Bioinformatika és genomanalízis az orvostudományban

Keresés adatbázisokban

Cserző Miklós

2020

https://semmelweis.zoom.us/j/96102872458?pwd=Rk1PL2tqS21sdlUwc3B4eDFCZkNKQT09

A mai előadás

- Szekvenciakeresés:
  - FASTA
  - BLAST család
  - HMMER
- Párhuzamos változatok a keresésre
- A keresési eredmények értékelése
- Szöveges keresés: eTBLAST
- Adatbányászat

Az adatbázis keresés jelentősége

- Össze tudunk hasonlítani két szekvenciát
- El tudjuk dönteni, hogy közös őstől származnak-e
- Tudunk illeszteni több, hasonló szekvenciát
- De hogyan találjuk meg a hasonlókat több millió szekvencia közt???

Mit keresünk?

- Keresés az annotációban
  - Kódok
  - Kulcsszavak
  - Azonosított funkciók
- Ez alapkövetelmény minden adatbázissal szemben
- Keresés a szekvenciában
- Ez a biológiai adatbázisok jellegzetessége

Hogy lehet gyorsan keresni?

- A szekvenciát feltördeljük rövid, összefüggő karakter-sorozatokra – *szavakra* (*wordsize, k-mer, k-tuple*)
- Az adatbázist indexeljük: elkészítjük a *k-merek* listáját és hozzárendeljük a helyüket
- Ezt csak egyszer kell megcsinálni
- A kereső szekvenciát is felbontjuk *k-merekre*
- És így keresünk az adatbázisban

GTCTGACAGCAGCCGCTGCAGCAGCTGCTGCTGCTACCAATGCAG
CTATTGCTGAAGCAA

GTC:1

GTCTGACAGCAGCCGCTGCAGCAGCTGCTGCTGCTACCAATGCAG
CTATTGCTGAAGCAA

GTC:1 TCT:2

GTCTGACAGCAGCCGCTGCAGCAGCTGCTGCTGCTACCAATGCAG
CTATTGCTGAAGCAA

GTC:1 TCT:2 CTG:3

GTCTGACAGCAGCCGCTGCAGCAGCTGCTGCTGCTACCAATGCAG
CTATTGCTGAAGCAA

Táblázatos forma (lookup table):

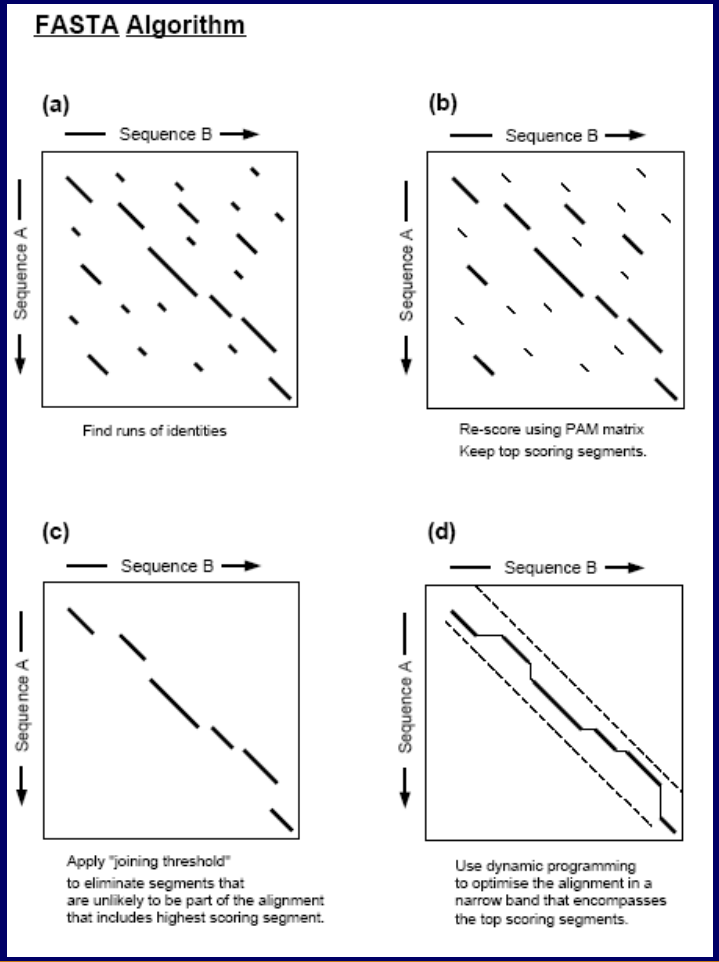
AAG:55, AAT:39, ACA:6, ACC:36, AGC:8:11:20:23:44:56, ATG:40,
ATT:48, CAA:38:58, CAG:7:10:19:22:43, CCA:37, CCG:13, CGC:14,
CTA:34:46, CTG:3:16:25:28:31:52, GAA:54, GAC:5, GCA:9:18:21:42:57,
GCC:12, GCT:15:24:27:30:33:45:51, GTC:1, TAC:35, TAT:47, TCT:2,
TGA:4:53, TGC:17:26:29:32:41:50, TTG:49

A FASTA család

- Honlap: http://fasta.bioch.virginia.edu/
- Szolgáltatás itt és az EBI oldalán is
- Letölthető Linuxra és Windowsra is
- Dokumentáció és tutorial is elérhető
- Az egyik legnépszerűbb kereső
- Az algoritmus régi, de azóta is folyamatosan fejlesztik a programot

Az algoritmus négy lépése

a) Gyors keresés táblázat alapján (lookup table)
b) Pontozás mátrix alapján
c) Kiválasztás
d) S-W illesztés, végeredmény



A programcsalád tagjai

- A legfrissebb verzió a 3.6-os
  - „fasta”: egy szekvenciát összehasonlít egy databázis szekvenciái ellenében (fehérjét fehérjével vagy DNS-t DNS-sel), gyors algoritmust használ – táblázatos keresés
  - „ssearch”: S-W algoritmust használ a keresésre (fehérjét fehérjével vagy DNS-t DNS-sel), lassabb, de pontosabb

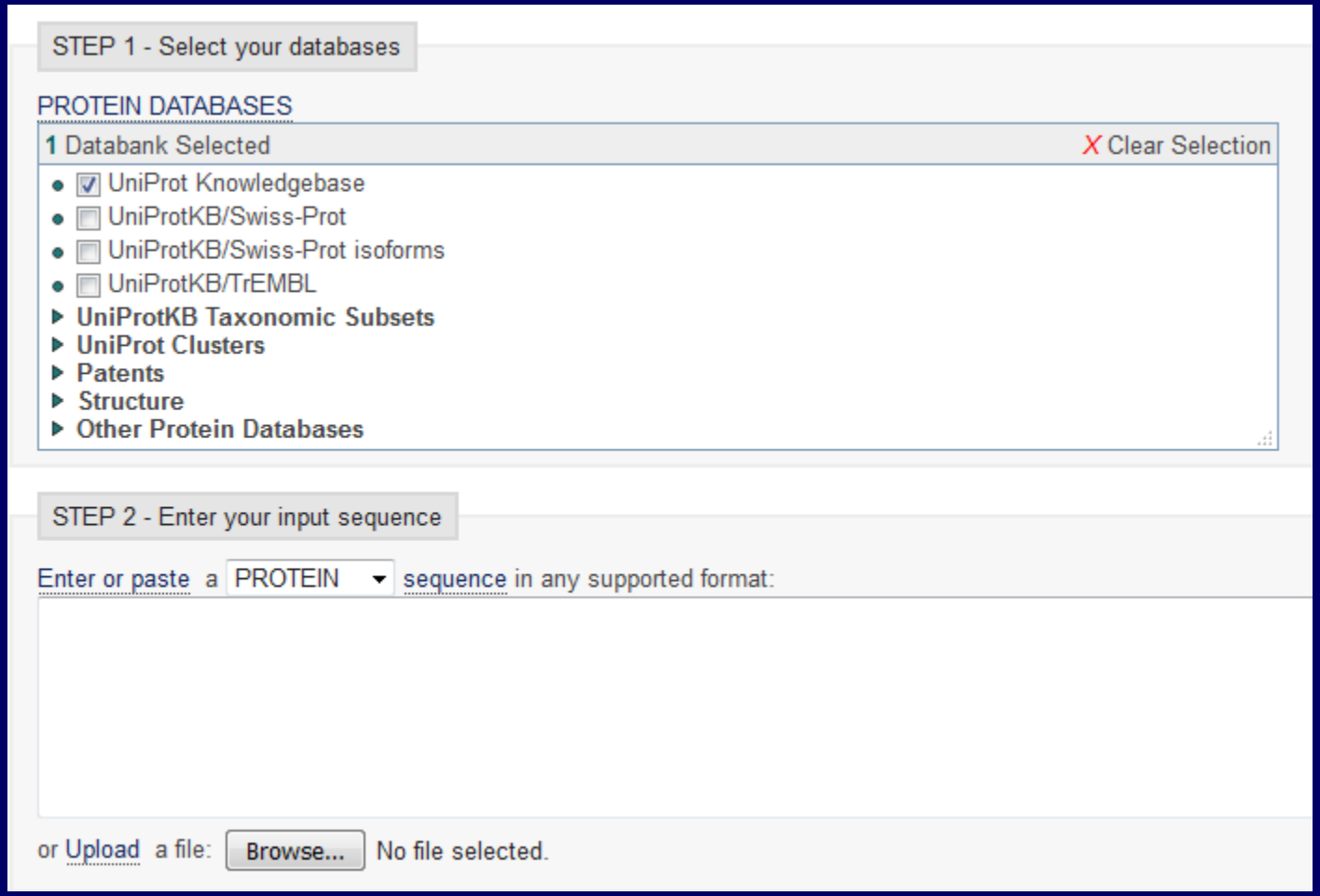
Folytatás ...

- „ggsearch”: global:global keresés (fehérje és DNS)
- „glsearch”: global:local keresés (fehérje és DNS)
- „fastx”: lefordított DNS szekvenciát keres fehérje adatbázis ellen, három frame-et fordít, toldás és frame-eltolás is megengedett
- „fasty”: mint a „fastx”, de kódonon belüli eltolás is engedett

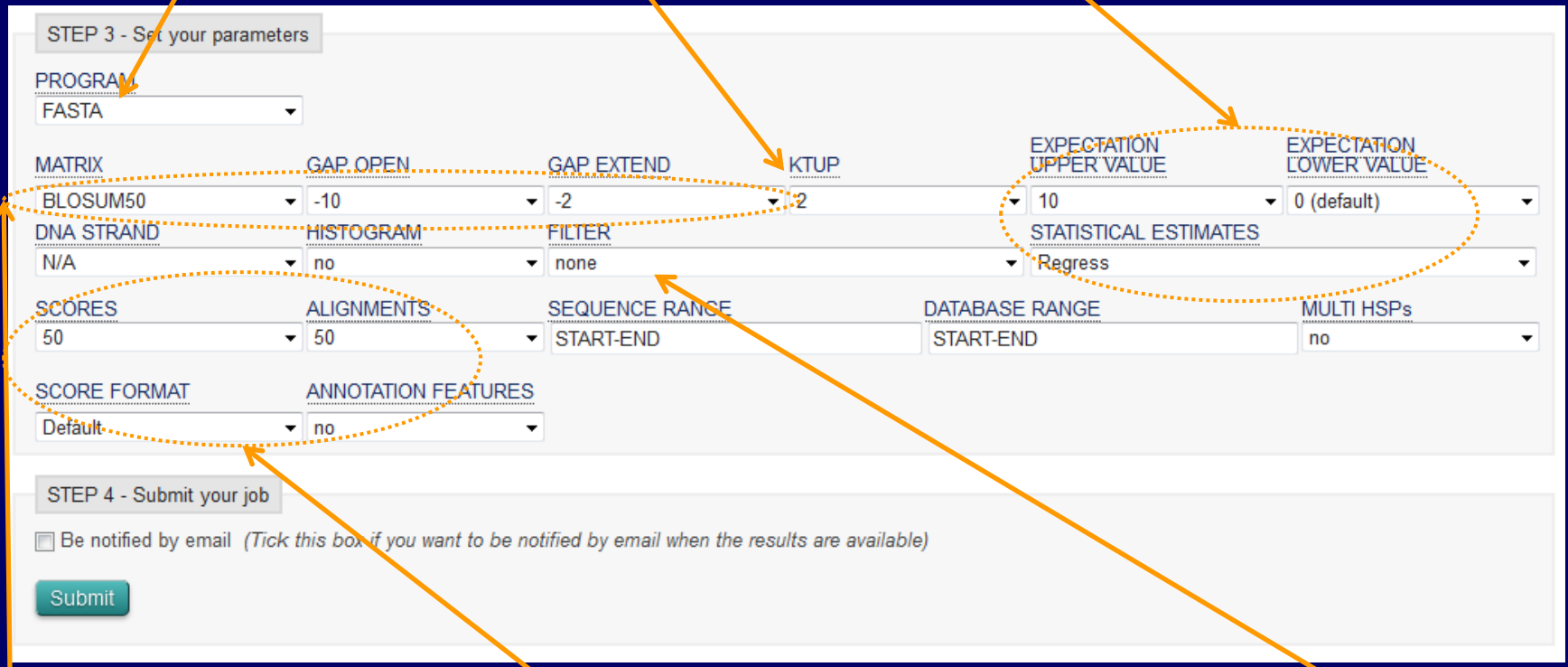
Folytatás ...

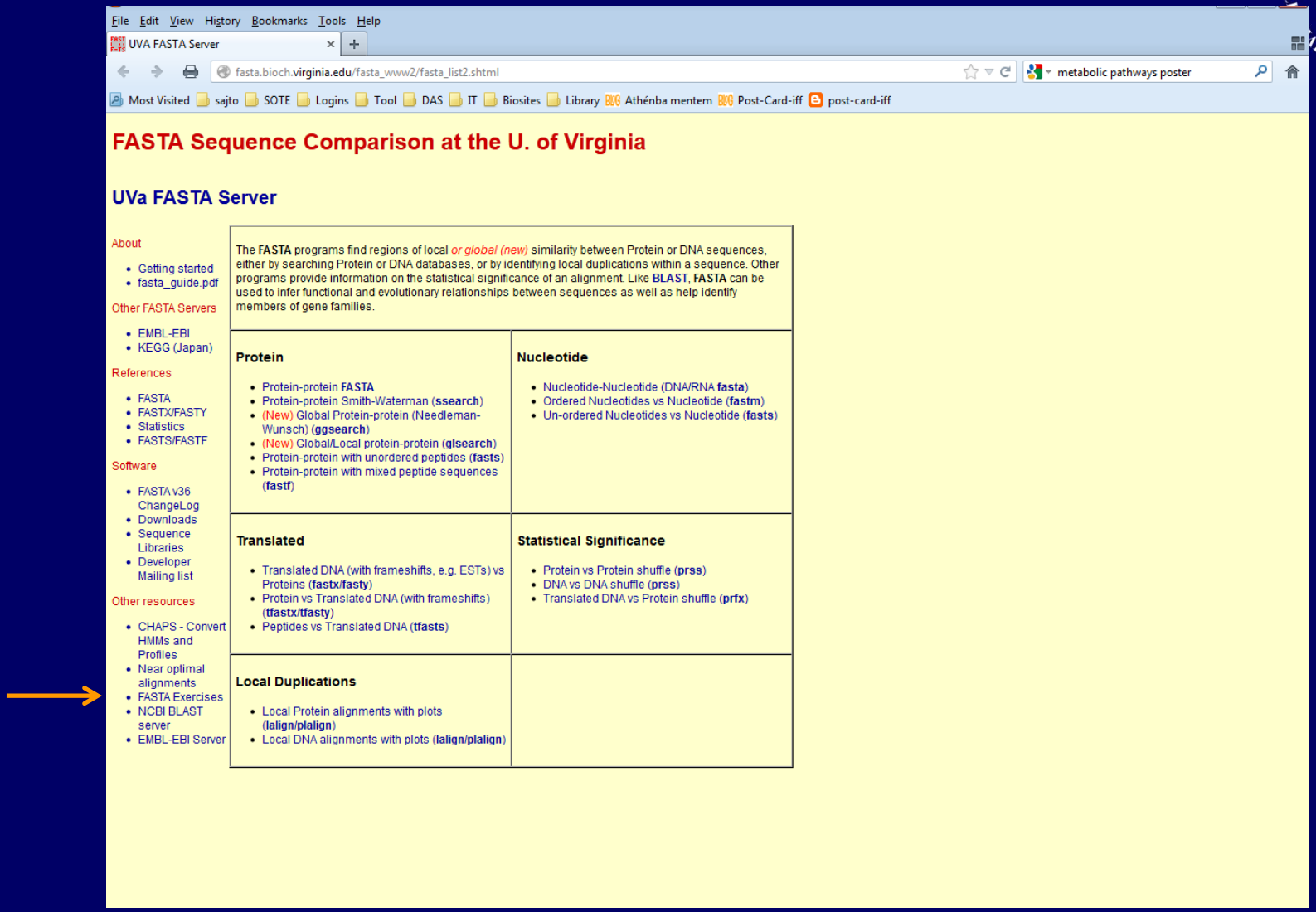
- „tfastx”, „tfasty”: fehérje szekvenciát keres DNS adatbázis ellen
- „fastf”, „tfastf”: peptid keverék listáját keresi fehérje, illetve DNS adatbázis ellen, mintha részlegesen emésztett minta lenne
- „fasts”, „tfasts”: peptid fragmens listát keres fehérje, illetve DNS adatbázis ellen, mintha tömegspektrométerből származó minta lenne
- „lalign”: többszörös illesztő program

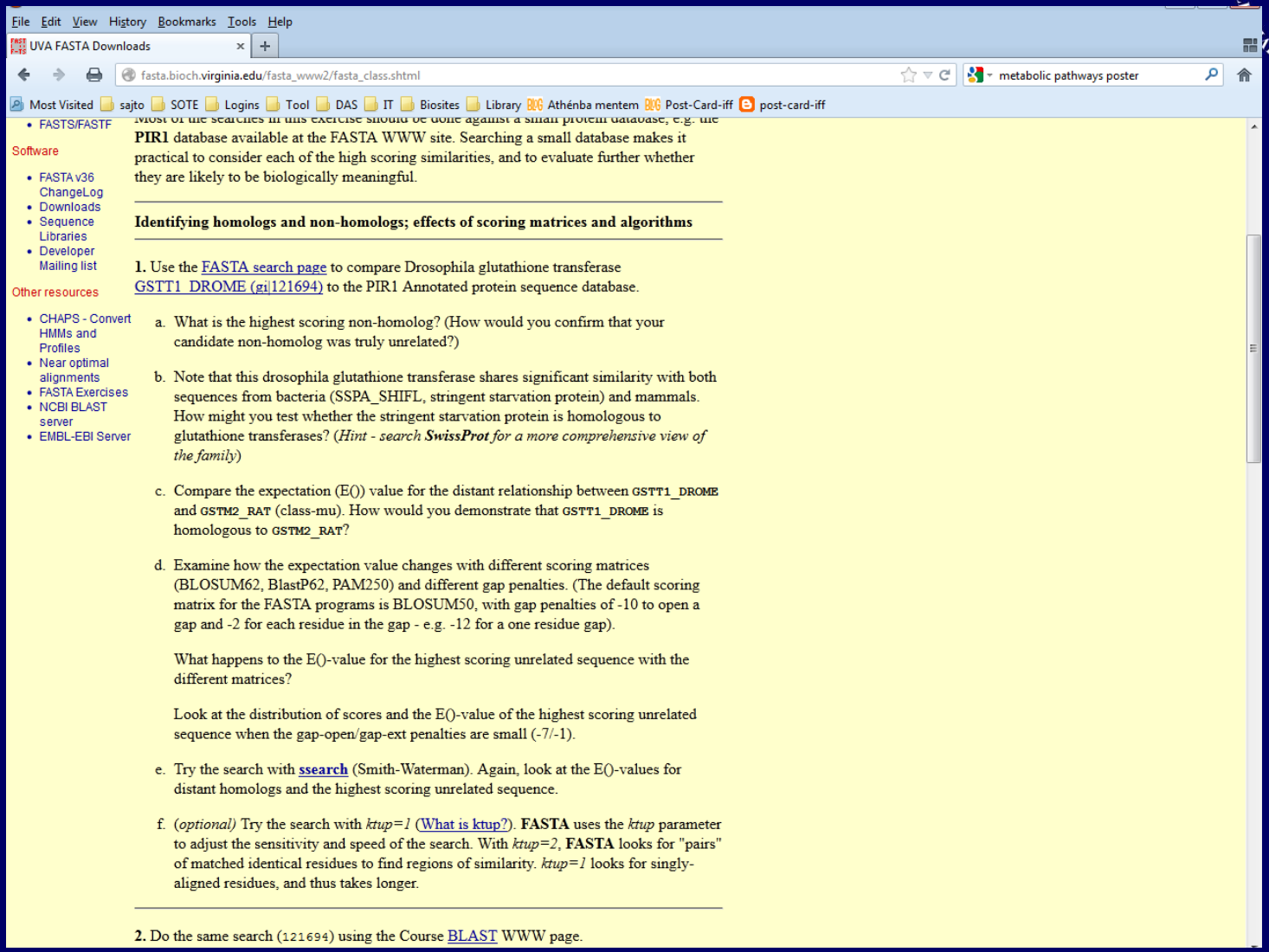
A FASTA felület – EBI



Paraméterek




File Edit View History Bookmarks Tools Help
UVA FASTA Server
fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml
metabolic pathways poster
Most Visited sajto SOTE Logins Tool DAS IT Biosites Library Athénba mentem Post-Card-iff post-card-iff
FASTA Sequence Comparison at the U. of Virginia
UVa FASTA Server
About
Getting started
fasta_guide.pdf
Other FASTA Servers
EMBL-EBI
KEGG (Japan)
References
FASTA
FASTX/FASTY
Statistics
FASTS/FASTF
Software
FASTA v36 ChangeLog
Downloads
Sequence Libraries
Developer Mailing list
Other resources
CHAPS - Convert HMMs and Profiles
Near optimal alignments
FASTA Exercises
NCBI BLAST server
EMBL-EBI Server
The FASTA programs find regions of local or global (new) similarity between Protein or DNA sequences, either by searching Protein or DNA databases, or by identifying local duplications within a sequence. Other programs provide information on the statistical significance of an alignment. Like BLAST, FASTA can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.
Protein
Protein-protein FASTA
Protein-protein Smith-Waterman (ssearch)
(New) Global Protein-protein (Needleman-Wunsch) (ggsearch)
(New) Global/Local protein-protein (glsearch)
Protein-protein with unordered peptides (fasts)
Protein-protein with mixed peptide sequences (fastf)
Nucleotide
Nucleotide-Nucleotide (DNA/RNA fasta)
Ordered Nucleotides vs Nucleotide (fastm)
Un-ordered Nucleotides vs Nucleotide (fasts)
Translated
Translated DNA (with frameshifts, e.g. ESTs) vs Proteins (fastx/fasty)
Protein vs Translated DNA (with frameshifts) (tfastx/tfasty)
Peptides vs Translated DNA (tfasts)
Statistical Significance
Protein vs Protein shuffle (prss)
DNA vs DNA shuffle (prss)
Translated DNA vs Protein shuffle (prfx)
Local Duplications
Local Protein alignments with plots (lalign/plalign)
Local DNA alignments with plots (lalign/plalign)



File Edit View History Bookmarks Tools Help

UVA FASTA Downloads

fasta.bioch.virginia.edu/fasta_www2/fasta_class.shtml

metabolic pathways poster

Most Visited sajto SOTE Logins Tool DAS IT Biosites Library Athénba mentem Post-Card-iff post-card-iff

- FASTS/FASTF

Software

- FASTA v36 ChangeLog
- Downloads
- Sequence Libraries
- Developer Mailing list

Other resources

- CHAPS - Convert HMMs and Profiles
- Near optimal alignments
- FASTA Exercises
- NCBI BLAST server
- EMBL-EBI Server

Most of the searches in this exercise should be done against a small protein database, e.g. the **PIR1** database available at the FASTA WWW site. Searching a small database makes it practical to consider each of the high scoring similarities, and to evaluate further whether they are likely to be biologically meaningful.

