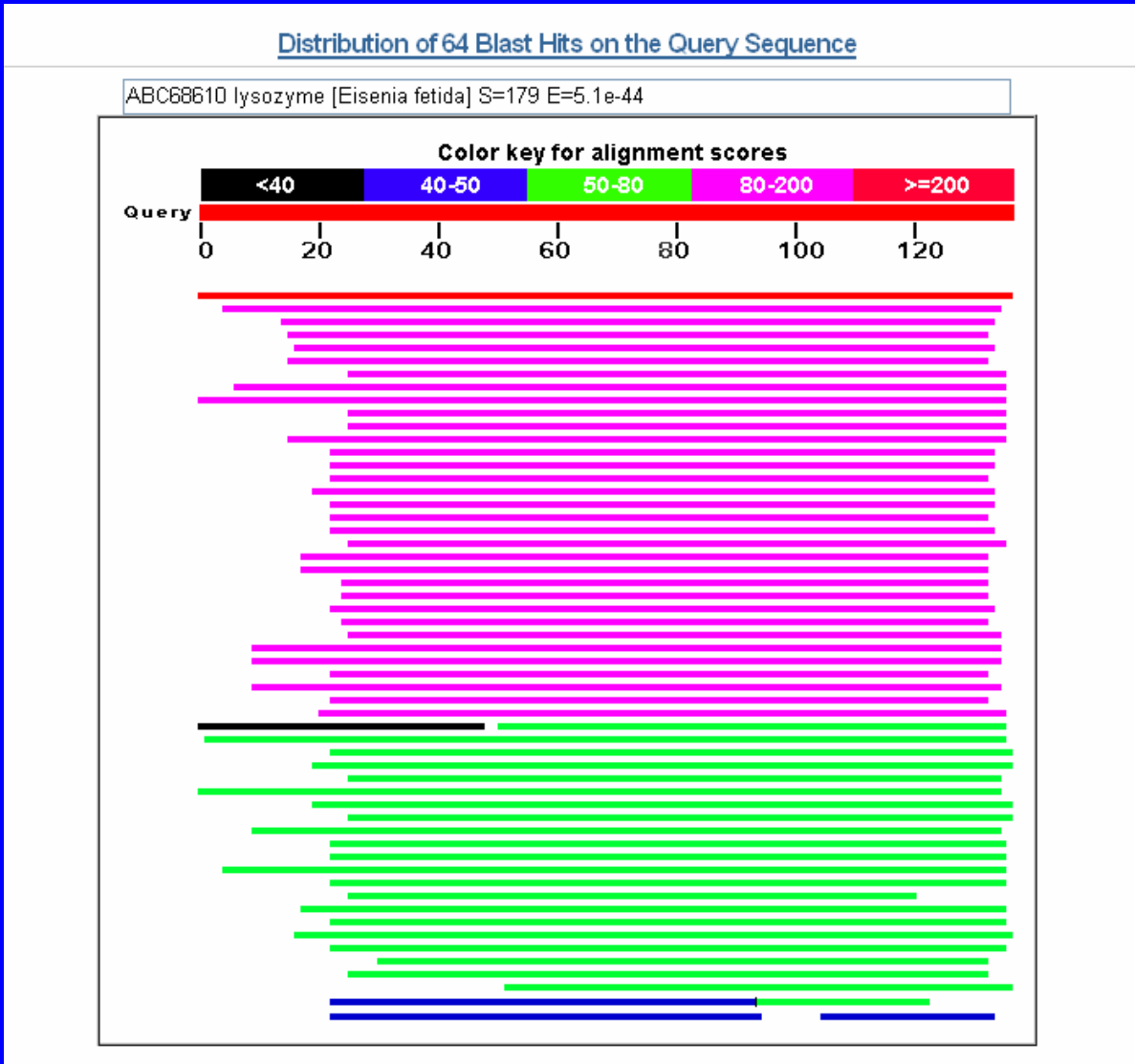
Промежуточная страница - CD