Identifying homologs and non-homologs; effects of scoring matrices and algorithms

---

1. Use the FASTA search page to compare Drosophila glutathione transferase GSTT1_DROME (gi|121694) to the PIR1 Annotated protein sequence database.

a. What is the highest scoring non-homolog? (How would you confirm that your candidate non-homolog was truly unrelated?)

b. Note that this drosophila glutathione transferase shares significant similarity with both sequences from bacteria (SSPA_SHIFL, stringent starvation protein) and mammals. How might you test whether the stringent starvation protein is homologous to glutathione transferases? (*Hint - search **SwissProt** for a more comprehensive view of the family*)

c. Compare the expectation (E()) value for the distant relationship between `GSTT1_DROME` and `GSTM2_RAT` (class-mu). How would you demonstrate that `GSTT1_DROME` is homologous to `GSTM2_RAT`?

d. Examine how the expectation value changes with different scoring matrices (BLOSUM62, BlastP62, PAM250) and different gap penalties. (The default scoring matrix for the FASTA programs is BLOSUM50, with gap penalties of -10 to open a gap and -2 for each residue in the gap - e.g. -12 for a one residue gap).

What happens to the E()-value for the highest scoring unrelated sequence with the different matrices?

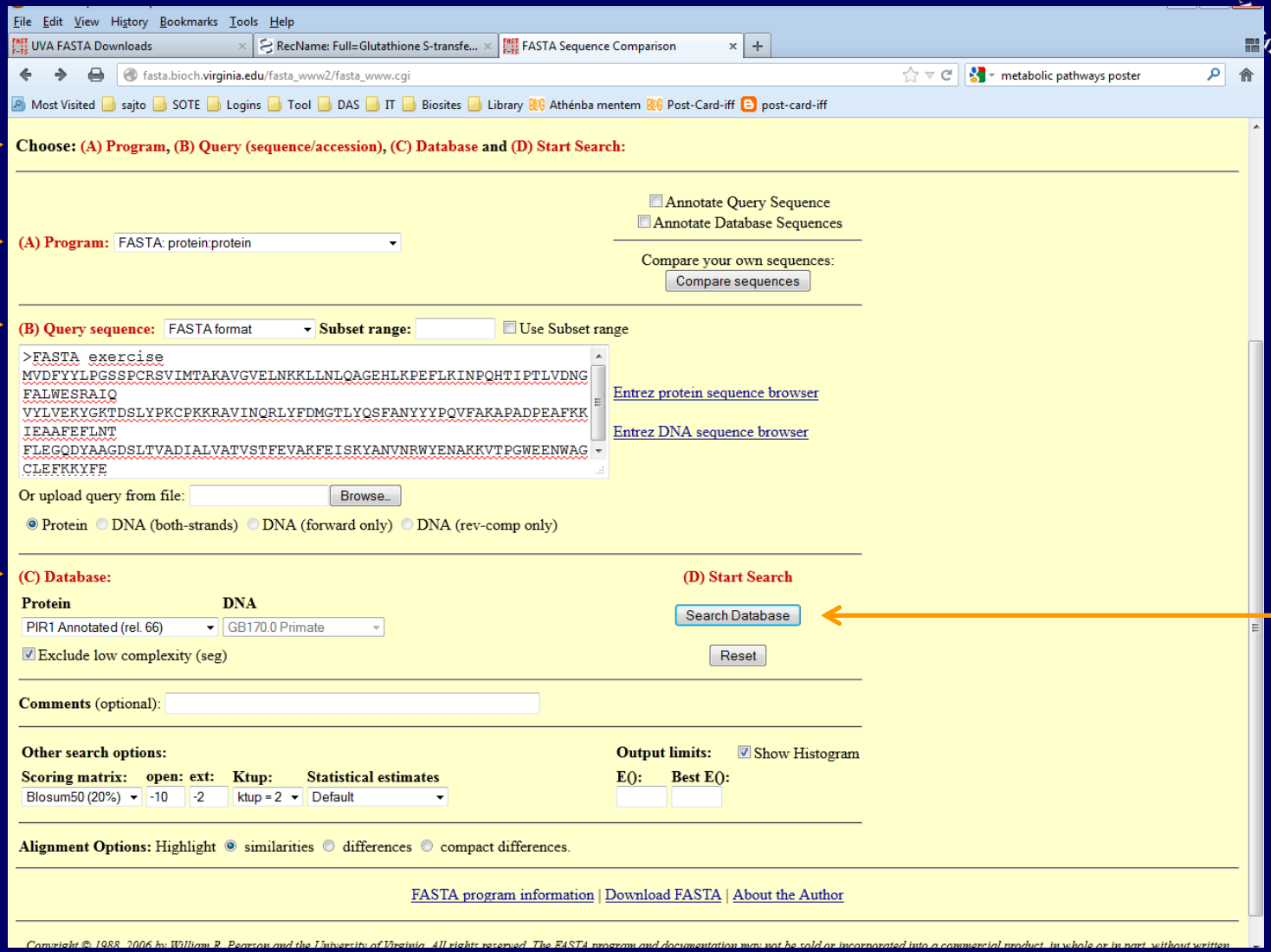
Look at the distribution of scores and the E()-value of the highest scoring unrelated sequence when the gap-open/gap-ext penalties are small (-7/-1).

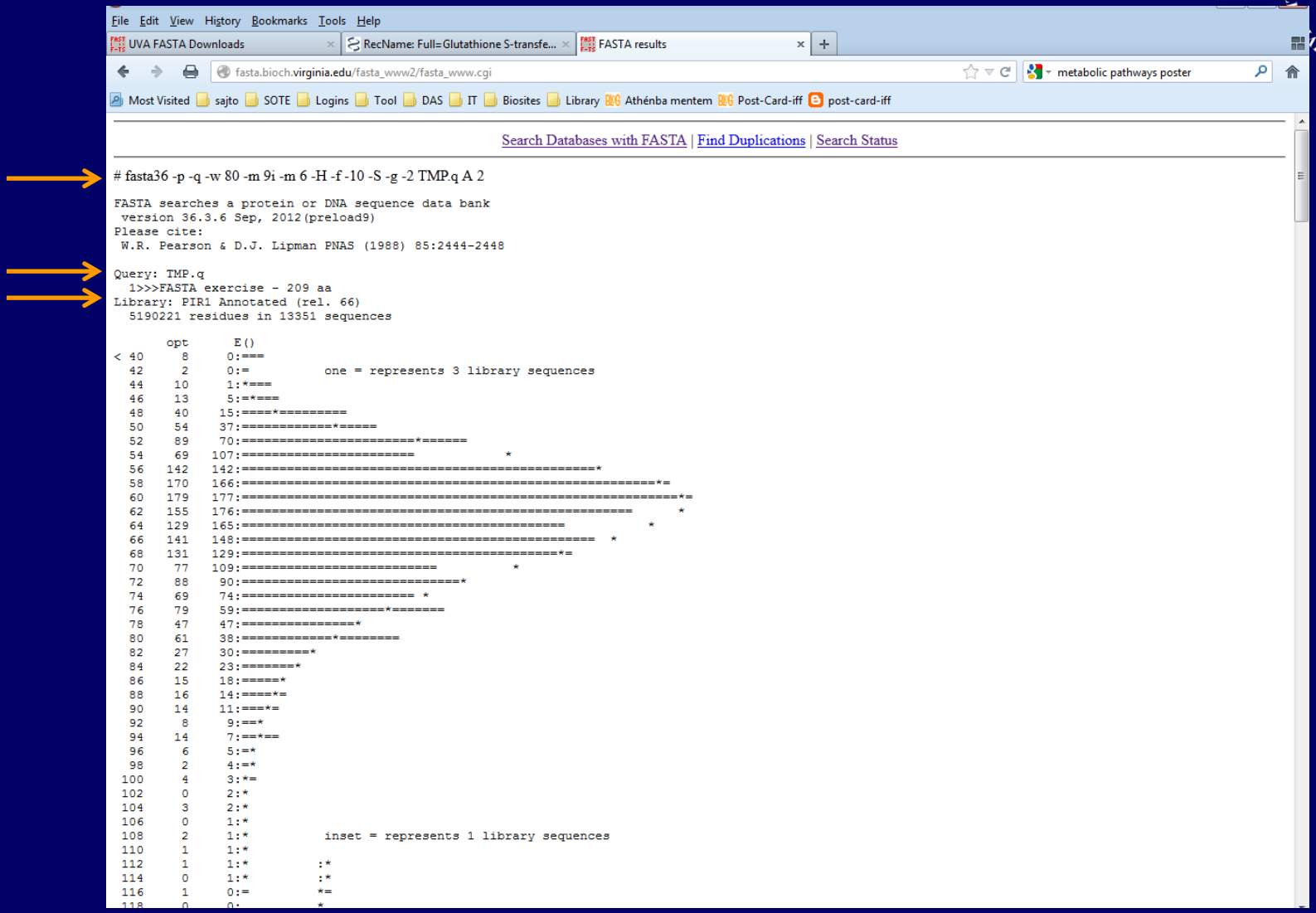
e. Try the search with **ssearch** (Smith-Waterman). Again, look at the E()-values for distant homologs and the highest scoring unrelated sequence.

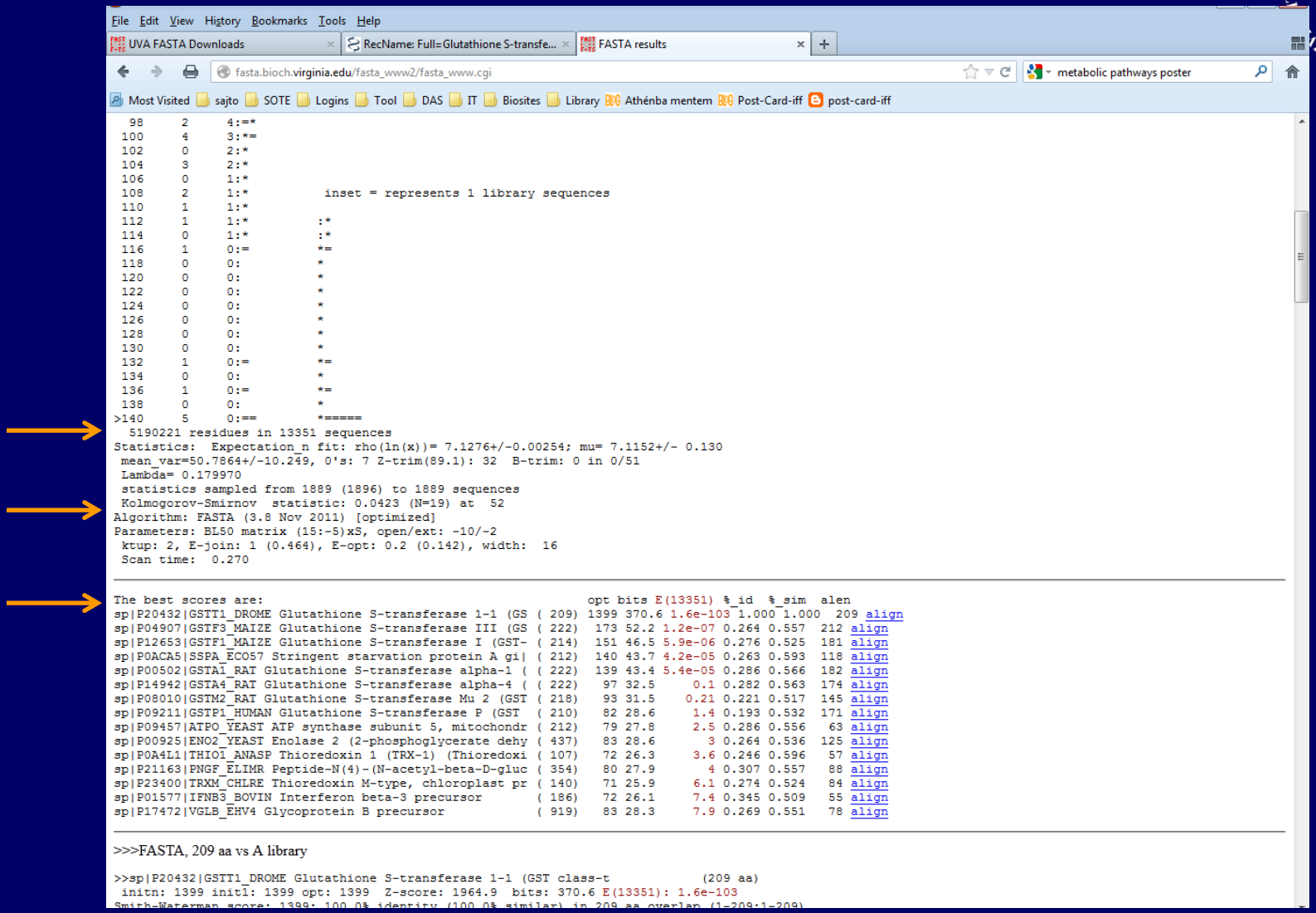
f. (*optional*) Try the search with *ktup=1* (What is ktup?). **FASTA** uses the *ktup* parameter to adjust the sensitivity and speed of the search. With *ktup=2*, **FASTA** looks for "pairs" of matched identical residues to find regions of similarity. *ktup=1* looks for singly-aligned residues, and thus takes longer.

---

2. Do the same search (`121694`) using the Course BLAST WWW page.


File Edit View History Bookmarks Tools Help
UVA FASTA Downloads
RecName: Full=Glutathione S-transfe...
FASTA Sequence Comparison
fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi
metabolic pathways poster
Most Visited sajto SOTE Logins Tool DAS IT Biosites Library Athénba mentem Post-Card-iff post-card-iff
Choose: (A) Program, (B) Query (sequence/accession), (C) Database and (D) Start Search:
Annotate Query Sequence
Annotate Database Sequences
(A) Program: FASTA: protein:protein
Compare your own sequences:
Compare sequences
(B) Query sequence: FASTA format
Subset range:
Use Subset range
>FASTA exercise
MVDFYYLPGSSPCRSVIMTAKAVGVELNKKLLNLQAGEHLKPEFLKINPQHTIPTLVDNG
FALWESRAIQ
VYLVEKYGKTDSLYPKCPKKRAVINQRLYFDMGTLYQSFANYYYPQVFAKAPADPEAFKK
IEAAFEFLNT
FLEGQDYAAGDSLTVADIALVATVSTFEVAKFEISKYANVNRWYENAKKVTPGWEENWAG
CLEFKKYFE
Entrez protein sequence browser
Entrez DNA sequence browser
Or upload query from file:
Browse...
Protein DNA (both-strands) DNA (forward only) DNA (rev-comp only)
(C) Database:
Protein
DNA
PIR1 Annotated (rel. 66)
GB170.0 Primate
Exclude low complexity (seg)
(D) Start Search
Search Database
Reset
Comments (optional):
Other search options:
Scoring matrix: open: ext: Ktup: Statistical estimates
Blosum50 (20%) -10 -2 ktup = 2 Default
Output limits: Show Histogram
E(): Best E():
Alignment Options: Highlight similarities differences compact differences.
FASTA program information | Download FASTA | About the Author
Copyright © 1988–2006 by William R. Pearson and the University of Virginia. All rights reserved. The FASTA program and documentation may not be sold or incorporated into a commercial product, in whole or in part, without written


File Edit View History Bookmarks Tools Help
UVA FASTA Downloads
RecName: Full=Glutathione S-transfe...
FASTA results
fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi
metabolic pathways poster
Most Visited sajto SOTE Logins Tool DAS IT Biosites Library Athénba mentem Post-Card-iff post-card-iff
Search Databases with FASTA | Find Duplications | Search Status
fasta36 -p -q -w 80 -m 9i -m 6 -H -f -10 -S -g -2 TMP.q A 2
FASTA searches a protein or DNA sequence data bank
version 36.3.6 Sep, 2012(preload9)
Please cite:
W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448
Query: TMP.q
1>>>FASTA exercise - 209 aa
Library: PIR1 Annotated (rel. 66)
5190221 residues in 13351 sequences
opt E()
< 40 8 0:===
42 2 0:= one = represents 3 library sequences
44 10 1:*===
46 13 5:=*===
48 40 15:====*==========
50 54 37:============*======
52 89 70:========================*=======
54 69 107:======================== *
56 142 142:==*
58 170 166:==*=
60 179 177:==*=
62 155 176:=== *
64 129 165:=== *
66 141 148:== *
68 131 129:===*=
70 77 109:========================== *
72 88 90:==============================*
74 69 74:======================== *
76 79 59:====================*=======
78 47 47:===============*
80 61 38:============*========
82 27 30:==========*
84 22 23:========*
86 15 18:======*
88 16 14:=====*=
90 14 11:====*=
92 8 9:==*
94 14 7:==*==
96 6 5:=*
98 2 4:=*
100 4 3:*=
102 0 2:*
104 3 2:*
106 0 1:*
108 2 1:* inset = represents 1 library sequences
110 1 1:*
112 1 1:* :*
114 0 1:* :*
116 1 0:= *=
118 0 0: *



File Edit View History Bookmarks Tools Help

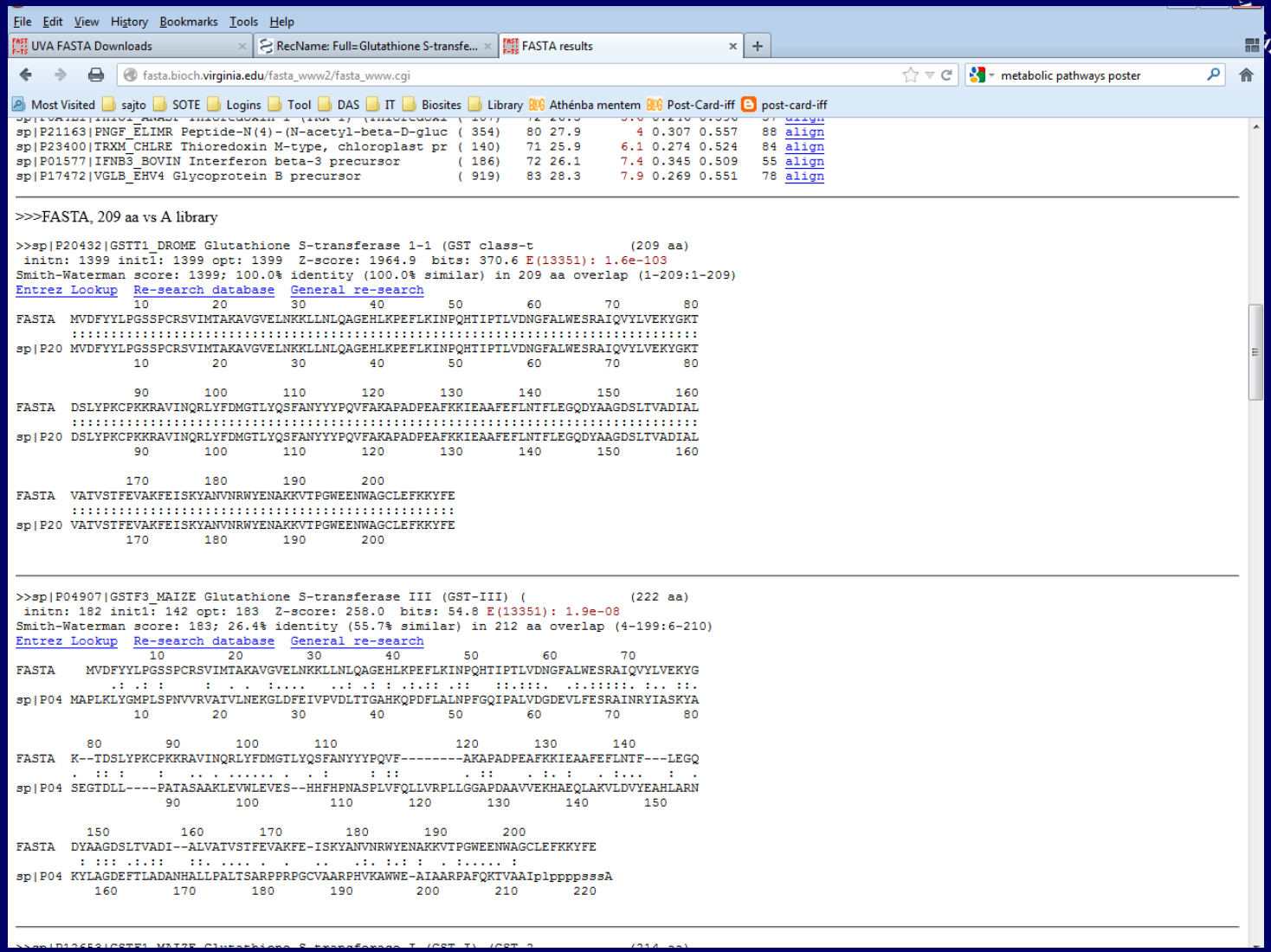
UVA FASTA Downloads | RecName: Full=Glutathione S-transfe... | FASTA results

fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi

metabolic pathways poster

Most Visited sajto SOTE Logins Tool DAS IT Biosites Library Athénba mentem Post-Card-iff post-card-iff

```
  98      2      4:=*
 100      4      3:*=
 102      0      2:*
 104      3      2:*
 106      0      1:*
 108      2      1:*          inset = represents 1 library sequences
 110      1      1:*
 112      1      1:*         :*
 114      0      1:*         :*
 116      1      0:=         *=
 118      0      0:          *
 120      0      0:          *
 122      0      0:          *
 124      0      0:          *
 126      0      0:          *
 128      0      0:          *
 130      0      0:          *
 132      1      0:=         *=
 134      0      0:          *
 136      1      0:=         *=
 138      0      0:          *
>140      5      0:==        *=====

  5190221 residues in 13351 sequences
Statistics:  Expectation_n fit: rho(ln(x))= 7.1276+/-0.00254; mu= 7.1152+/- 0.130
 mean_var=50.7864+/-10.249, 0's: 7 Z-trim(89.1): 32  B-trim: 0 in 0/51
 Lambda= 0.179970
 statistics sampled from 1889 (1896) to 1889 sequences
 Kolmogorov-Smirnov  statistic: 0.0423 (N=19) at  52
Algorithm: FASTA (3.8 Nov 2011) [optimized]
Parameters: BL50 matrix (15:-5)xS, open/ext: -10/-2
 ktup: 2, E-join: 1 (0.464), E-opt: 0.2 (0.142), width:  16
 Scan time:  0.270

The best scores are:                                      opt bits E(13351) %_id  %_sim  alen
sp|P20432|GSTT1_DROME Glutathione S-transferase 1-1 (GS ( 209) 1399 370.6 1.6e-103 1.000 1.000  209 align
sp|P04907|GSTF3_MAIZE Glutathione S-transferase III (GS ( 222)  173 52.2 1.2e-07 0.264 0.557  212 align
sp|P12653|GSTF1_MAIZE Glutathione S-transferase I (GST- ( 214)  151 46.5 5.9e-06 0.276 0.525  181 align
sp|P0ACA5|SSPA_ECO57 Stringent starvation protein A gi| ( 212)  140 43.7 4.2e-05 0.263 0.593  118 align
sp|P00502|GSTA1_RAT Glutathione S-transferase alpha-1 ( 222)  139 43.4 5.4e-05 0.286 0.566  182 align
sp|P14942|GSTA4_RAT Glutathione S-transferase alpha-4 ( 222)   97 32.5     0.1 0.282 0.563  174 align
sp|P08010|GSTM2_RAT Glutathione S-transferase Mu 2 (GST ( 218)   93 31.5    0.21 0.221 0.517  145 align
sp|P09211|GSTP1_HUMAN Glutathione S-transferase P (GST  ( 210)   82 28.6     1.4 0.193 0.532  171 align
sp|P09457|ATPO_YEAST ATP synthase subunit 5, mitochondr ( 212)   79 27.8     2.5 0.286 0.556   63 align
sp|P00925|ENO2_YEAST Enolase 2 (2-phosphoglycerate dehy ( 437)   83 28.6       3 0.264 0.536  125 align
sp|P0A4L1|THIO1_ANASP Thioredoxin 1 (TRX-1) (Thioredoxi ( 107)   72 26.3     3.6 0.246 0.596   57 align
sp|P21163|PNGF_ELIMR Peptide-N(4)-(N-acetyl-beta-D-gluc ( 354)   80 27.9       4 0.307 0.557   88 align
sp|P23400|TRXM_CHLRE Thioredoxin M-type, chloroplast pr ( 140)   71 25.9     6.1 0.274 0.524   84 align
sp|P01577|IFNB3_BOVIN Interferon beta-3 precursor       ( 186)   72 26.1     7.4 0.345 0.509   55 align
sp|P17472|VGLB_EHV4 Glycoprotein B precursor            ( 919)   83 28.3     7.9 0.269 0.551   78 align

>>>FASTA, 209 aa vs A library

>>sp|P20432|GSTT1_DROME Glutathione S-transferase 1-1 (GST class-t            (209 aa)
 initn: 1399 init1: 1399 opt: 1399  Z-score: 1964.9  bits: 370.6 E(13351): 1.6e-103
Smith-Waterman score: 1399; 100.0% identity (100.0% similar) in 209 aa overlap (1-209:1-209)
```



File Edit View History Bookmarks Tools Help

UVA FASTA Downloads | RecName: Full=Glutathione S-transfe... | FASTA results

fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi

metabolic pathways poster

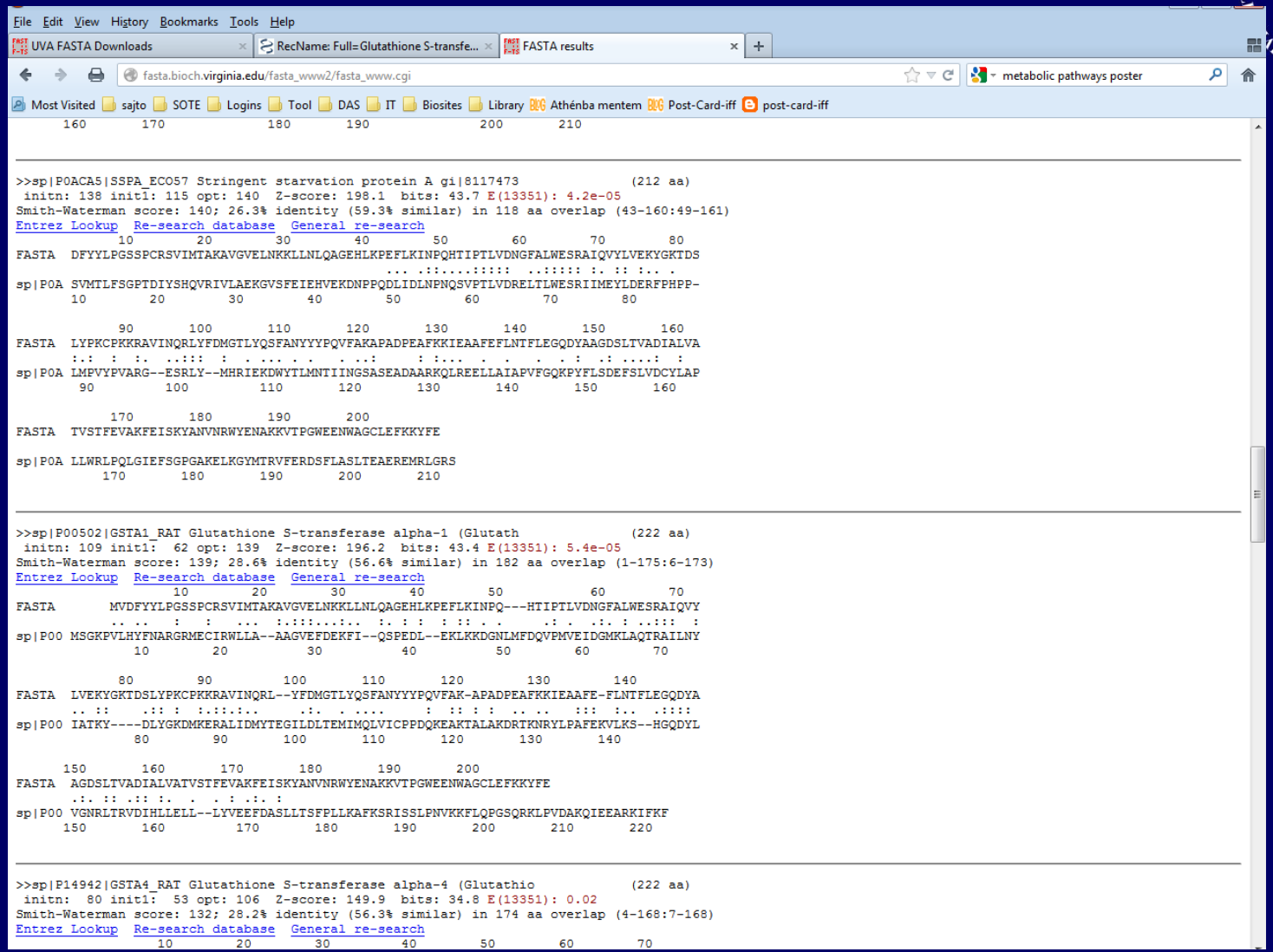
Most Visited sajto SOTE Logins Tool DAS IT Biosites Library Athénba mentem Post-Card-iff post-card-iff

```
sp|P21163|PNGF_ELIMR Peptide-N(4)-(N-acetyl-beta-D-gluc ( 354)   80 27.9      4 0.307 0.557   88 align
sp|P23400|TRXM_CHLRE Thioredoxin M-type, chloroplast pr ( 140)   71 25.9    6.1 0.274 0.524   84 align
sp|P01577|IFNB3_BOVIN Interferon beta-3 precursor       ( 186)   72 26.1    7.4 0.345 0.509   55 align
sp|P17472|VGLB_EHV4 Glycoprotein B precursor            ( 919)   83 28.3    7.9 0.269 0.551   78 align
```

>>>FASTA, 209 aa vs A library

```
>>sp|P20432|GSTT1_DROME Glutathione S-transferase 1-1 (GST class-t           (209 aa)
 initn: 1399 init1: 1399 opt: 1399  Z-score: 1964.9  bits: 370.6 E(13351): 1.6e-103
Smith-Waterman score: 1399; 100.0% identity (100.0% similar) in 209 aa overlap (1-209:1-209)
Entrez Lookup  Re-search database  General re-search
               10        20        30        40        50        60        70        80
FASTA  MVDFYYLPGSSPCRSVIMTAKAVGVELNKKLLNLQAGEHLKPEFLKINPQHTIPTLVDNGFALWESRAIQVYLVEKYGKT
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
sp|P20 MVDFYYLPGSSPCRSVIMTAKAVGVELNKKLLNLQAGEHLKPEFLKINPQHTIPTLVDNGFALWESRAIQVYLVEKYGKT
               10        20        30        40        50        60        70        80

               90       100       110       120       130       140       150       160
FASTA  DSLYPKCPKKRAVINQRLYFDMGTLYQSFANYYYPQVFAKAPADPEAFKKIEAAFEFLNTFLEGQDYAAGDSLTVADIAL
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
sp|P20 DSLYPKCPKKRAVINQRLYFDMGTLYQSFANYYYPQVFAKAPADPEAFKKIEAAFEFLNTFLEGQDYAAGDSLTVADIAL
               90       100       110       120       130       140       150       160

              170       180       190       200
FASTA  VATVSTFEVAKFEISKYANVNRWYENAKKVTPGWEENWAGCLEFKKYFE
       ::::::::::::::::::::::::::::::::::::::::::::::::::
sp|P20 VATVSTFEVAKFEISKYANVNRWYENAKKVTPGWEENWAGCLEFKKYFE
              170       180       190       200
```

```
>>sp|P04907|GSTF3_MAIZE Glutathione S-transferase III (GST-III) (           (222 aa)
 initn: 182 init1: 142 opt: 183  Z-score: 258.0  bits: 54.8 E(13351): 1.9e-08
Smith-Waterman score: 183; 26.4% identity (55.7% similar) in 212 aa overlap (4-199:6-210)
Entrez Lookup  Re-search database  General re-search
                10        20        30        40        50        60        70
FASTA    MVDFYYLPGSSPCRSVIMTAKAVGVELNKKLLNLQAGEHLKPEFLKINPQHTIPTLVDNGFALWESRAIQVYLVEKYG
            .: .: :     :  . .  :....  ..: .: : .:.:: .::   ::.:::.  .:.::::. :.. ::.
sp|P04 MAPLKLYGMPLSPNVVRVATVLNEKGLDFEIVPVDLTTGAHKQPDFLALNPFGQIPALVDGDEVLFESRAINRYIASKYA
                10        20        30        40        50        60        70        80

         80        90       100       110               120       130       140
FASTA  K--TDSLYPKCPKKRAVINQRLYFDMGTLYQSFANYYYPQVF--------AKAPADPEAFKKIEAAFEFLNTF---LEGQ
       .  :: :    :   .. . ..... .  . :     : ::         . ::    . :. :   . :...   : .
sp|P04 SEGTDLL----PATASAAKLEVWLEVES--HHFHPNASPLVFQLLVRPLLGGAPDAAVVEKHAEQLAKVLDVYEAHLARN
                90       100       110       120       130       140       150

         150         160       170       180       190       200
FASTA  DYAAGDSLTVADI--ALVATVSTFEVAKFE-ISKYANVNRWYENAKKVTPGWEENWAGCLEFKKYFE
        : ::: .:.::   ::. .... .  .   ..  .:. :.: :  . :..... :
sp|P04 KYLAGDEFTLADANHALLPALTSARPPRPGCVAARPHVKAWWE-AIAARPAFQKTVAAIplpppsssA
          160       170       180       190       200       210       220
```



File Edit View History Bookmarks Tools Help

UVA FASTA Downloads | RecName: Full=Glutathione S-transfe... | FASTA results

fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi

metabolic pathways poster

Most Visited sajto SOTE Logins Tool DAS IT Biosites Library Athénba mentem Post-Card-iff post-card-iff

```
       160       170       180       190       200       210

--------------------------------------------------------------------------------

>>sp|P0ACA5|SSPA_ECO57 Stringent starvation protein A gi|8117473             (212 aa)
 initn: 138 init1: 115 opt: 140  Z-score: 198.1  bits: 43.7 E(13351): 4.2e-05
Smith-Waterman score: 140; 26.3% identity (59.3% similar) in 118 aa overlap (43-160:49-161)
Entrez Lookup  Re-search database  General re-search
               10        20        30        40        50        60        70        80
FASTA  DFYYLPGSSPCRSVIMTAKAVGVELNKKLLNLQAGEHLKPEFLKINPQHTIPTLVDNGFALWESRAIQVYLVEKYGKTDS
                                                  ... .::....::::  ..:::: :. :: :.. .
sp|P0A SVMTLFSGPTDIYSHQVRIVLAEKGVSFEIEHVEKDNPPQDLIDLNPNQSVPTLVDRELTLWESRIIMEYLDERFPHPP-
       10        20        30        40        50        60        70        80

              90       100       110       120       130       140       150       160
FASTA  LYPKCPKKRAVINQRLYFDMGTLYQSFANYYYPQVFAKAPADPEAFKKIEAAFEFLNTFLEGQDYAAGDSLTVADIALVA
       :.:  :  :.  ..:::  :  . ... . .   . ..:     : :...  .  .   .  . :  .: ....:  :
sp|P0A LMPVYPVARG--ESRLY--MHRIEKDWYTLMNTIINGSASEADAARKQLREELLAIAPVFGQKPYFLSDEFSLVDCYLAP
        90        100       110       120       130       140       150       160

             170       180       190       200
FASTA  TVSTFEVAKFEISKYANVNRWYENAKKVTPGWEENWAGCLEFKKYFE

sp|P0A LLWRLPQLGIEFSGPGAKELKGYMTRVFERDSFLASLTEAEREMRLGRS
            170       180       190       200       210

--------------------------------------------------------------------------------

>>sp|P00502|GSTA1_RAT Glutathione S-transferase alpha-1 (Glutath             (222 aa)
 initn: 109 init1:  62 opt: 139  Z-score: 196.2  bits: 43.4 E(13351): 5.4e-05
Smith-Waterman score: 139; 28.6% identity (56.6% similar) in 182 aa overlap (1-175:6-173)
Entrez Lookup  Re-search database  General re-search
                  10        20        30        40        50        60        70
FASTA       MVDFYYLPGSSPCRSVIMTAKAVGVELNKKLLNLQAGEHLKPEFLKINPQ---HTIPTLVDNGFALWESRAIQVY
            .. ..   :   :   ...  :.:::...:..  :. : :  : :: . .     .: .  .:. : ..:::  :
sp|P00 MSGKPVLHYFNARGRMECIRWLLA--AAGVEFDEKFI--QSPEDL--EKLKKDGNLMFDQVPMVEIDGMKLAQTRAILNY
               10        20        30          40          50        60        70

              80        90         100       110       120       130       140
FASTA  LVEKYGKTDSLYPKCPKKRAVINQRL--YFDMGTLYQSFANYYYPQVFAK-APADPEAFKKIEAAFE-FLNTFLEGQDYA
       .. ::    .:: :  :.::.:..     .:.  . ....     :  :: : :  .. ..   :::  :..  .::::
sp|P00 IATKY----DLYGKDMKERALIDMYTEGILDLTEMIMQLVICPPDQKEAKTALAKDRTKNRYLPAFEKVLKS--HGQDYL
                80        90        100       110       120       130       140

      150       160       170       180       190       200
FASTA  AGDSLTVADIALVATVSTFEVAKFEISKYANVNRWYENAKKVTPGWEENWAGCLEFKKYFE
       .:. :: .:: :.  .  . : .:. :
sp|P00 VGNRLTRVDIHLLELL--LYVEEFDASLLTSFPLLKAFKSRISSLPNVKKFLQPGSQRKLPVDAKQIEEARKIFKF
      150       160       170       180       190       200       210       220

--------------------------------------------------------------------------------

>>sp|P14942|GSTA4_RAT Glutathione S-transferase alpha-4 (Glutathio            (222 aa)
 initn:  80 init1:  53 opt: 106  Z-score: 149.9  bits: 34.8 E(13351): 0.02
Smith-Waterman score: 132; 28.2% identity (56.3% similar) in 174 aa overlap (4-168:7-168)
Entrez Lookup  Re-search database  General re-search
                  10        20        30        40        50        60        70
```

A BLAST család

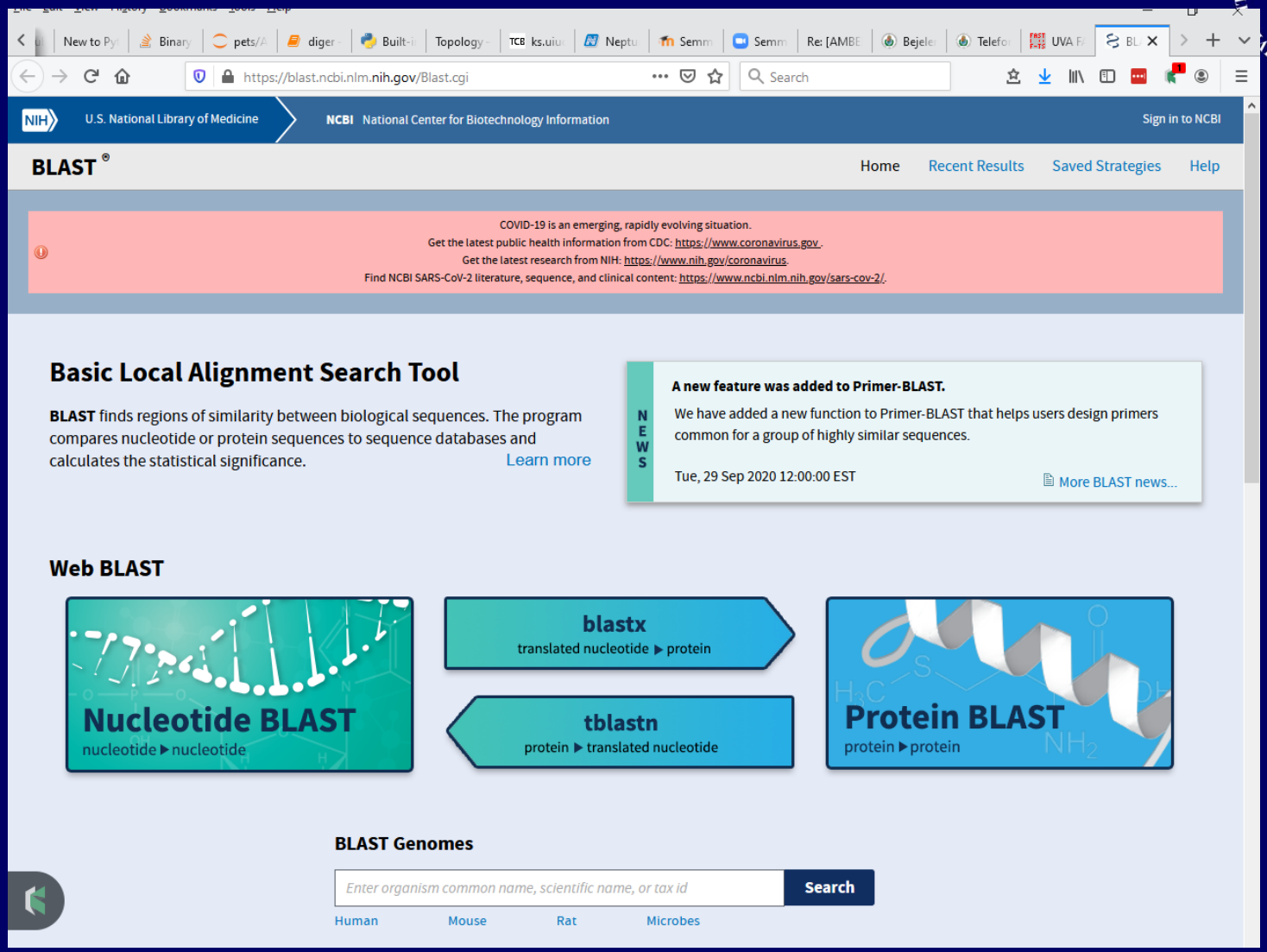
- Honlap: http://blast.ncbi.nlm.nih.gov/Blast.cgi
- Letülthető Linuxra és Windowsra is
- Web alapú szolgáltatás elérhető
- Alapos dokumentáció a honlapon
- Igen népszerű, megbízható
- Régóta van jelen az irodalomban, folyamatosan fejlesztik

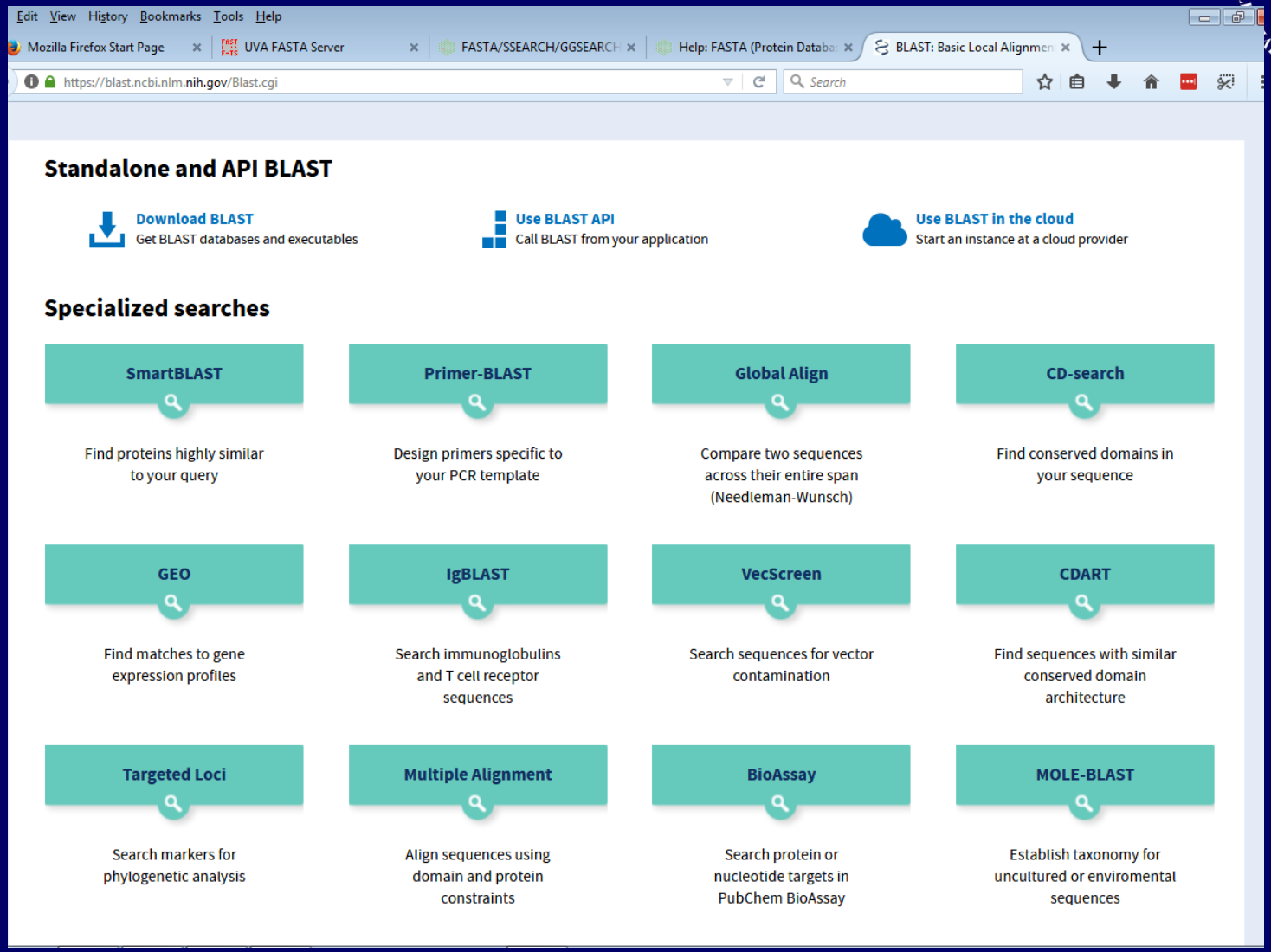
Az algoritmus

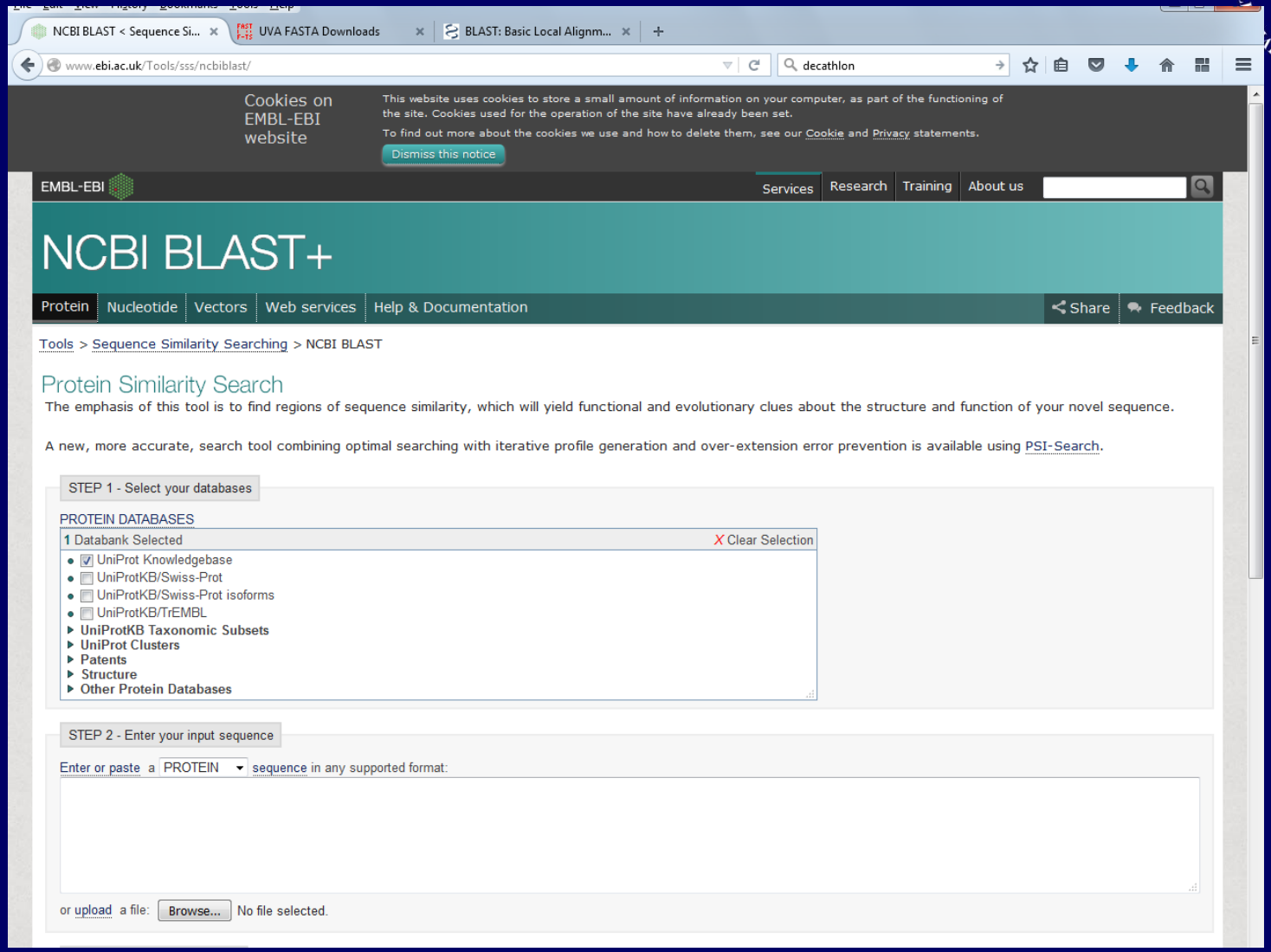
1. A szekvencia maszkolása
2. A szekvencia felbontása szavakra
3. A lista szűkítése a nagy pontértékű szavakra
4. Keresés az adatbázisban a lista alapján
5. A találatok kiterjesztése (HSP – high-scoring segment pair)
6. A HSP-k statisztikus kiértékelése
7. HSP-k összefűzése hosszabb illesztéssé