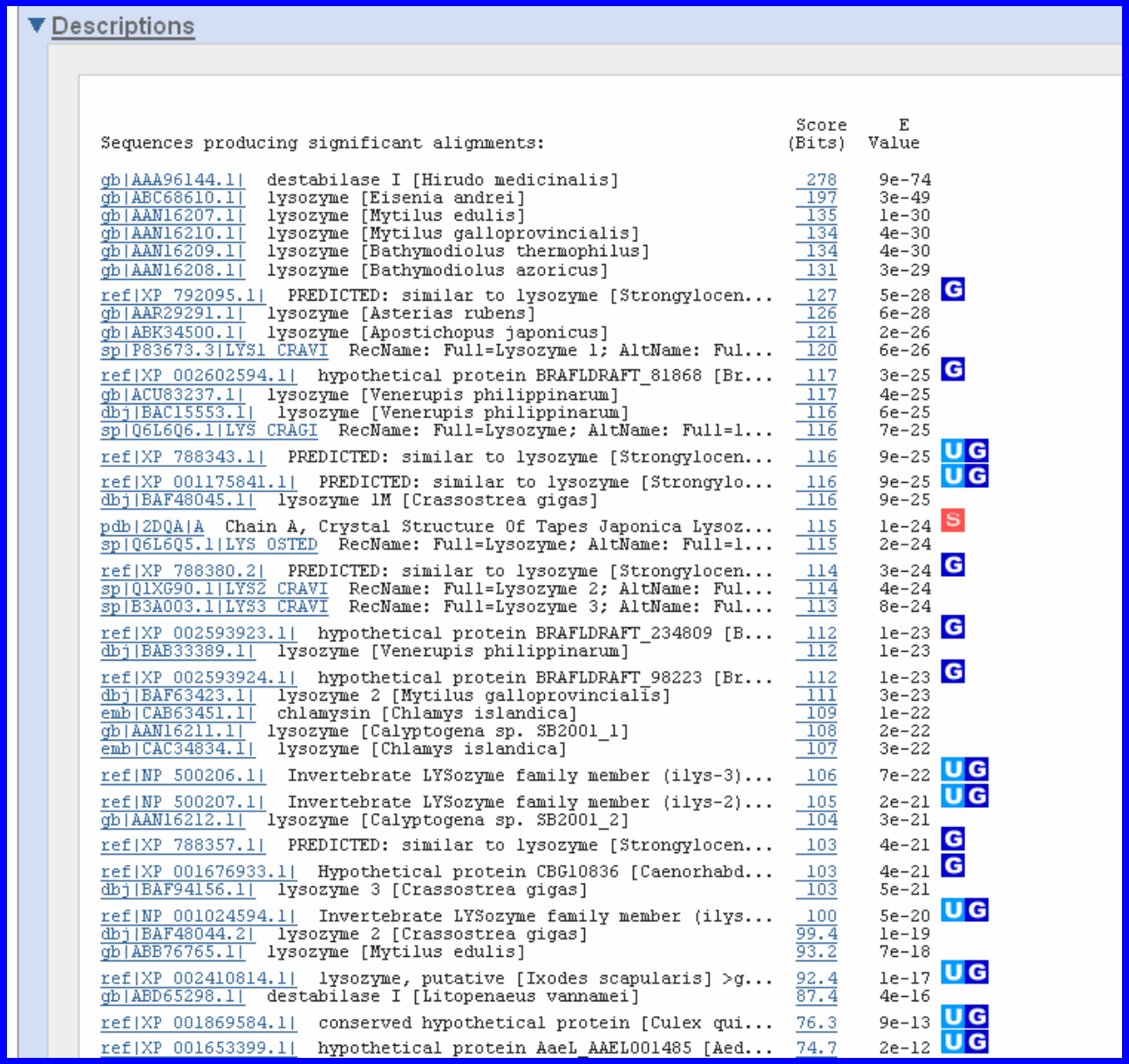
Output – I (parameters and options)



Output – II (color chart)