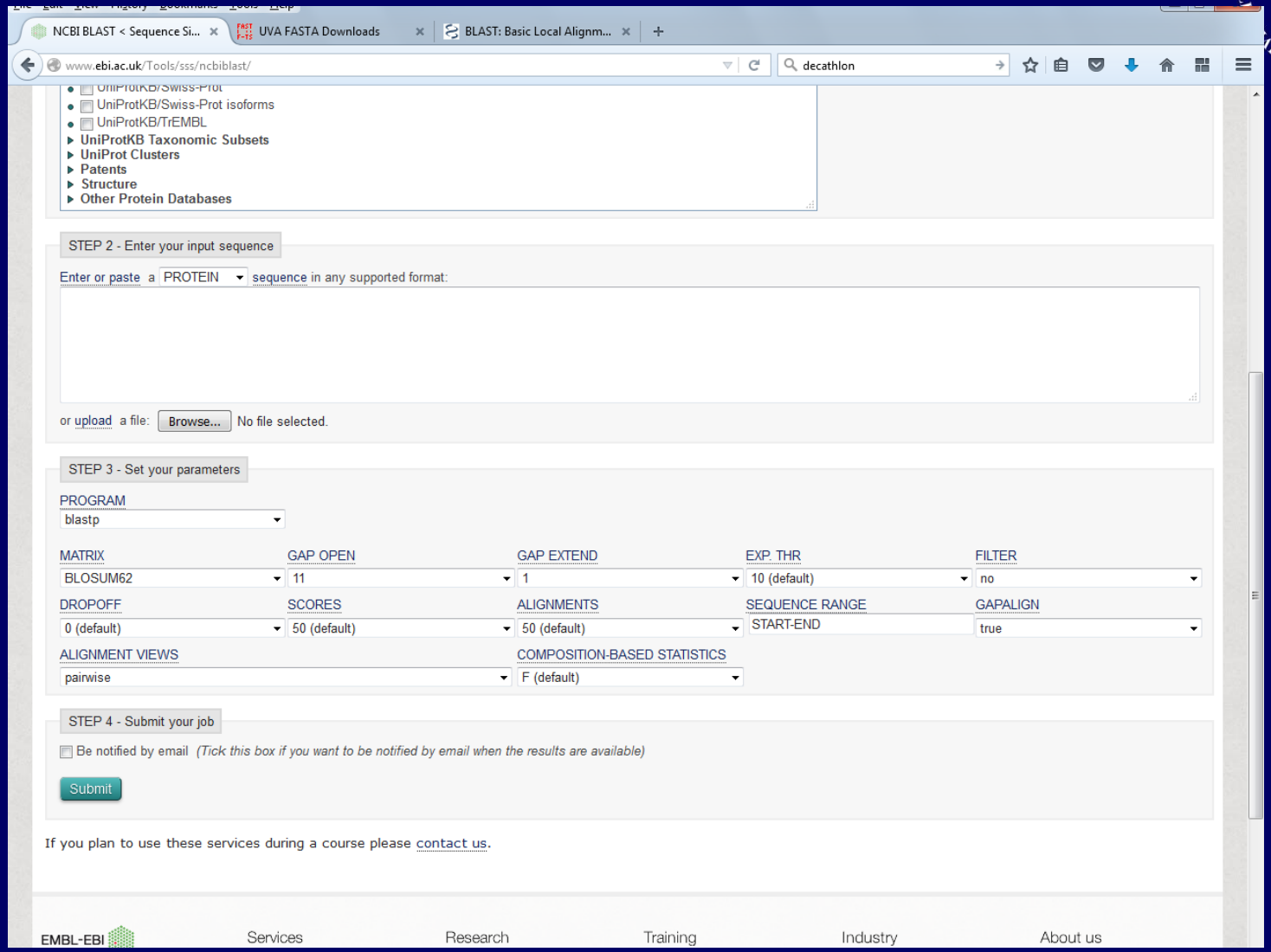
A programcsomag tagjai

- „blastn”: DNS szekvencia keresése DNS adatbázis ellen
- „blastp”: fehérje szekvencia fehérje adatbázis ellen
- „psi-blast”: fehérjék iteratív keresése fehérje adatbázis ellen
- „blastx”: lefordított DNS szekvencia keresése fehérje adatbázis ellen
- „tblastx”: lefordított DNS szekvencia keresés lefordított DNS adatbázison
- „tblastn”: visszafordított fehérje keresés DNS adatbázis ellen
- „megablast”: sok szekvencia keresése egy futás során


https://blast.ncbi.nlm.nih.gov/Blast.cgi
Search
NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Sign in to NCBI
BLAST ®
Home
Recent Results
Saved Strategies
Help
COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.
Basic Local Alignment Search Tool
BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.
Learn more
NEWS
A new feature was added to Primer-BLAST.
We have added a new function to Primer-BLAST that helps users design primers common for a group of highly similar sequences.
Tue, 29 Sep 2020 12:00:00 EST
More BLAST news...
Web BLAST
Nucleotide BLAST
nucleotide ▶ nucleotide
blastx
translated nucleotide ▶ protein
tblastn
protein ▶ translated nucleotide
Protein BLAST
protein ▶ protein
BLAST Genomes
Enter organism common name, scientific name, or tax id
Search
Human
Mouse
Rat
Microbes


Edit View History Bookmarks Tools Help
Mozilla Firefox Start Page
UVA FASTA Server
FASTA/SSEARCH/GGSEARCH
Help: FASTA (Protein Databa
BLAST: Basic Local Alignmen
https://blast.ncbi.nlm.nih.gov/Blast.cgi
Search
Standalone and API BLAST
Download BLAST
Get BLAST databases and executables
Use BLAST API
Call BLAST from your application
Use BLAST in the cloud
Start an instance at a cloud provider
Specialized searches
SmartBLAST
Find proteins highly similar to your query
Primer-BLAST
Design primers specific to your PCR template
Global Align
Compare two sequences across their entire span (Needleman-Wunsch)
CD-search
Find conserved domains in your sequence
GEO
Find matches to gene expression profiles
IgBLAST
Search immunoglobulins and T cell receptor sequences
VecScreen
Search sequences for vector contamination
CDART
Find sequences with similar conserved domain architecture
Targeted Loci
Search markers for phylogenetic analysis
Multiple Alignment
Align sequences using domain and protein constraints
BioAssay
Search protein or nucleotide targets in PubChem BioAssay
MOLE-BLAST
Establish taxonomy for uncultured or enviromental sequences


NCBI BLAST < Sequence Si...
UVA FASTA Downloads
BLAST: Basic Local Alignm...
www.ebi.ac.uk/Tools/sss/ncbiblast/
decathlon
Cookies on EMBL-EBI website
This website uses cookies to store a small amount of information on your computer, as part of the functioning of the site. Cookies used for the operation of the site have already been set.
To find out more about the cookies we use and how to delete them, see our Cookie and Privacy statements.
Dismiss this notice
EMBL-EBI
Services
Research
Training
About us
NCBI BLAST+
Protein
Nucleotide
Vectors
Web services
Help & Documentation
Share
Feedback
Tools > Sequence Similarity Searching > NCBI BLAST
Protein Similarity Search
The emphasis of this tool is to find regions of sequence similarity, which will yield functional and evolutionary clues about the structure and function of your novel sequence.
A new, more accurate, search tool combining optimal searching with iterative profile generation and over-extension error prevention is available using PSI-Search.
STEP 1 - Select your databases
PROTEIN DATABASES
1 Databank Selected
X Clear Selection
UniProt Knowledgebase
UniProtKB/Swiss-Prot
UniProtKB/Swiss-Prot isoforms
UniProtKB/TrEMBL
UniProtKB Taxonomic Subsets
UniProt Clusters
Patents
Structure
Other Protein Databases
STEP 2 - Enter your input sequence
Enter or paste a PROTEIN sequence in any supported format:
or upload a file: Browse... No file selected.


NCBI BLAST < Sequence Si...
UVA FASTA Downloads
BLAST: Basic Local Alignm...
www.ebi.ac.uk/Tools/sss/ncbiblast/
decathlon
UniProtKB/Swiss-Prot
UniProtKB/Swiss-Prot isoforms
UniProtKB/TrEMBL
UniProtKB Taxonomic Subsets
UniProt Clusters
Patents
Structure
Other Protein Databases
STEP 2 - Enter your input sequence
Enter or paste a PROTEIN sequence in any supported format:
or upload a file: Browse... No file selected.
STEP 3 - Set your parameters
PROGRAM
blastp
MATRIX
BLOSUM62
GAP OPEN
11
GAP EXTEND
1
EXP. THR
10 (default)
FILTER
no
DROPOFF
0 (default)
SCORES
50 (default)
ALIGNMENTS
50 (default)
SEQUENCE RANGE
START-END
GAPALIGN
true
ALIGNMENT VIEWS
pairwise
COMPOSITION-BASED STATISTICS
F (default)
STEP 4 - Submit your job
Be notified by email (Tick this box if you want to be notified by email when the results are available)
Submit
If you plan to use these services during a course please contact us.
EMBL-EBI
Services
Research
Training
Industry
About us

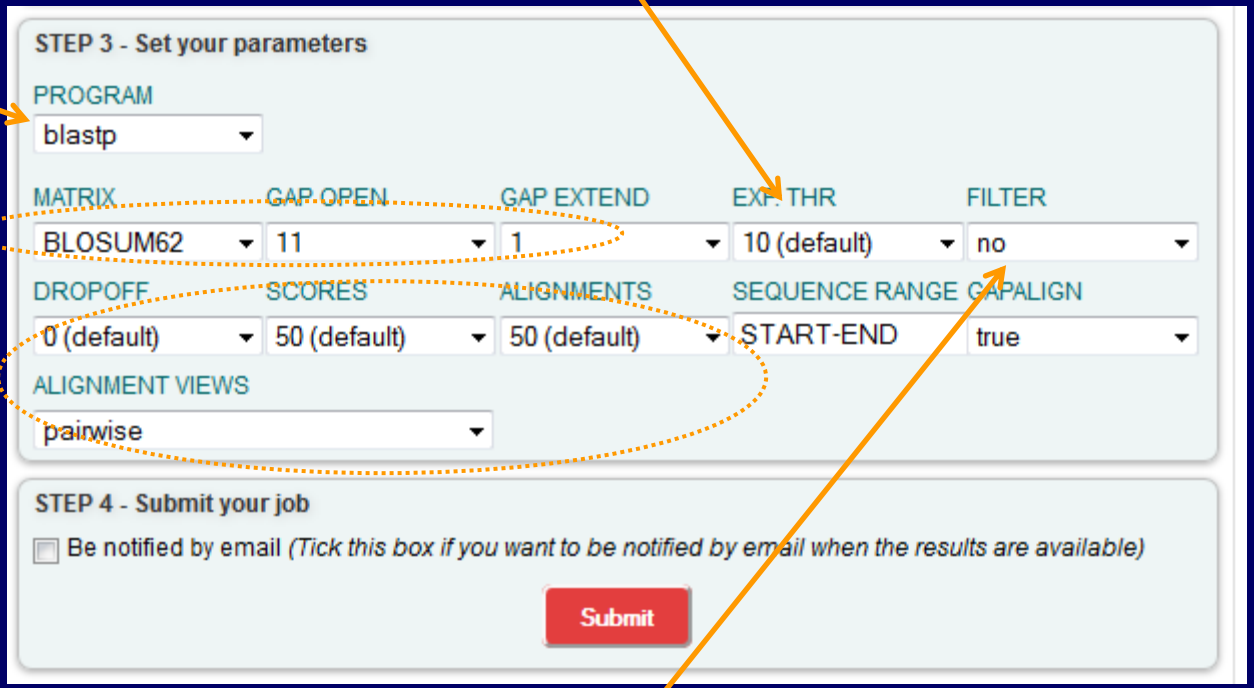
Paraméterek

A program neve

Mátrix paraméterek

Eredmény

Statisztika



Maszkolás

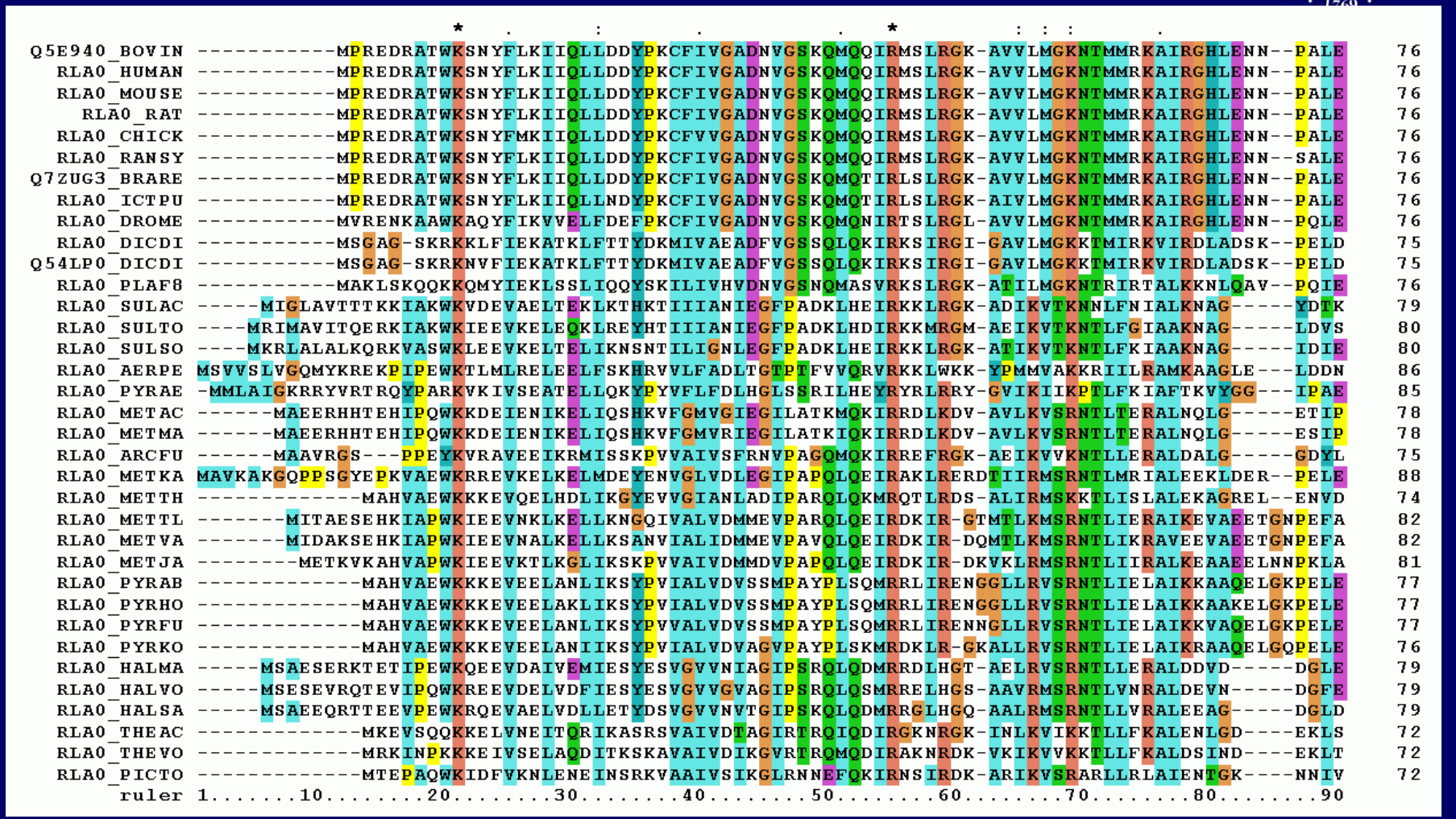
Egy további változat

BLAT:

- A cél a további gyorsítás
- Az ár: rosszabb érzékenység
  - Csak a nagyon hasonló szegmenseket talalálja meg: 95% egyezés DNS-re, 80% fehérjére
  - Rövid egyezéseket nem talál meg
- Új generációs szekvenálásnál igen hasznos

Profilkeresés

- Az aminósavak helyettesítési hajlandósága függ a pozíciótól
- Egy jó többszörös illesztés megadja ezt az információt
- A többszörös illesztést közvetlenül „profillá” alakítjuk
- Ezt használjuk a kereséshez
- Megnő az eljárás érzékenysége


Q5E940_BOVIN
RLA0_HUMAN
RLA0_MOUSE
RLA0_RAT
RLA0_CHICK
RLA0_RANSY
Q7ZUG3_BRARE
RLA0_ICTPU
RLA0_DROME
RLA0_DICDI
Q54LP0_DICDI
RLA0_PLAF8
RLA0_SULAC
RLA0_SULTO
RLA0_SULSO
RLA0_AERPE
RLA0_PYRAE
RLA0_METAC
RLA0_METMA
RLA0_ARCFU
RLA0_METKA
RLA0_METTH
RLA0_METTL
RLA0_METVA
RLA0_METJA
RLA0_PYRAB
RLA0_PYRHO
RLA0_PYRFU
RLA0_PYRKO
RLA0_HALMA
RLA0_HALVO
RLA0_HALSA
RLA0_THEAC
RLA0_THEVO
RLA0_PICTO
ruler
1.....10.....20.....30.....40.....50.....60.....70.....80.....90

	1	2	3	4	5	6	7	8	9	...
A	a_1									
C										
D										
E										
F										
...										

	1	2	3	4	5	6	7	8	9	...
A	a_1									
C	c_1									
D										
E										
F										
...										

	1	2	3	4	5	6	7	8	9	...
A	a_1									
C	c_1									
D	d_1									
E	e_1									
F	f_1									
...										

	1	2	3	4	5	6	7	8	9	...
A	a_1	a_2								
C	c_1	c_2								
D	d_1	d_2								
E	e_1	e_2								
F	f_1	f_2								
...										

	1	2	3	4	5	6	7	8	9	...
A	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	
C	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	
D	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	
E	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	
F	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	
...										