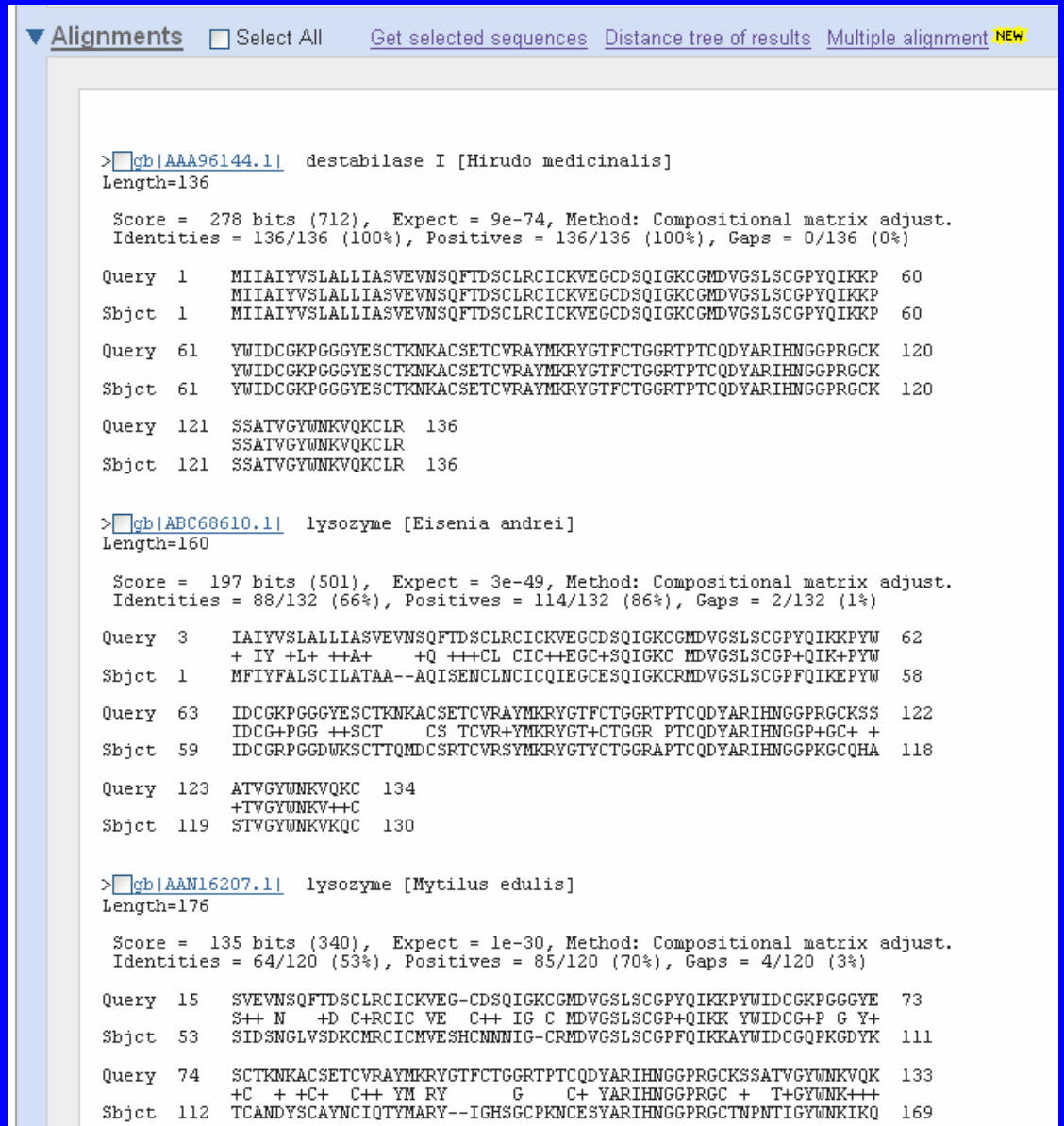
Output – III (descriptions)