Megvalósítás: PSI-BLAST

- Egy BLAST kereséssel megtaláljuk a közeli homológokat
- Ezekből többszörös illesztést készítünk
- Ebből származtatjuk a kerső profilt a következő BLAST kereséshez
- Ezzel bővül a homológok listája a távolabbi rokonokkal
- Az eljárást ismételjük

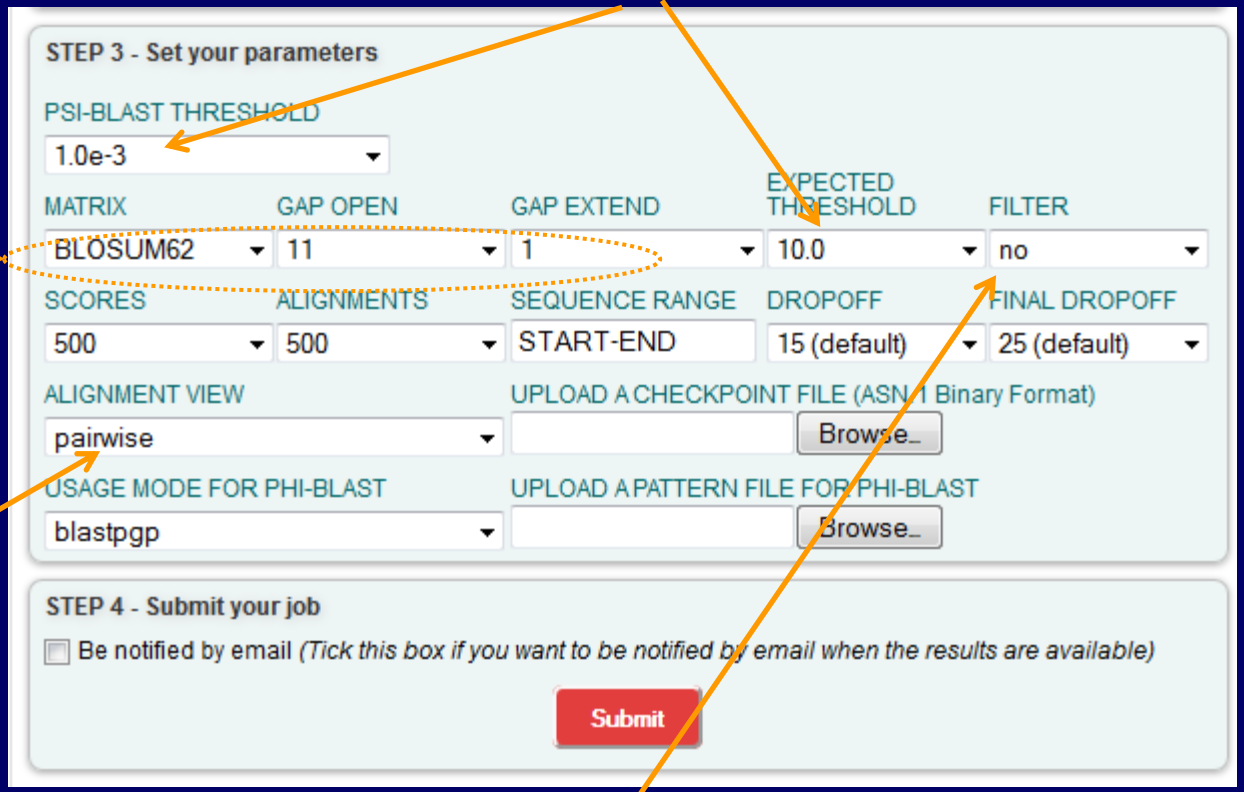
Paraméterek

Statisztika

Mátrix paraméterek

Eredmény

Maszkolás

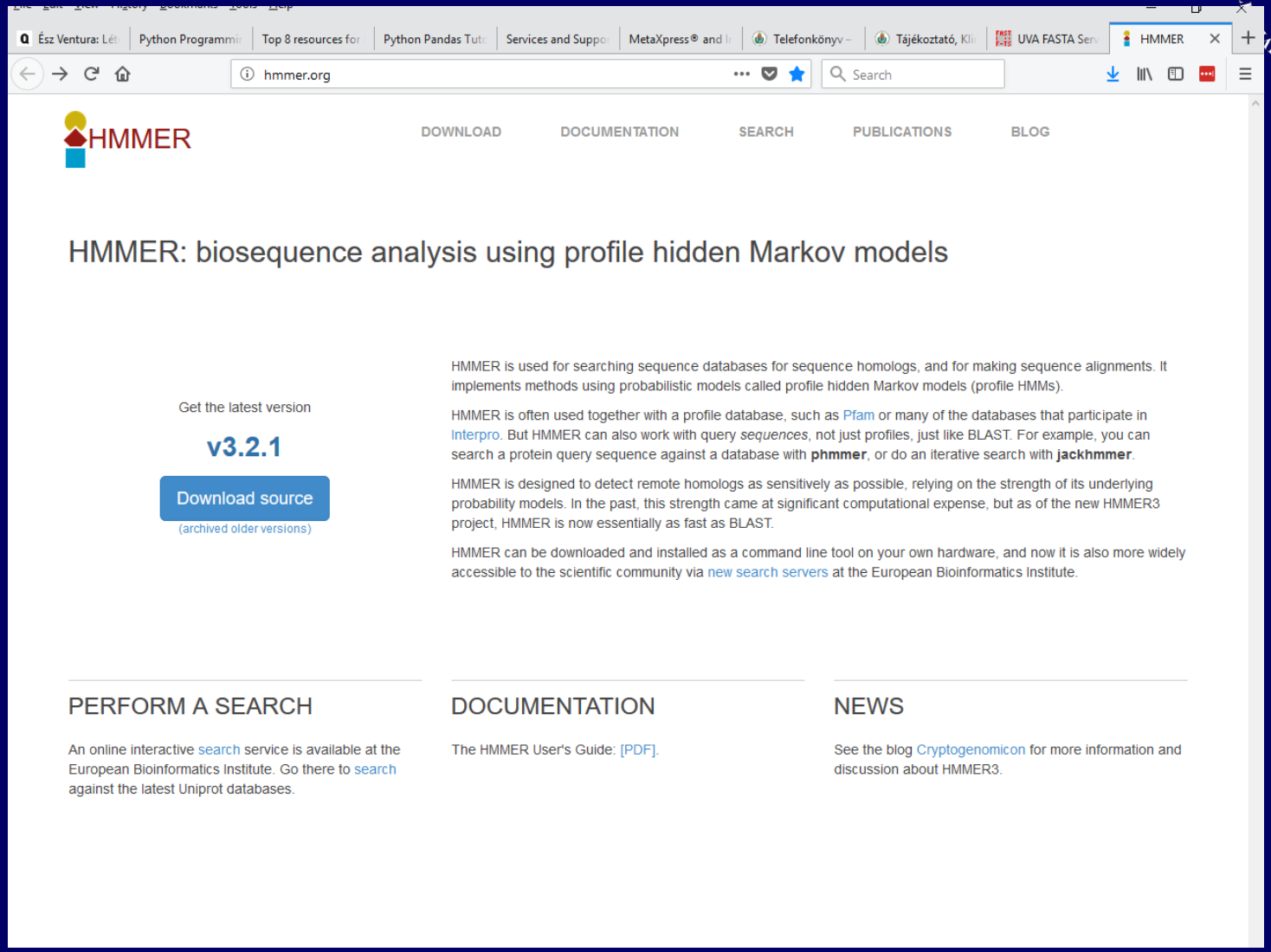


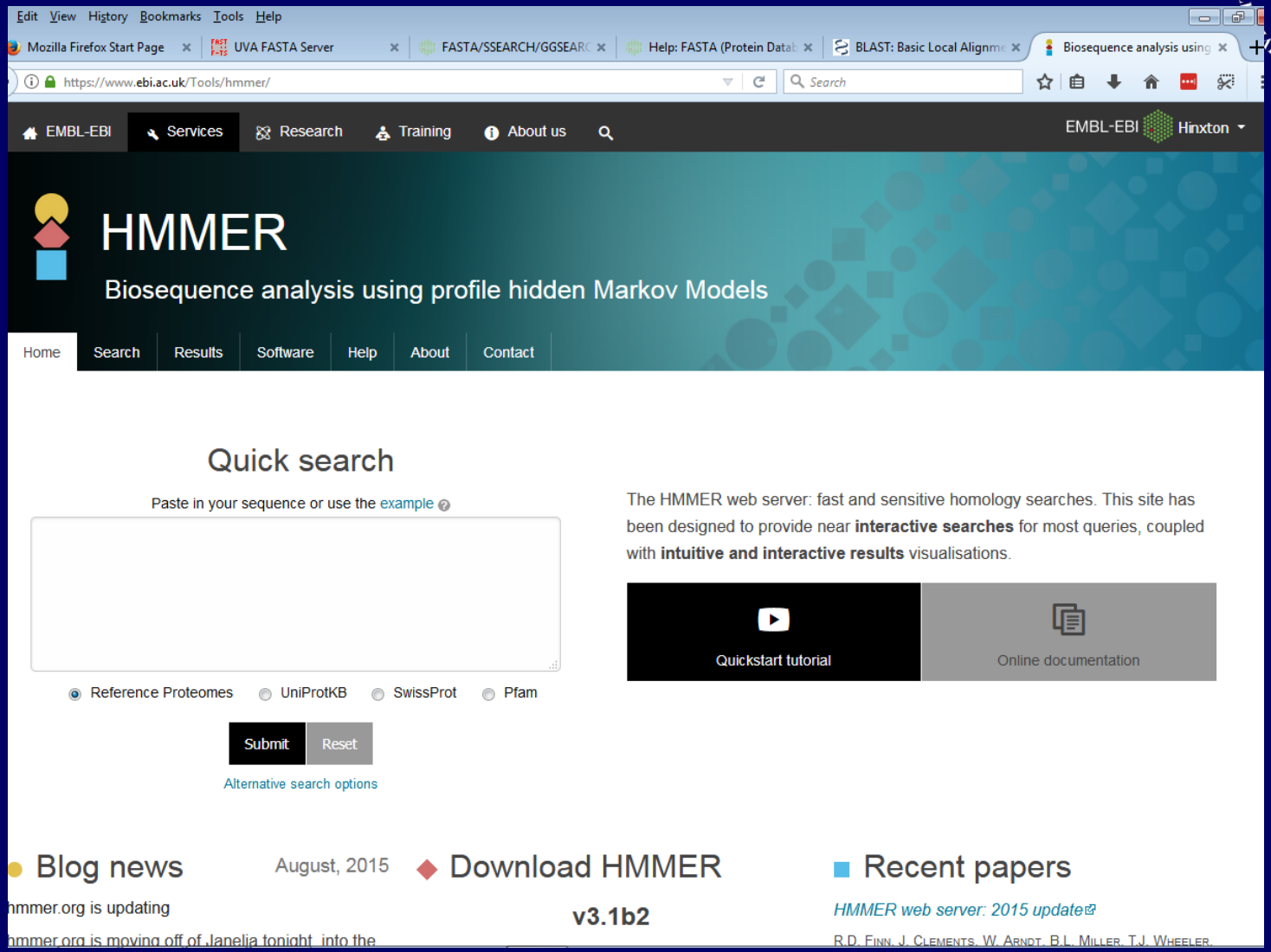
Másik lehetőség HMMER

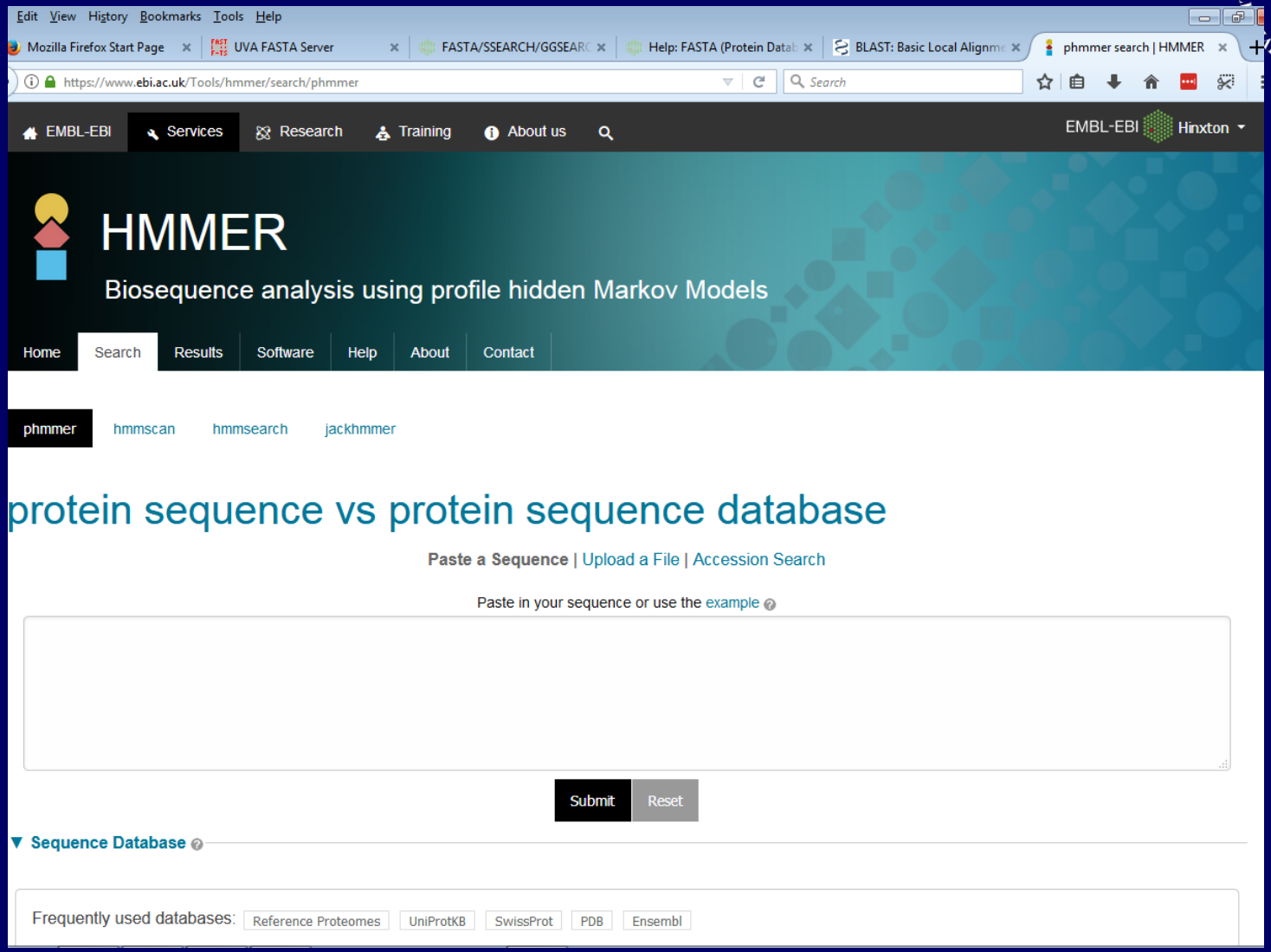
- Honlap: http://hmmer.org/
- Letölthető Linux és Windows verzióban
- Bőséges dokumentáció a honlapon
- Fehérje szekvenciákat kezel
- Egyszerre gyors és pontos módszer
- Egy hatékony statisztikai modellen alapul – „Markov modell”

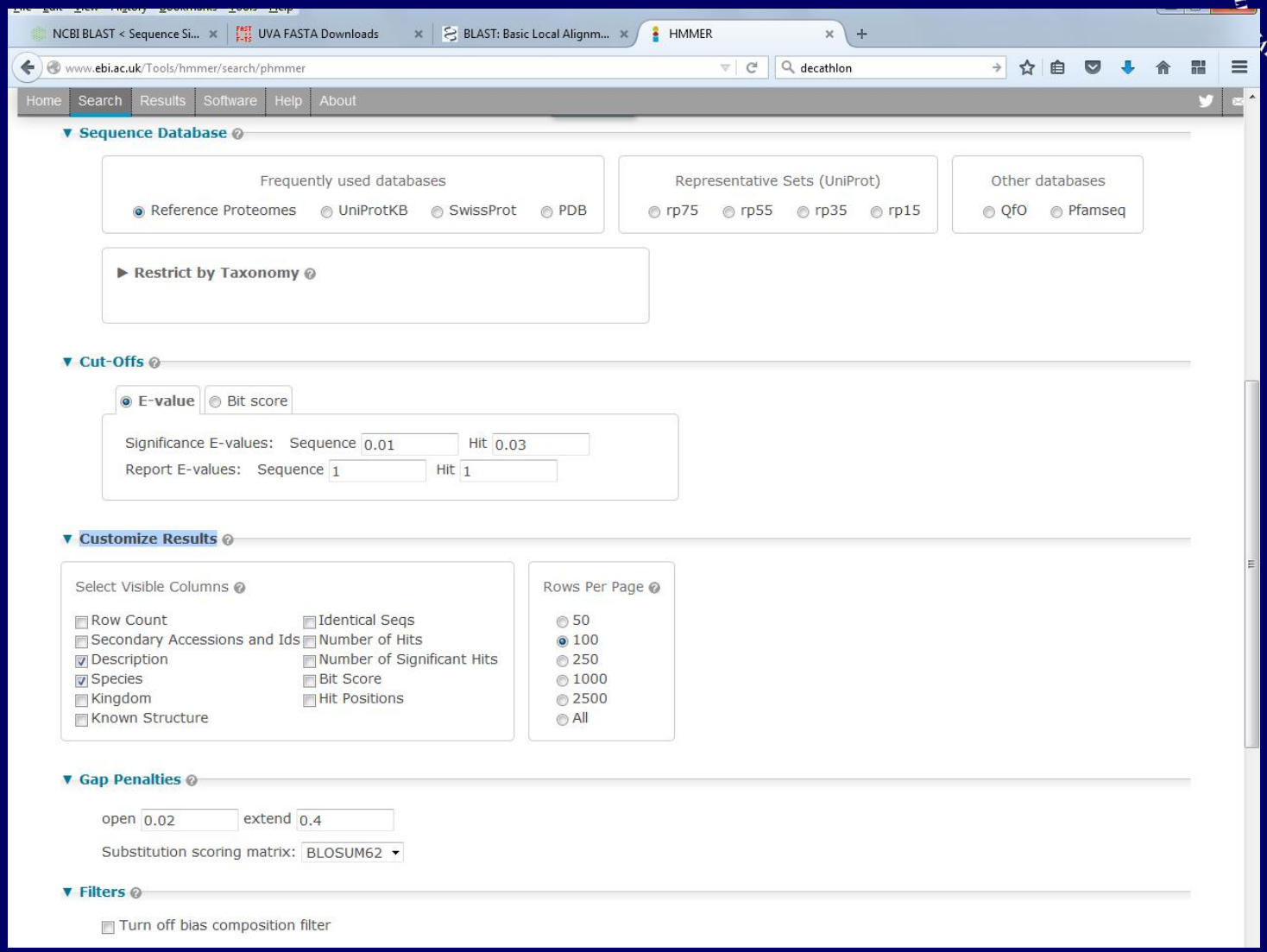
A programcsalád tagjai

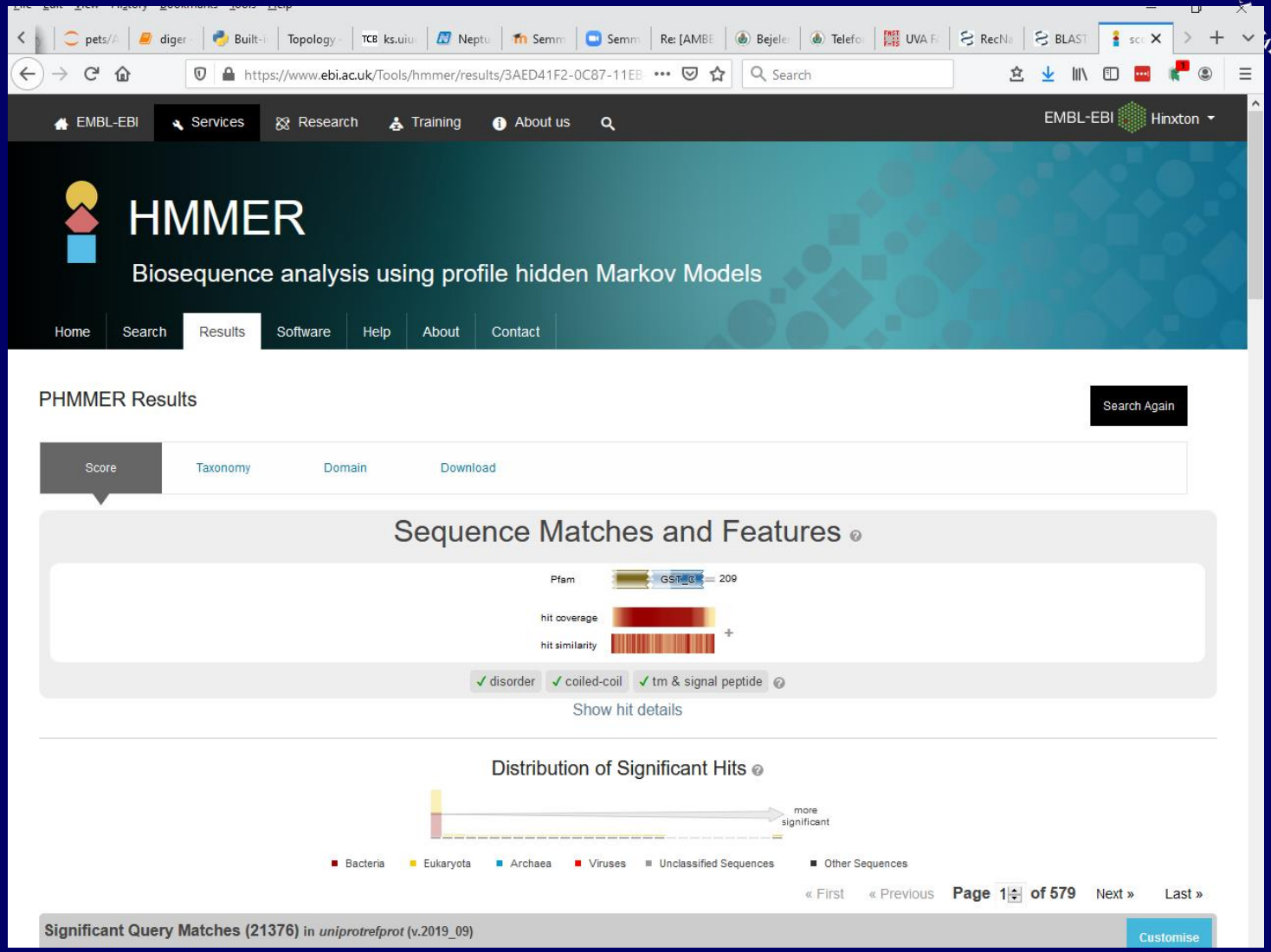
- „phmmer”: egy vagy több fehérje szekvenciát keres a fehérje adatbázis ellenében
- „hmmscan”: fehérje szekvenciákat keres profil adatbázis ellen
- „hmmsearch”: profilokat keres fehérje adatbázis ellenében
- “jackhmmer”: interaktív változat

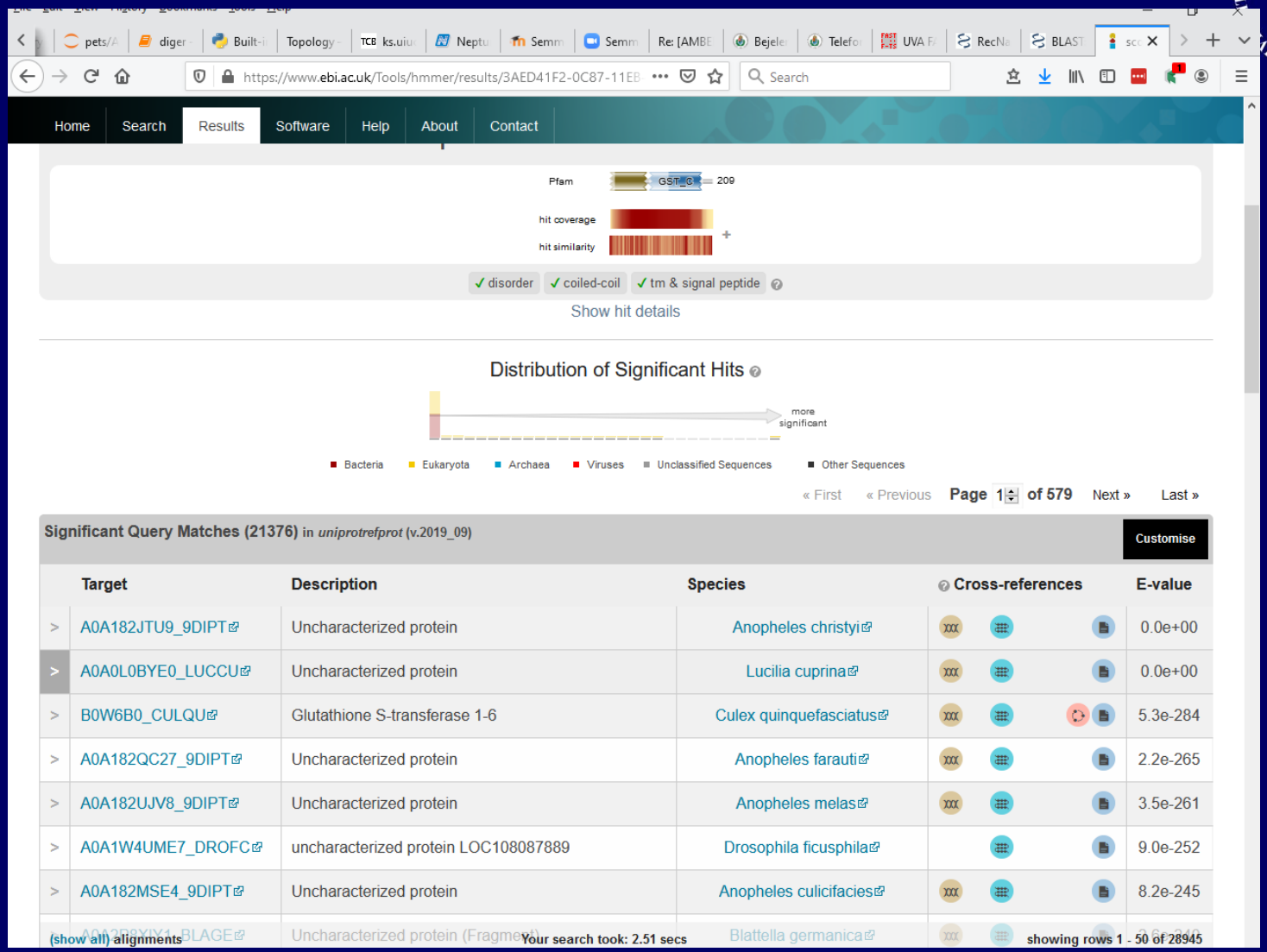

HMMER
hmmer.org
Search
DOWNLOAD
DOCUMENTATION
SEARCH
PUBLICATIONS
BLOG
HMMER: biosequence analysis using profile hidden Markov models
Get the latest version
v3.2.1
Download source
(archived older versions)
HMMER is used for searching sequence databases for sequence homologs, and for making sequence alignments. It implements methods using probabilistic models called profile hidden Markov models (profile HMMs).
HMMER is often used together with a profile database, such as Pfam or many of the databases that participate in Interpro. But HMMER can also work with query sequences, not just profiles, just like BLAST. For example, you can search a protein query sequence against a database with phmmer, or do an iterative search with jackhmmer.
HMMER is designed to detect remote homologs as sensitively as possible, relying on the strength of its underlying probability models. In the past, this strength came at significant computational expense, but as of the new HMMER3 project, HMMER is now essentially as fast as BLAST.
HMMER can be downloaded and installed as a command line tool on your own hardware, and now it is also more widely accessible to the scientific community via new search servers at the European Bioinformatics Institute.
PERFORM A SEARCH
An online interactive search service is available at the European Bioinformatics Institute. Go there to search against the latest Uniprot databases.
DOCUMENTATION
The HMMER User's Guide: [PDF].
NEWS
See the blog Cryptogenomicon for more information and discussion about HMMER3.