Descriptions

Sequences producing significant alignments:	Score (Bits)	E Value	
gb|AAA96144.1| destabilase I [Hirudo medicinalis]	278	9e-74	
gb|ABC68610.1| lysozyme [Eisenia andrei]	197	3e-49	
gb|AAN16207.1| lysozyme [Mytilus edulis]	135	1e-30	
gb|AAN16210.1| lysozyme [Mytilus galloprovincialis]	134	4e-30	
gb|AAN16209.1| lysozyme [Bathymodiolus thermophilus]	134	4e-30	
gb|AAN16208.1| lysozyme [Bathymodiolus azoricus]	131	3e-29	
ref|XP_792095.1| PREDICTED: similar to lysozyme [Strongylocen...	127	5e-28	G
gb|AAR29291.1| lysozyme [Asterias rubens]	126	6e-28	
gb|ABK34500.1| lysozyme [Apostichopus japonicus]	121	2e-26	
sp|P83673.3|LYS1_CRAVI RecName: Full=Lysozyme 1; AltName: Ful...	120	6e-26	
ref|XP_002602594.1| hypothetical protein BRAFLDRAFT_81868 [Br...	117	3e-25	G
gb|ACU83237.1| lysozyme [Venerupis philippinarum]	117	4e-25	
dbj|BAC15553.1| lysozyme [Venerupis philippinarum]	116	6e-25	
sp|Q6L6Q6.1|LYS_CRAGI RecName: Full=Lysozyme; AltName: Full=l...	116	7e-25	
ref|XP_788343.1| PREDICTED: similar to lysozyme [Strongylocen...	116	9e-25	UG
ref|XP_001175841.1| PREDICTED: similar to lysozyme [Strongylo...	116	9e-25	UG
dbj|BAF48045.1| lysozyme 1M [Crassostrea gigas]	116	9e-25	
pdb|2DQA|A Chain A, Crystal Structure Of Tapes Japonica Lysoz...	115	1e-24	S
sp|Q6L6Q5.1|LYS_OSTED RecName: Full=Lysozyme; AltName: Full=l...	115	2e-24	
ref|XP_788380.2| PREDICTED: similar to lysozyme [Strongylocen...	114	3e-24	G
sp|Q1XG90.1|LYS2_CRAVI RecName: Full=Lysozyme 2; AltName: Ful...	114	4e-24	
sp|B3A003.1|LYS3_CRAVI RecName: Full=Lysozyme 3; AltName: Ful...	113	8e-24	
ref|XP_002593923.1| hypothetical protein BRAFLDRAFT_234809 [B...	112	1e-23	G
dbj|BAB33389.1| lysozyme [Venerupis philippinarum]	112	1e-23	
ref|XP_002593924.1| hypothetical protein BRAFLDRAFT_98223 [Br...	112	1e-23	G
dbj|BAF63423.1| lysozyme 2 [Mytilus galloprovincialis]	111	3e-23	
emb|CAB63451.1| chlamysin [Chlamys islandica]	109	1e-22	
gb|AAN16211.1| lysozyme [Calyptogena sp. SB2001_1]	108	2e-22	
emb|CAC34834.1| lysozyme [Chlamys islandica]	107	3e-22	
ref|NP_500206.1| Invertebrate LYSozyme family member (ilys-3)...	106	7e-22	UG
ref|NP_500207.1| Invertebrate LYSozyme family member (ilys-2)...	105	2e-21	UG
gb|AAN16212.1| lysozyme [Calyptogena sp. SB2001_2]	104	3e-21	
ref|XP_788357.1| PREDICTED: similar to lysozyme [Strongylocen...	103	4e-21	G
ref|XP_001676933.1| Hypothetical protein CBG10836 [Caenorhabd...	103	4e-21	G
dbj|BAF94156.1| lysozyme 3 [Crassostrea gigas]	103	5e-21	
ref|NP_001024594.1| Invertebrate LYSozyme family member (ilys...	100	5e-20	UG
dbj|BAF48044.2| lysozyme 2 [Crassostrea gigas]	99.4	1e-19	
gb|ABB76765.1| lysozyme [Mytilus edulis]	93.2	7e-18	
ref|XP_002410814.1| lysozyme, putative [Ixodes scapularis] >g...	92.4	1e-17	UG
gb|ABD65298.1| destabilase I [Litopenaeus vannamei]	87.4	4e-16	
ref|XP_001869584.1| conserved hypothetical protein [Culex qui...	76.3	9e-13	UG
ref|XP_001653399.1| hypothetical protein AaeL_AAEL001485 [Aed...	74.7	2e-12	UG