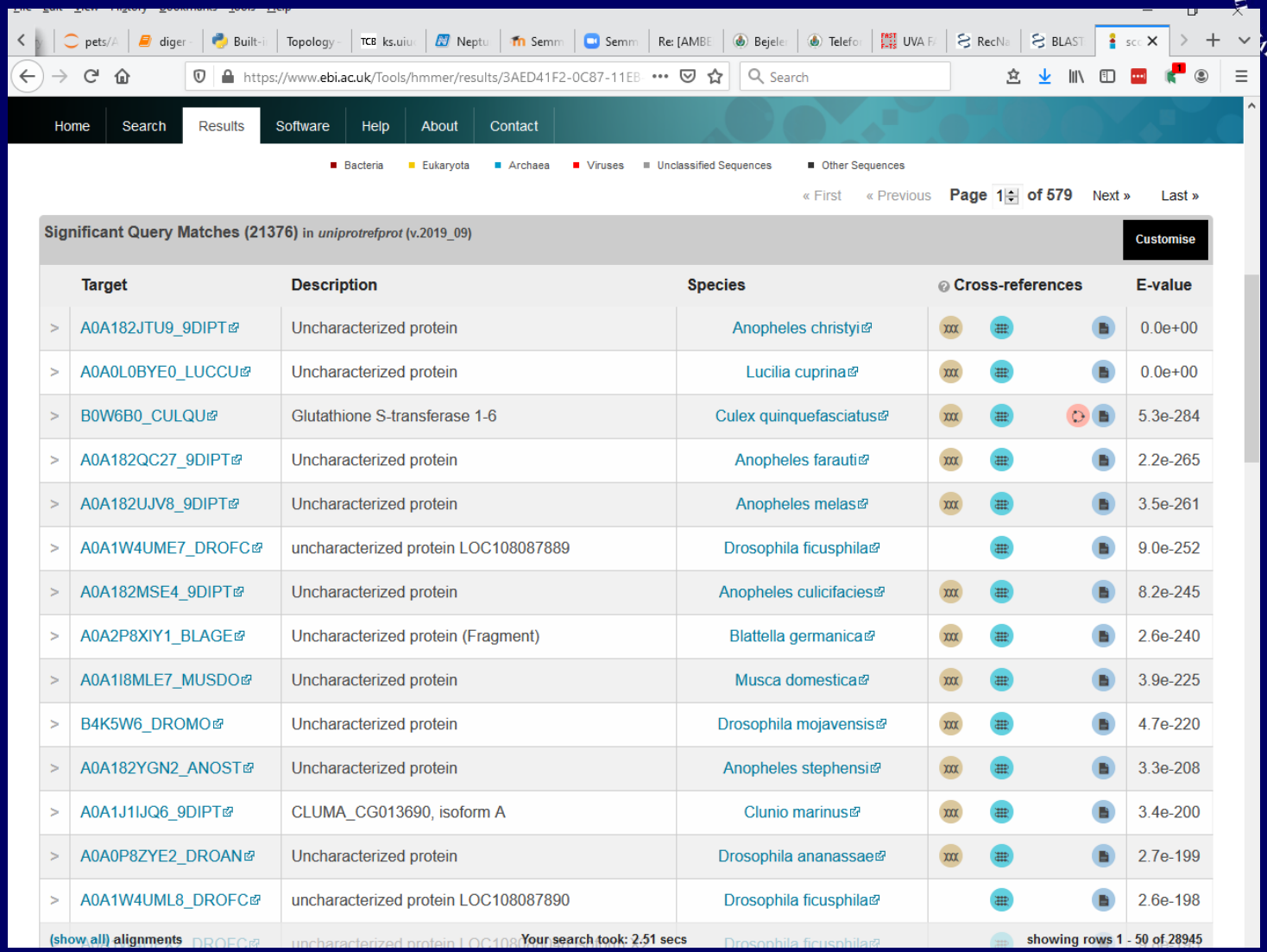

Edit View History Bookmarks Tools Help
Mozilla Firefox Start Page
UVA FASTA Server
FASTA/SSEARCH/GGSEARC
Help: FASTA (Protein Datab
BLAST: Basic Local Alignme
Biosequence analysis using
https://www.ebi.ac.uk/Tools/hmmer/
Search
EMBL-EBI
Services
Research
Training
About us
EMBL-EBI
Hinxton
HMMER
Biosequence analysis using profile hidden Markov Models
Home
Search
Results
Software
Help
About
Contact
Quick search
Paste in your sequence or use the example
Reference Proteomes
UniProtKB
SwissProt
Pfam
Submit
Reset
Alternative search options
The HMMER web server: fast and sensitive homology searches. This site has been designed to provide near interactive searches for most queries, coupled with intuitive and interactive results visualisations.
Quickstart tutorial
Online documentation
Blog news
August, 2015
hmmer.org is updating
hmmer.org is moving off of Janelia tonight, into the
Download HMMER
v3.1b2
Recent papers
HMMER web server: 2015 update
R.D. Finn, J. Clements, W. Arndt, B.L. Miller, T.J. Wheeler,

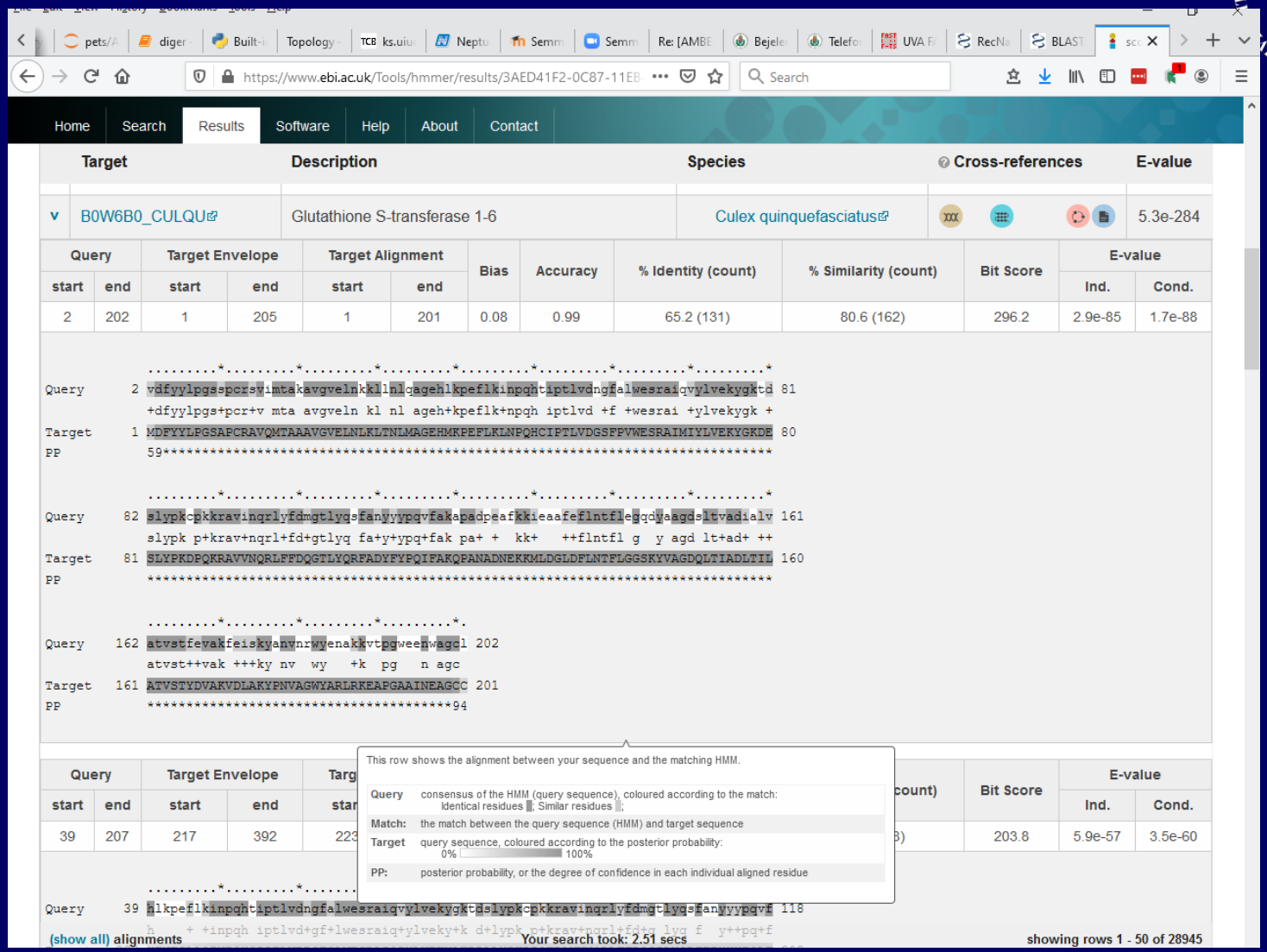

Edit View History Bookmarks Tools Help
Mozilla Firefox Start Page
UVA FASTA Server
FASTA/SSEARCH/GGSEARC
Help: FASTA (Protein Datab
BLAST: Basic Local Alignme
phmmer search | HMMER
https://www.ebi.ac.uk/Tools/hmmer/search/phmmer
Search
EMBL-EBI
Services
Research
Training
About us
EMBL-EBI
Hinxton
HMMER
Biosequence analysis using profile hidden Markov Models
Home
Search
Results
Software
Help
About
Contact
phmmer
hmmscan
hmmsearch
jackhmmer
protein sequence vs protein sequence database
Paste a Sequence | Upload a File | Accession Search
Paste in your sequence or use the example
Submit
Reset
Sequence Database
Frequently used databases:
Reference Proteomes
UniProtKB
SwissProt
PDB
Ensembl

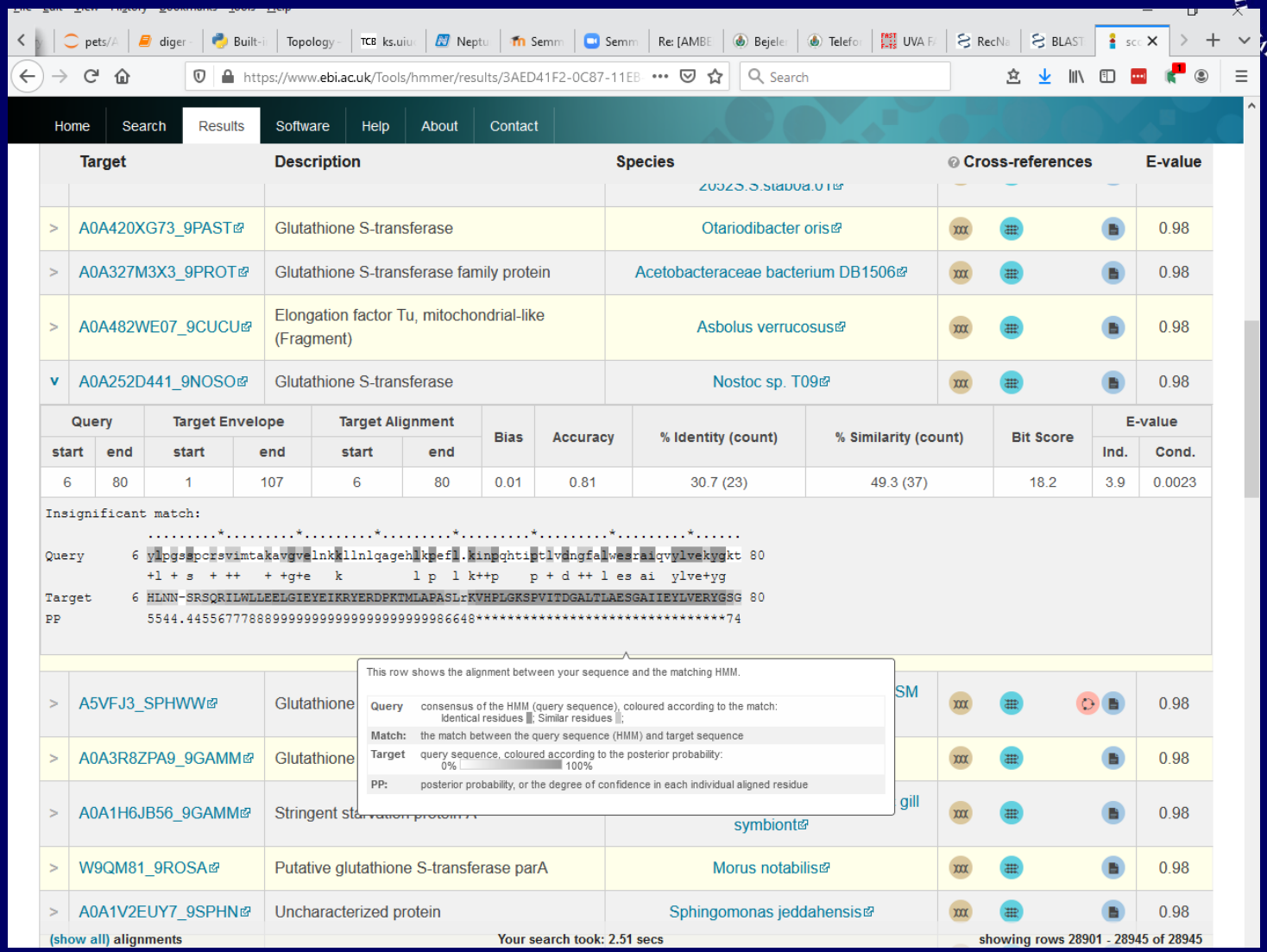

NCBI BLAST < Sequence Si...
UVA FASTA Downloads
BLAST: Basic Local Alignm...
HMMER
www.ebi.ac.uk/Tools/hmmer/search/phmmer
decathlon
Home
Search
Results
Software
Help
About
Sequence Database
Frequently used databases
Reference Proteomes
UniProtKB
SwissProt
PDB
Representative Sets (UniProt)
rp75
rp55
rp35
rp15
Other databases
QfO
Pfamseq
Restrict by Taxonomy
Cut-Offs
E-value
Bit score
Significance E-values: Sequence 0.01 Hit 0.03
Report E-values: Sequence 1 Hit 1
Customize Results
Select Visible Columns
Row Count
Secondary Accessions and Ids
Description
Species
Kingdom
Known Structure
Identical Seqs
Number of Hits
Number of Significant Hits
Bit Score
Hit Positions
Rows Per Page
50
100
250
1000
2500
All
Gap Penalties
open 0.02 extend 0.4
Substitution scoring matrix: BLOSUM62
Filters
Turn off bias composition filter

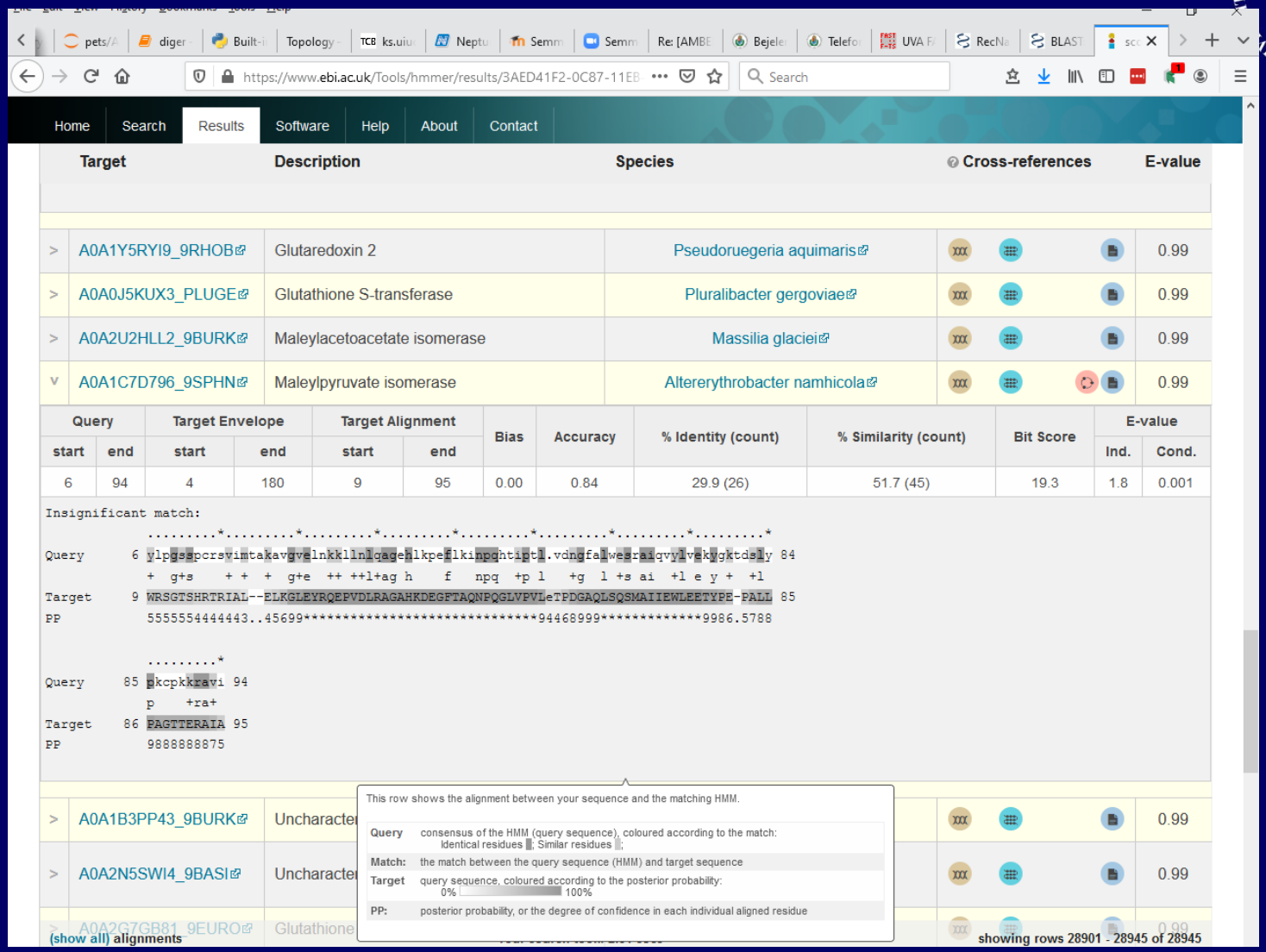

https://www.ebi.ac.uk/Tools/hmmer/results/3AED41F2-0C87-11EB
Search
EMBL-EBI
Services
Research
Training
About us
EMBL-EBI
Hinxton
HMMER
Biosequence analysis using profile hidden Markov Models
Home
Search
Results
Software
Help
About
Contact
PHMMER Results
Search Again
Score
Taxonomy
Domain
Download
Sequence Matches and Features
Pfam
GST_C
209
hit coverage
hit similarity
✓ disorder
✓ coiled-coil
✓ tm & signal peptide
Show hit details
Distribution of Significant Hits
more significant
Bacteria
Eukaryota
Archaea
Viruses
Unclassified Sequences
Other Sequences
« First
« Previous
Page 1
of 579
Next »
Last »
Significant Query Matches (21376) in uniprotrefprot (v.2019_09)
Customise


https://www.ebi.ac.uk/Tools/hmmer/results/3AED41F2-0C87-11EB
Search
Home
Search
Results
Software
Help
About
Contact
Pfam
GST_C
209
hit coverage
hit similarity
✓ disorder
✓ coiled-coil
✓ tm & signal peptide
Show hit details
Distribution of Significant Hits
more significant
Bacteria
Eukaryota
Archaea
Viruses
Unclassified Sequences
Other Sequences
« First
« Previous
Page 1 of 579
Next »
Last »
Significant Query Matches (21376) in uniprotrefprot (v.2019_09)
Customise
Target
Description
Species
Cross-references
E-value
A0A182JTU9_9DIPT
Uncharacterized protein
Anopheles christyi
0.0e+00
A0A0L0BYE0_LUCCU
Uncharacterized protein
Lucilia cuprina
0.0e+00
B0W6B0_CULQU
Glutathione S-transferase 1-6
Culex quinquefasciatus
5.3e-284
A0A182QC27_9DIPT
Uncharacterized protein
Anopheles farauti
2.2e-265
A0A182UJV8_9DIPT
Uncharacterized protein
Anopheles melas
3.5e-261
A0A1W4UME7_DROFC
uncharacterized protein LOC108087889
Drosophila ficusphila
9.0e-252
A0A182MSE4_9DIPT
Uncharacterized protein
Anopheles culicifacies
8.2e-245
(show all) alignments
Your search took: 2.51 secs
showing rows 1 - 50 of 28945


https://www.ebi.ac.uk/Tools/hmmer/results/3AED41F2-0C87-11EB
Search
Home
Search
Results
Software
Help
About
Contact
Bacteria
Eukaryota
Archaea
Viruses
Unclassified Sequences
Other Sequences
« First
« Previous
Page 1 of 579
Next »
Last »
Significant Query Matches (21376) in uniprotrefprot (v.2019_09)
Customise
Target
Description
Species
Cross-references
E-value
A0A182JTU9_9DIPT Uncharacterized protein Anopheles christyi 0.0e+00
A0A0L0BYE0_LUCCU Uncharacterized protein Lucilia cuprina 0.0e+00
B0W6B0_CULQU Glutathione S-transferase 1-6 Culex quinquefasciatus 5.3e-284
A0A182QC27_9DIPT Uncharacterized protein Anopheles farauti 2.2e-265
A0A182UJV8_9DIPT Uncharacterized protein Anopheles melas 3.5e-261
A0A1W4UME7_DROFC uncharacterized protein LOC108087889 Drosophila ficusphila 9.0e-252
A0A182MSE4_9DIPT Uncharacterized protein Anopheles culicifacies 8.2e-245
A0A2P8XIY1_BLAGE Uncharacterized protein (Fragment) Blattella germanica 2.6e-240
A0A1I8MLE7_MUSDO Uncharacterized protein Musca domestica 3.9e-225
B4K5W6_DROMO Uncharacterized protein Drosophila mojavensis 4.7e-220
A0A182YGN2_ANOST Uncharacterized protein Anopheles stephensi 3.3e-208
A0A1J1IJQ6_9DIPT CLUMA_CG013690, isoform A Clunio marinus 3.4e-200
A0A0P8ZYE2_DROAN Uncharacterized protein Drosophila ananassae 2.7e-199
A0A1W4UML8_DROFC uncharacterized protein LOC108087890 Drosophila ficusphila 2.6e-198
(show all) alignments
Your search took: 2.51 secs
showing rows 1 - 50 of 28945


https://www.ebi.ac.uk/Tools/hmmer/results/3AED41F2-0C87-11EB
Search
Home
Search
Results
Software
Help
About
Contact
Target
Description
Species
Cross-references
E-value
B0W6B0_CULQU
Glutathione S-transferase 1-6
Culex quinquefasciatus
5.3e-284
Query
Target Envelope
Target Alignment
Bias
Accuracy
% Identity (count)
% Similarity (count)
Bit Score
E-value
start
end
Ind.
Cond.
2 202 1 205 1 201 0.08 0.99 65.2 (131) 80.6 (162) 296.2 2.9e-85 1.7e-88
Query 2 vdfyylpgssprsvimtakavgvelnkkllnlqageklkpeflkinpqhtiptlvdngfalwesraiqvylvekygktd 81
+dfyylpgs+pcr+v mta avgveln kl nl ageh+kpeflk+npqh iptlvd +f +wesrai +ylvekygk +
Target 1 MDFYYLPGSAPCRAVQMTAAAVGVELNLKLTNLMAGEHMKPEFLKLNPQHCIPTLVDGSFPVWESRAIMIYLVEKYGKDE 80
PP 59**
Query 82 slypkcpkkravingrlyfdmgtlygsfanyyppqvfakapadpeafkkieaafeflntflegqdyaagdsltvadialv 161
slypk p+krav+nqrl+fd+gtlyq fa+y+ypq+fak pa+ + kk+ ++flntfl g y agd lt+ad+ ++
Target 81 SLYPKDPQKRAVVNQRLFFDQGTLYQRFADYFYPQIFAKQPANADNEKKMLDGLDFLNTFLGGSKYVAGDQLTIADLTIL 160
PP **
Query 162 atvstfevakfeiskyanvnrwyenakkvtpgweenwagcl 202
atvst++vak +++ky nv wy +k pg n agc
Target 161 ATVSTYDVAKVDLAKYPNVAGWYARLRKEAPGAAINEAGCC 201
PP ***************************************94
This row shows the alignment between your sequence and the matching HMM.
Query consensus of the HMM (query sequence), coloured according to the match:
Identical residues ; Similar residues ;
Match: the match between the query sequence (HMM) and target sequence
Target query sequence, coloured according to the posterior probability:
0% 100%
PP: posterior probability, or the degree of confidence in each individual aligned residue
39 207 217 392 203.8 5.9e-57 3.5e-60
Query 39 klkpeflkinpqhtiptlvdngfalwesraiqvylvekygktdslypkcpkkravingrlyfdmgtlygsfanyyppqvf 118
(show all) alignments
Your search took: 2.51 secs
showing rows 1 - 50 of 28945


https://www.ebi.ac.uk/Tools/hmmer/results/3AED41F2-0C87-11EB
Search
Home
Search
Results
Software
Help
About
Contact
Target
Description
Species
Cross-references
E-value
> A0A420XG73_9PAST Glutathione S-transferase Otariodibacter oris 0.98
> A0A327M3X3_9PROT Glutathione S-transferase family protein Acetobacteraceae bacterium DB1506 0.98
> A0A482WE07_9CUCU Elongation factor Tu, mitochondrial-like (Fragment) Asbolus verrucosus 0.98
v A0A252D441_9NOSO Glutathione S-transferase Nostoc sp. T09 0.98
Query
Target Envelope
Target Alignment
Bias
Accuracy
% Identity (count)
% Similarity (count)
Bit Score
E-value
start
end
Ind.
Cond.
6 80 1 107 6 80 0.01 0.81 30.7 (23) 49.3 (37) 18.2 3.9 0.0023
Insignificant match:
Query 6 ylpgssporsvimtakavgvelnkkllnlqagehlkpefl.kinpqhtiptlvdngfalwesraiqvylvekygkt 80
Target 6 HLNN-SRSQRILWLLEELGIEYEIKRYERDPKTMLAPASLrKVHPLGKSPVITDGALTLAESGAIIEYLVERYGSG 80
PP 5544.44556777888999999999999999999986648**********74
This row shows the alignment between your sequence and the matching HMM.
Query consensus of the HMM (query sequence), coloured according to the match: Identical residues ; Similar residues ;
Match: the match between the query sequence (HMM) and target sequence
Target query sequence, coloured according to the posterior probability: 0% 100%
PP: posterior probability, or the degree of confidence in each individual aligned residue
> A5VFJ3_SPHWW Glutathione
> A0A3R8ZPA9_9GAMM Glutathione
> A0A1H6JB56_9GAMM Stringent
symbiont
> W9QM81_9ROSA Putative glutathione S-transferase parA Morus notabilis 0.98
> A0A1V2EUY7_9SPHN Uncharacterized protein Sphingomonas jeddahensis 0.98
(show all) alignments
Your search took: 2.51 secs
showing rows 28901 - 28945 of 28945



https://www.ebi.ac.uk/Tools/hmmer/results/3AED41F2-0C87-11EB

Home | Search | Results | Software | Help | About | Contact

	Target	Description	Species	Cross-references	E-value
>	A0A1Y5RYI9_9RHOB	Glutaredoxin 2	Pseudoruegeria aquimaris		0.99
>	A0A0J5KUX3_PLUGE	Glutathione S-transferase	Pluralibacter gergoviae		0.99
>	A0A2U2HLL2_9BURK	Maleylacetoacetate isomerase	Massilia glaciei		0.99
v	A0A1C7D796_9SPHN	Maleylpyruvate isomerase	Altererythrobacter namhicola		0.99

Query start	Query end	Target Envelope start	Target Envelope end	Target Alignment start	Target Alignment end	Bias	Accuracy	% Identity (count)	% Similarity (count)	Bit Score	E-value Ind.	E-value Cond.
6	94	4	180	9	95	0.00	0.84	29.9 (26)	51.7 (45)	19.3	1.8	0.001

```
Insignificant match:
                .........*.........*.........*.........*.........*.........*.........*.........*
Query         6 ylpgsspcrsvimtakavgvelnkkllnlqageklkpeflkinpqhtiptl.vdngfalwesraiqvylvekygktdsly 84
                +  g+s    + +  +  g+e  ++ ++l+ag h    f   npq  +p l   +g  l +s ai  +l e y +  +l
Target        9 WRSGTSHRTRIAL--ELKGLEYRQEPVDLRAGAHKDEGFTAQNPQGLVPVLeTPDGAQLSQSMAIIEWLEETYPE-PALL 85
PP            5555554444443..45699*****************************94468999************9986.5788

                .........*
Query        85 gkcpkkravi 94
                p    +ra+
Target       86 PAGTTERAIA 95
PP            9888888875
```

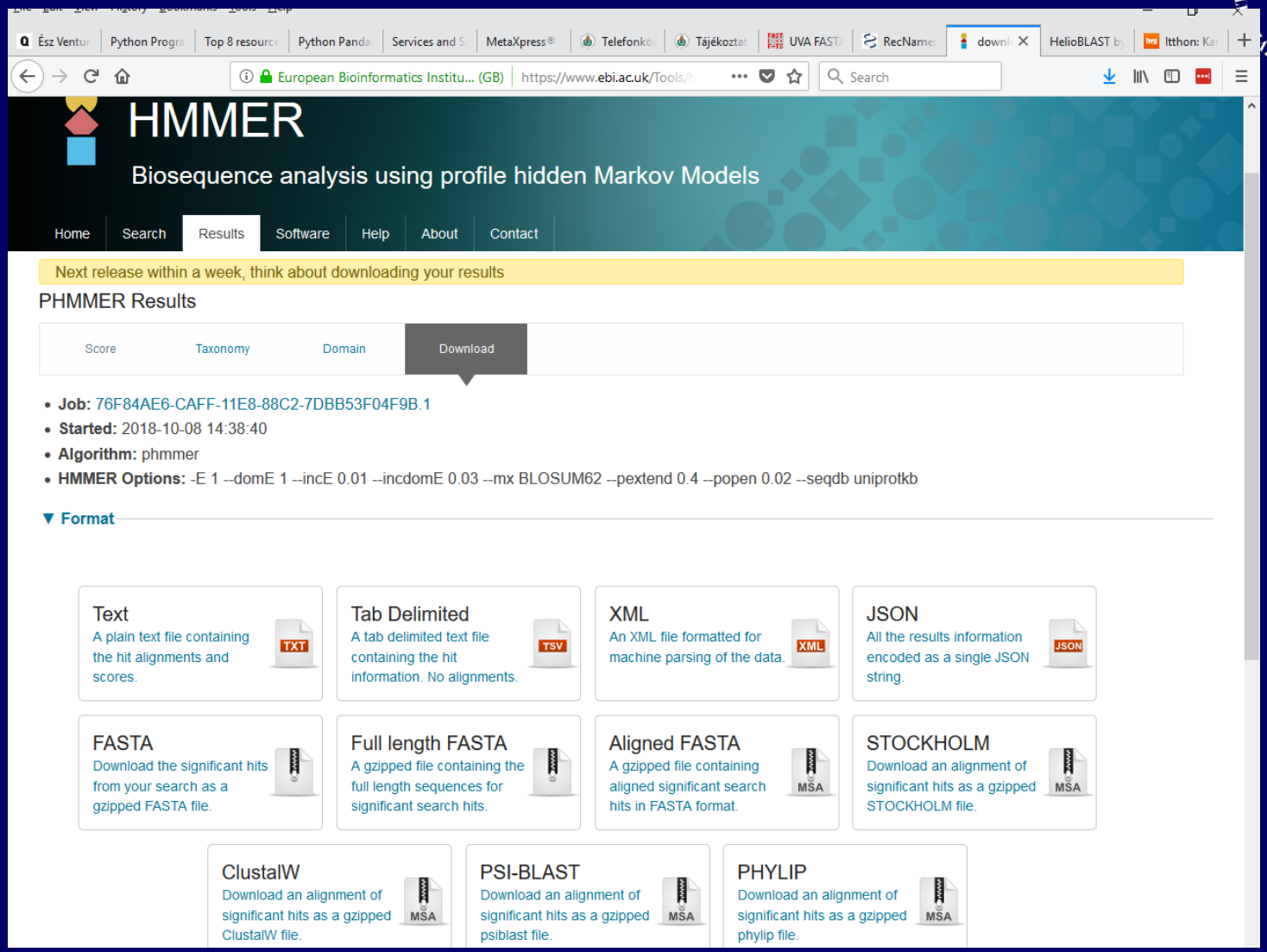
This row shows the alignment between your sequence and the matching HMM.

Query	consensus of the HMM (query sequence), coloured according to the match: Identical residues ; Similar residues ;
Match:	the match between the query sequence (HMM) and target sequence
Target	query sequence, coloured according to the posterior probability: 0% 100%
PP:	posterior probability, or the degree of confidence in each individual aligned residue

>	A0A1B3PP43_9BURK	Uncharacte			0.99
>	A0A2N5SWI4_9BASI	Uncharacte			0.99
>	A0A2G7GB81_9EURO	Glutathione			0.99

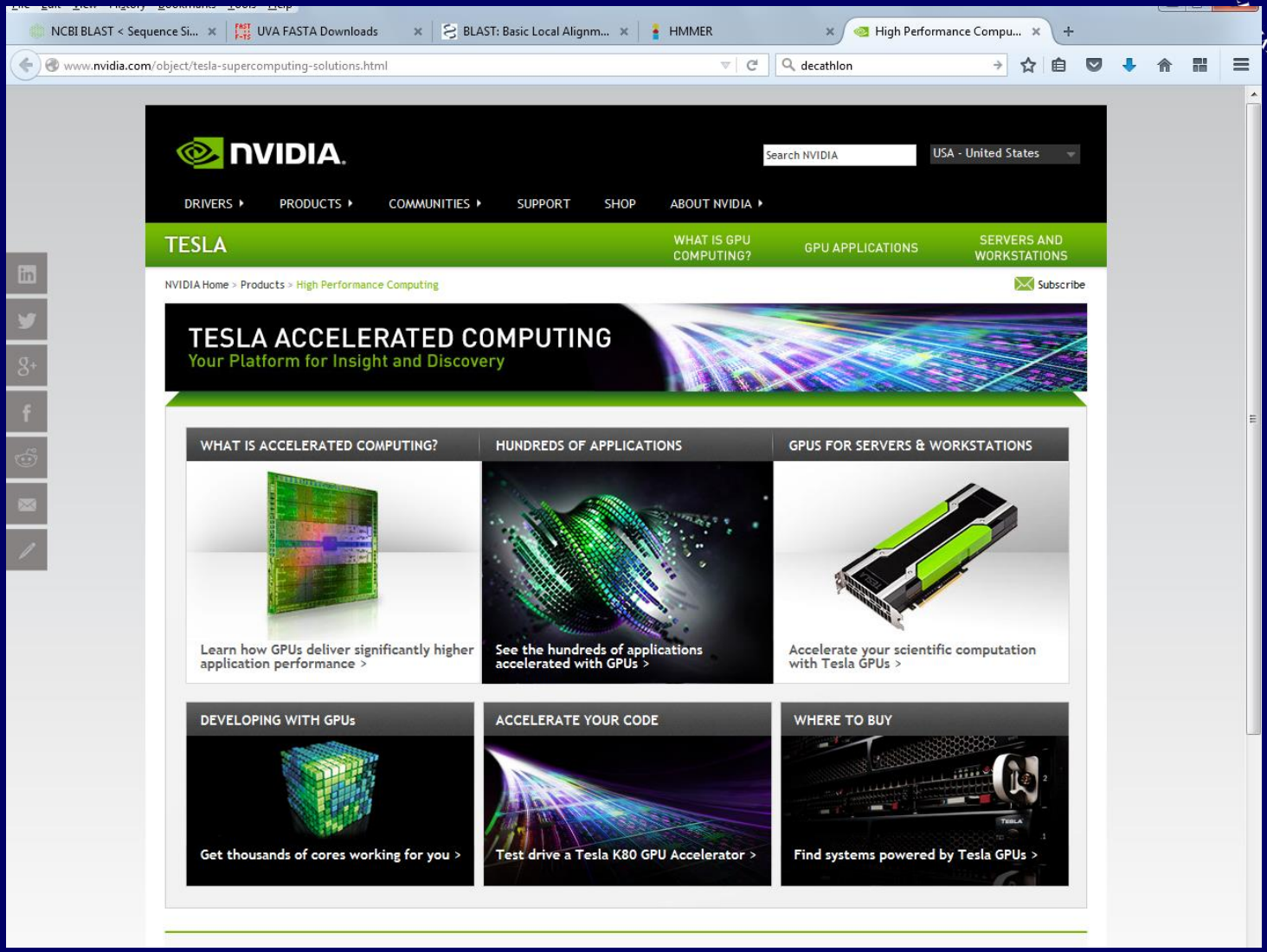
(show all) alignments

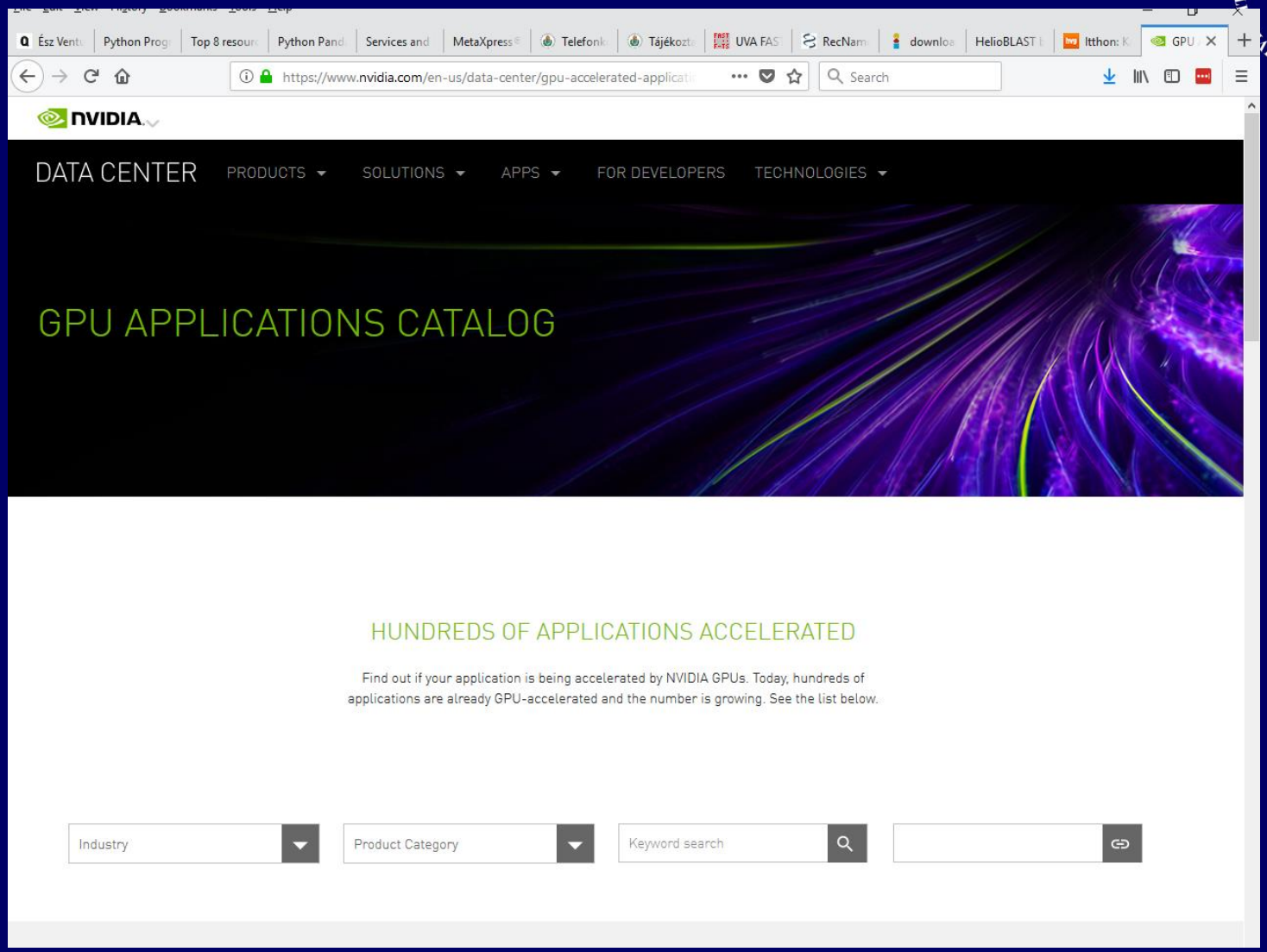
showing rows 28901 - 28945 of 28945

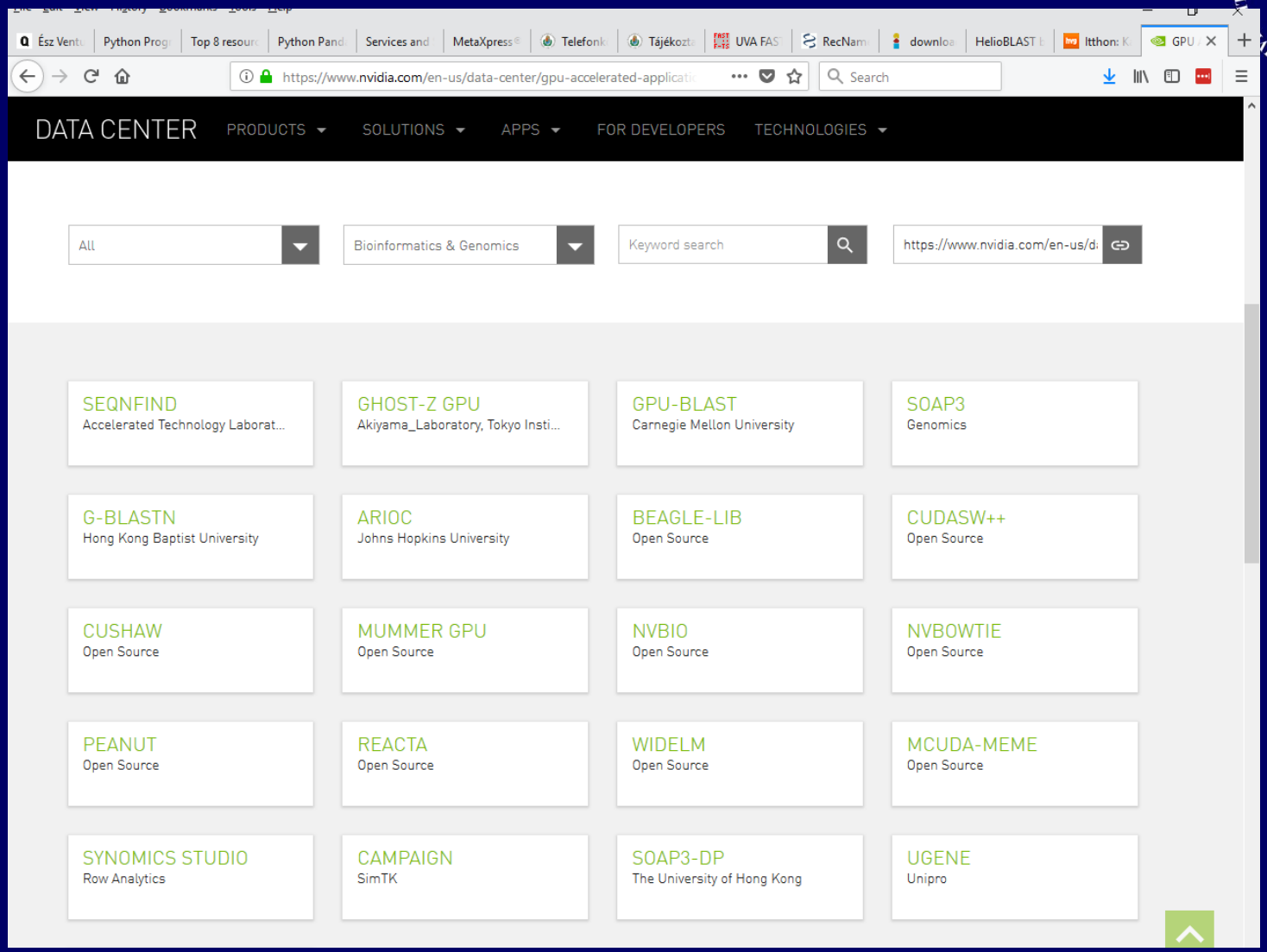

European Bioinformatics Institu... (GB)
https://www.ebi.ac.uk/Tools/
Search
HMMER
Biosequence analysis using profile hidden Markov Models
Home
Search
Results
Software
Help
About
Contact
Next release within a week, think about downloading your results
PHMMER Results
Score
Taxonomy
Domain
Download
Job: 76F84AE6-CAFF-11E8-88C2-7DBB53F04F9B.1
Started: 2018-10-08 14:38:40
Algorithm: phmmer
HMMER Options: -E 1 --domE 1 --incE 0.01 --incdomE 0.03 --mx BLOSUM62 --pextend 0.4 --popen 0.02 --seqdb uniprotkb
▼ Format
Text
A plain text file containing the hit alignments and scores.
TXT
Tab Delimited
A tab delimited text file containing the hit information. No alignments.
TSV
XML
An XML file formatted for machine parsing of the data.
XML
JSON
All the results information encoded as a single JSON string.
JSON
FASTA
Download the significant hits from your search as a gzipped FASTA file.
Full length FASTA
A gzipped file containing the full length sequences for significant search hits.
Aligned FASTA
A gzipped file containing aligned significant search hits in FASTA format.
MSA
STOCKHOLM
Download an alignment of significant hits as a gzipped STOCKHOLM file.
MSA
ClustalW
Download an alignment of significant hits as a gzipped ClustalW file.
MSA
PSI-BLAST
Download an alignment of significant hits as a gzipped psiblast file.
MSA
PHYLIP
Download an alignment of significant hits as a gzipped phylip file.
MSA