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    MIIAIYVSLALLIASVEVNSQFTDSCLRCICKVEGCDSQIGKCGMDVGSLSCGPYQIKKP  60
            MIIAIYVSLALLIASVEVNSQFTDSCLRCICKVEGCDSQIGKCGMDVGSLSCGPYQIKKP
Sbjct  1    MIIAIYVSLALLIASVEVNSQFTDSCLRCICKVEGCDSQIGKCGMDVGSLSCGPYQIKKP  60

Query  61   YWIDCGKPGGGYESCTKNKACSETCVRAYMKRYGTFCTGGRTPTCQDYARIHNGGPRGCK  120
            YWIDCGKPGGGYESCTKNKACSETCVRAYMKRYGTFCTGGRTPTCQDYARIHNGGPRGCK
Sbjct  61   YWIDCGKPGGGYESCTKNKACSETCVRAYMKRYGTFCTGGRTPTCQDYARIHNGGPRGCK  120

Query  121  SSATVGYWNKVQKCLR  136
            SSATVGYWNKVQKCLR
Sbjct  121  SSATVGYWNKVQKCLR  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    IAIYVSLALLIASVEVNSQFTDSCLRCICKVEGCDSQIGKCGMDVGSLSCGPYQIKKPYW  62
            + IY +L+ ++A+    +Q +++CL CIC++EGC+SQIGKC MDVGSLSCGP+QIK+PYW
Sbjct  1    MFIYFALSCILATAA--AQISENCLNCICQIEGCESQIGKCRMDVGSLSCGPFQIKEPYW  58

Query  63   IDCGKPGGGYESCTKNKACSETCVRAYMKRYGTFCTGGRTPTCQDYARIHNGGPRGCKSS  122
            IDCG+PGG ++SCT    CS TCVR+YMKRYGT+CTGGR PTCQDYARIHNGGP+GC+ +
Sbjct  59   IDCGRPGGDWKSCTTQMDCSRTCVRSYMKRYGTYCTGGRAPTCQDYARIHNGGPKGCQHA  118

Query  123  ATVGYWNKVQKC  134
            +TVGYWNKV++C
Sbjct  119  STVGYWNKVKQC  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   SVEVNSQFTDSCLRCICKVEG-CDSQIGKCGMDVGSLSCGPYQIKKPYWIDCGKPGGGYE  73
            S++ N   +D C+RCIC VE  C++ IG C MDVGSLSCGP+QIKK YWIDCG+P G Y+
Sbjct  53   SIDSNGLVSDKCMRCICMVESHCNNNIG-CRMDVGSLSCGPFQIKKAYWIDCGQPKGDYK  111

Query  74   SCTKNKACSETCVRAYMKRYGTFCTGGRTPTCQDYARIHNGGPRGCKSSATVGYWNKVQK  133
            +C  + +C+  C++ YM RY      G    C+ YARIHNGGPRGC +  T+GYWNK+++
Sbjct  112  TCANDYSCAYNCIQTYMARY--IGHSGCPKNCESYARIHNGGPRGCTNPNTIGYWNKIKQ  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 линки
- ✓ 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 аа-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 - Expect (если хитов мало, то можно смотреть 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
LLYPTILGHEGAGIVESIGEGVSTVKPGDKVITLFLPQCGECTSCLNSEGNFCIQFKQSK
TQLMSDGTSRFTCKGKSIYHFGNTSTFCEYTVIKEISVAKIDAVAPLEKVCLISCGFSTG
FGAAINTAKVTPGSTCAVFGLGGVGLSVVMGCKAAGAARIIGVDVNKEKFKKAQELGATE
CLNPQDLKKPIQEVLFDMTDAGIDFCFEAIGNLDVLAAALASCNESYGVCVVVGVLPASV
QLKISGQLFFSGRSLKGSVFGGWKSRQHIPKLVADYMAEKLNLDPLITHTLNLDKINEAV
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'a:
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/)
  - Ssearch (алгоритм Smith-Waterman) (www.ch.embnet.org)
  - BLAT (genome.ucsc.edu)
- ✓ Попарное выравнивание:
  - Lalign (www.ch.embnet.org)
  - Любая программа из следующей лекции