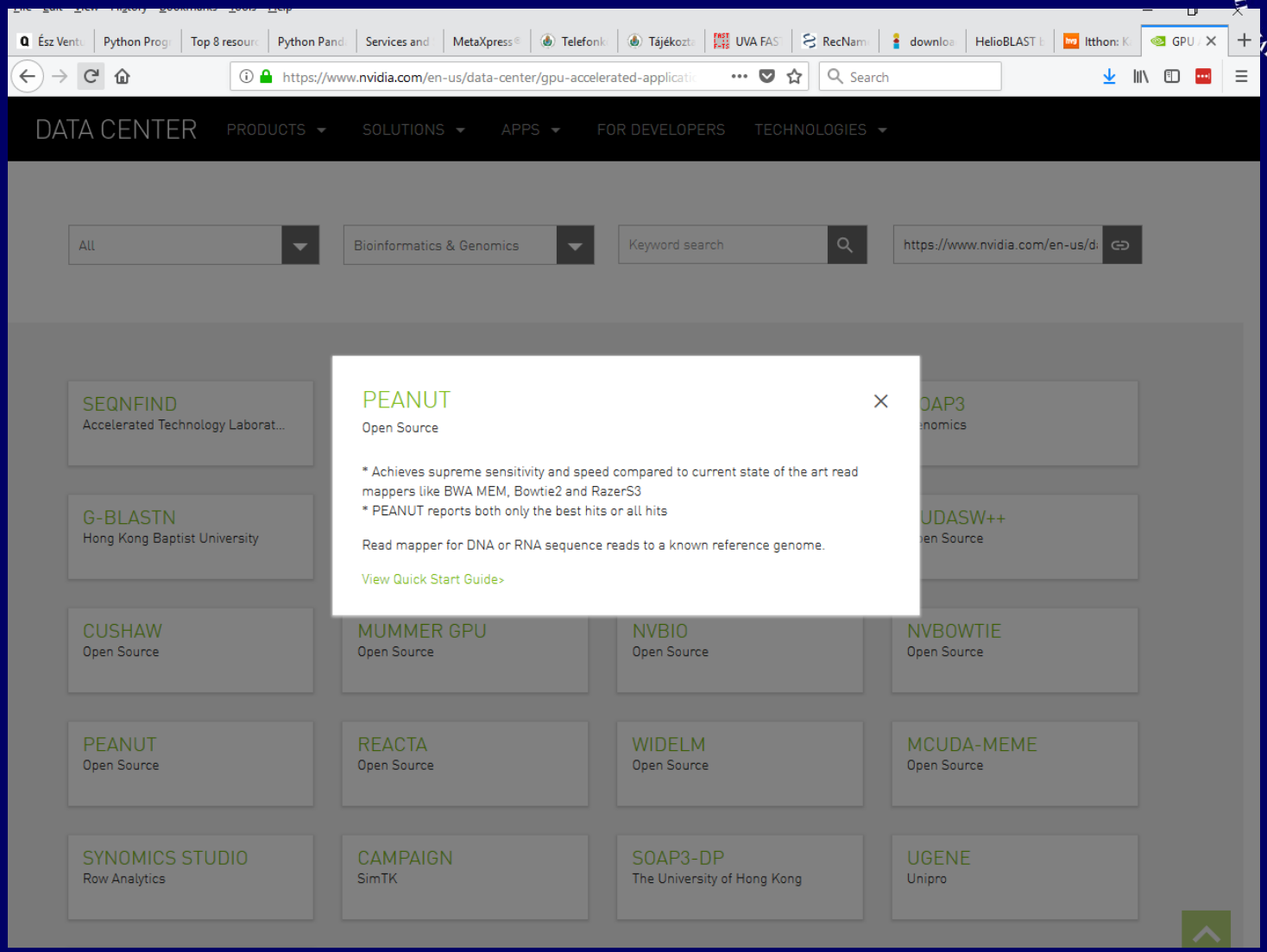
Paralell keresés

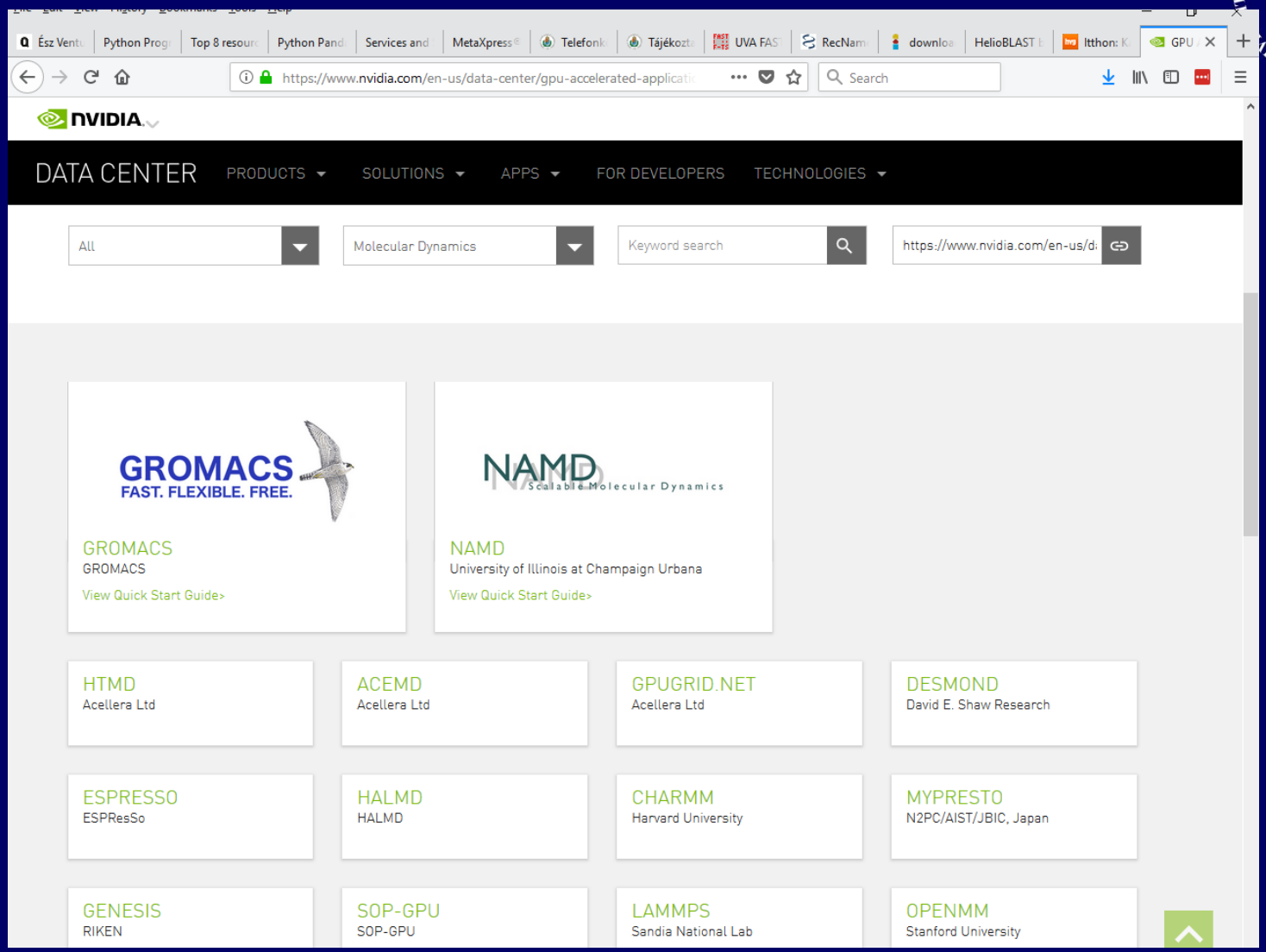
- A szekvenciák sorrendje egy adatbázisban esetleges
- Nem kell sorban haladni a keresés során
- Több processzoros, több magos architekturán párhuzamosan lehet futtatni a keresést – GPU computing
- A párhuzamos futtatáshoz nemcsak a kódot kell átírni, de az algoritmust is

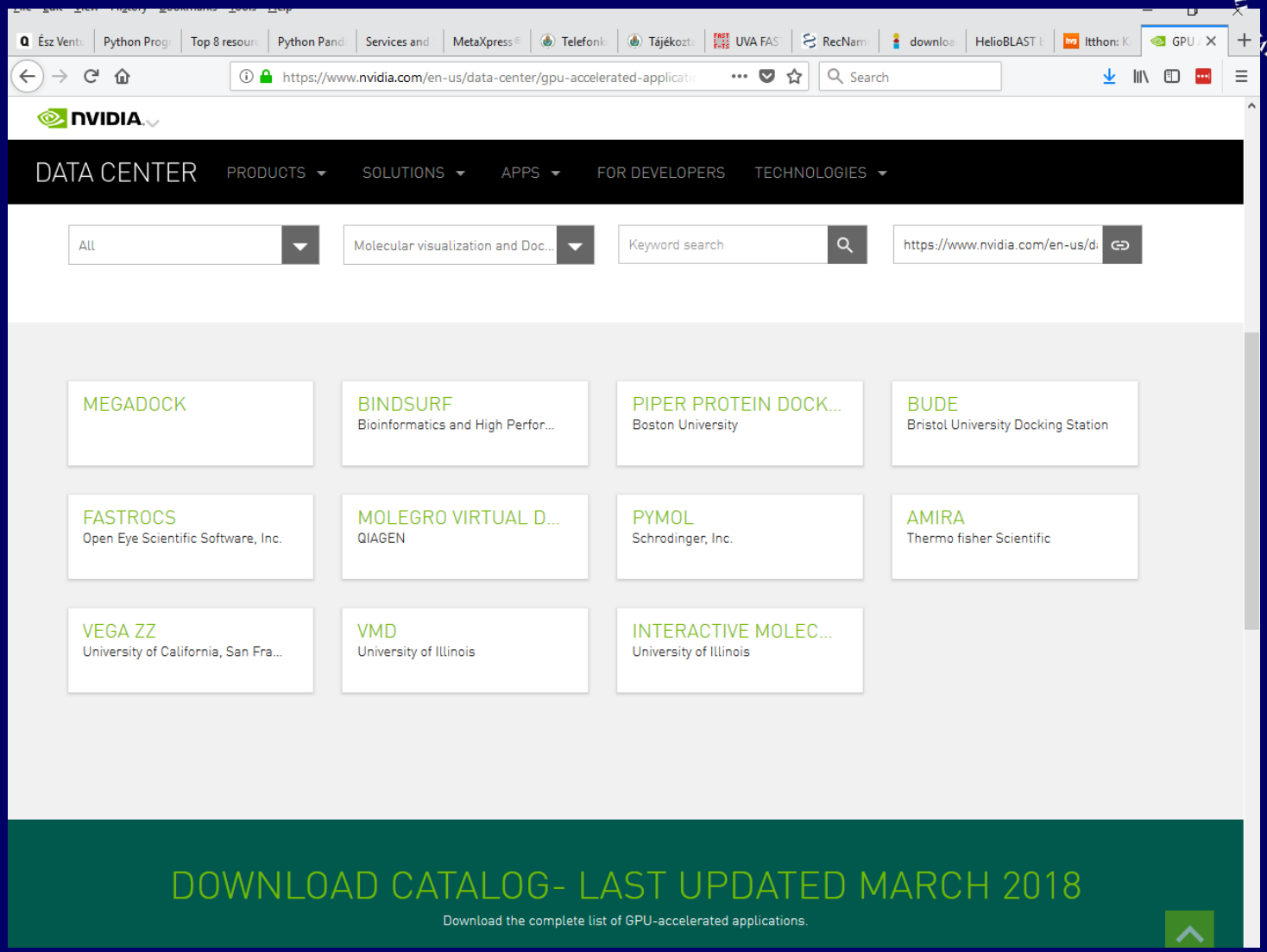

NCBI BLAST < Sequence Si...
UVA FASTA Downloads
BLAST: Basic Local Alignm...
HMMER
High Performance Compu...
www.nvidia.com/object/tesla-supercomputing-solutions.html
decathlon
NVIDIA
Search NVIDIA
USA - United States
DRIVERS
PRODUCTS
COMMUNITIES
SUPPORT
SHOP
ABOUT NVIDIA
TESLA
WHAT IS GPU COMPUTING?
GPU APPLICATIONS
SERVERS AND WORKSTATIONS
NVIDIA Home > Products > High Performance Computing
Subscribe
TESLA ACCELERATED COMPUTING
Your Platform for Insight and Discovery
WHAT IS ACCELERATED COMPUTING?
HUNDREDS OF APPLICATIONS
GPUS FOR SERVERS & WORKSTATIONS
Learn how GPUs deliver significantly higher application performance >
See the hundreds of applications accelerated with GPUs >
Accelerate your scientific computation with Tesla GPUs >
DEVELOPING WITH GPUs
ACCELERATE YOUR CODE
WHERE TO BUY
Get thousands of cores working for you >
Test drive a Tesla K80 GPU Accelerator >
Find systems powered by Tesla GPUs >


https://www.nvidia.com/en-us/data-center/gpu-accelerated-applicati
Search
NVIDIA
DATA CENTER
PRODUCTS
SOLUTIONS
APPS
FOR DEVELOPERS
TECHNOLOGIES
GPU APPLICATIONS CATALOG
HUNDREDS OF APPLICATIONS ACCELERATED
Find out if your application is being accelerated by NVIDIA GPUs. Today, hundreds of applications are already GPU-accelerated and the number is growing. See the list below.
Industry
Product Category
Keyword search


DATA CENTER
PRODUCTS
SOLUTIONS
APPS
FOR DEVELOPERS
TECHNOLOGIES
All
Bioinformatics & Genomics
Keyword search
SEQNFIND
Accelerated Technology Laborat...
GHOST-Z GPU
Akiyama_Laboratory, Tokyo Insti...
GPU-BLAST
Carnegie Mellon University
SOAP3
Genomics
G-BLASTN
Hong Kong Baptist University
ARIOC
Johns Hopkins University
BEAGLE-LIB
Open Source
CUDASW++
Open Source
CUSHAW
Open Source
MUMMER GPU
Open Source
NVBIO
Open Source
NVBOWTIE
Open Source
PEANUT
Open Source
REACTA
Open Source
WIDELM
Open Source
MCUDA-MEME
Open Source
SYNOMICS STUDIO
Row Analytics
CAMPAIGN
SimTK
SOAP3-DP
The University of Hong Kong
UGENE
Unipro


DATA CENTER
PRODUCTS
SOLUTIONS
APPS
FOR DEVELOPERS
TECHNOLOGIES
All
Bioinformatics & Genomics
Keyword search
SEQNFIND
Accelerated Technology Laborat...
G-BLASTN
Hong Kong Baptist University
CUSHAW
Open Source
MUMMER GPU
Open Source
NVBIO
Open Source
NVBOWTIE
Open Source
PEANUT
Open Source
REACTA
Open Source
WIDELM
Open Source
MCUDA-MEME
Open Source
SYNOMICS STUDIO
Row Analytics
CAMPAIGN
SimTK
SOAP3-DP
The University of Hong Kong
UGENE
Unipro
PEANUT
Open Source
* Achieves supreme sensitivity and speed compared to current state of the art read mappers like BWA MEM, Bowtie2 and RazerS3
* PEANUT reports both only the best hits or all hits
Read mapper for DNA or RNA sequence reads to a known reference genome.
View Quick Start Guide>


https://www.nvidia.com/en-us/data-center/gpu-accelerated-applicatic
Search
NVIDIA
DATA CENTER
PRODUCTS
SOLUTIONS
APPS
FOR DEVELOPERS
TECHNOLOGIES
All
Molecular Dynamics
Keyword search
https://www.nvidia.com/en-us/d
GROMACS
FAST. FLEXIBLE. FREE.
GROMACS
GROMACS
View Quick Start Guide>
NAMD
Scalable Molecular Dynamics
NAMD
University of Illinois at Champaign Urbana
View Quick Start Guide>
HTMD
Acellera Ltd
ACEMD
Acellera Ltd
GPUGRID.NET
Acellera Ltd
DESMOND
David E. Shaw Research
ESPRESSO
ESPResSo
HALMD
HALMD
CHARMM
Harvard University
MYPRESTO
N2PC/AIST/JBIC, Japan
GENESIS
RIKEN
SOP-GPU
SOP-GPU
LAMMPS
Sandia National Lab
OPENMM
Stanford University


https://www.nvidia.com/en-us/data-center/gpu-accelerated-applicati
Search
NVIDIA
DATA CENTER
PRODUCTS
SOLUTIONS
APPS
FOR DEVELOPERS
TECHNOLOGIES
All
Molecular visualization and Doc...
Keyword search
https://www.nvidia.com/en-us/d
MEGADOCK
BINDSURF
Bioinformatics and High Perfor...
PIPER PROTEIN DOCK...
Boston University
BUDE
Bristol University Docking Station
FASTROCS
Open Eye Scientific Software, Inc.
MOLEGRO VIRTUAL D...
QIAGEN
PYMOL
Schrodinger, Inc.
AMIRA
Thermo fisher Scientific
VEGA ZZ
University of California, San Fra...
VMD
University of Illinois
INTERACTIVE MOLEC...
University of Illinois
DOWNLOAD CATALOG- LAST UPDATED MARCH 2018
Download the complete list of GPU-accelerated applications.

Adatbányászat

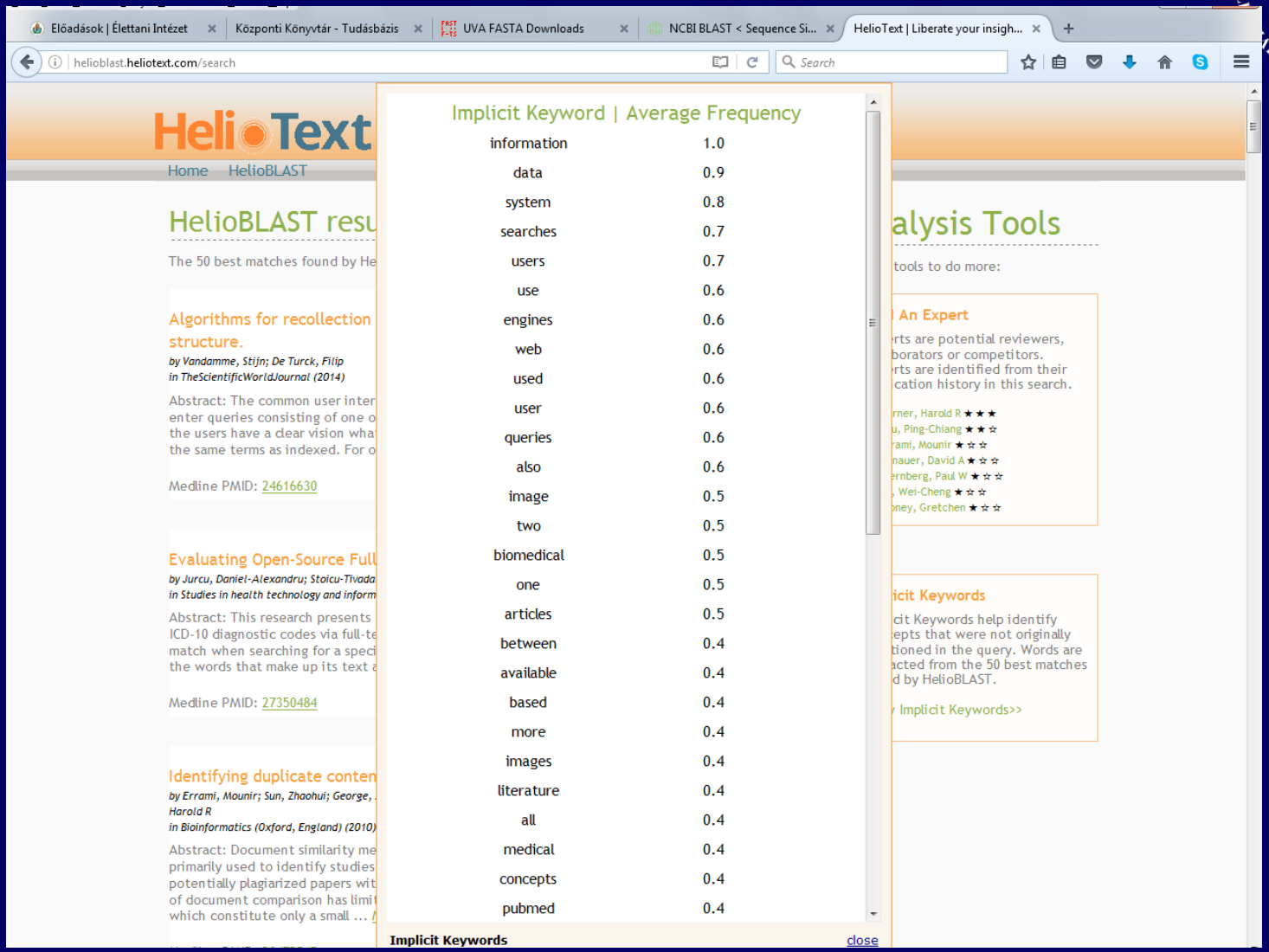
- Keresés szöveges adatbázisokban
- MEDLINE: tudományos szövegek kivonatai
- Ez is elsődleges adatbázis
- Igen nagy és ingyenesen elérhető
- Egy cikk mennyire hasonlít egy másikra?
- „számítógépes nyelvészet” – adatbányászat

eTBLAST

- Első fázis: súlyozott kulcsszó keresés
- Ez gyors, de nem túl érzékeny
- Második fázis: „mondat illesztő” lépés
- Ez az érzékenyebb
- Tartalmuk szerint hasonló cikkeket talál
- Javaslatot tesz a megfelelő újságra
- Hasonló érdeklődésű kutatókat azonosít
- http://helioblast.heliotext.com/


Előadások | Élettani Intézet
Központi Könyvtár - Tudásbázis
UVA FASTA Downloads
NCBI BLAST < Sequence Si...
HelioBLAST by HelioText
helioblast.heliotext.com
Search
HelioBLAST
HelioText
Home
HelioBLAST
Ask HelioBLAST
Search in: Medline Would you like to add your own database? Get in touch with us.
HelioBLAST is a free service provided by HelioText. The HelioBLAST text similarity engine finds text records that are similar to the submitted query. HelioBLAST uses our super-fast search engine, and this service is provided to demonstrate what can be done using text similarity searching. HelioBLAST can be customized to search a particular database or multiple ones; and proprietary databases can be created for individual clients.
Submit to HelioBLAST
MEDLINE index last updated Friday, October 07, 2016.
NLM data are produced by a U.S. Government agency and include works of the United States Government that are not protected by U.S. copyright law but may be protected by non-US copyright law, as well as abstracts originating from publications that may be protected by U.S. copyright law.
NLM assumes no responsibility or liability associated with use of copyrighted material, including transmitting, reproducing, redistributing, or making commercial use of the data. NLM does not provide legal advice regarding copyright, fair use, or other aspects of intellectual property rights. Persons contemplating any type of transmission or reproduction of copyrighted material such as abstracts are advised to consult legal counsel.
Welcome to HelioBLAST
HelioBLAST is a free service provided by HelioText. The HelioBLAST text similarity engine finds text records that are similar to the submitted query. HelioBLAST uses our super-fast search engine, and this service is provided to demonstrate what can be done using text similarity searching. HelioBLAST can be customized to search a particular database or multiple ones; and proprietary databases can be created for individual clients.
Here, your query is searched against the citations (abstract and titles) in Medline/PubMed. Submissions are limited to 1,000 words, so we recommend your query should be an abstract or paragraph.
Copyright HelioText 2015. email:kmenier@heliotext.com. Phone: +1-214-924-6334.


Előadások | Élettani Intézet
Központi Könyvtár - Tudásbázis
UVA FASTA Downloads
NCBI BLAST < Sequence Si...
HelioText | Liberate your insigh...
helioblast.heliotext.com/search
Search
HelioBLAST
HelioText
Home
HelioBLAST
HelioBLAST results similar to your query
The 50 best matches found by HelioBLAST in 2.829 sec:
Algorithms for recollection of search terms based on the Wikipedia category structure.
Score:0.067
by Vandamme, Stijn; De Turck, Filip
in TheScientificWorldJournal (2014)
Abstract: The common user interface for a search engine consists of a text field where the user can enter queries consisting of one or more keywords. Keyword query based search engines work well when the users have a clear vision what they are looking for and are capable of articulating their query using the same terms as indexed. For our multimedia ... More>>
Medline PMID: 24616630
Evaluating Open-Source Full-Text Search Engines for Matching ICD-10 Codes.
Score:0.066
by Jurcu, Daniel-Alexandru; Stoicu-Tivadar, Vasile
in Studies in health technology and informatics (2016)
Abstract: This research presents the results of evaluating multiple free, open-source engines on matching ICD-10 diagnostic codes via full-text searches. The study investigates what it takes to get an accurate match when searching for a specific diagnostic code. For each code the evaluation starts by extracting the words that make up its text and continues with building full-text search queries from ... More>>
Medline PMID: 27350484
Identifying duplicate content using statistically improbable phrases.
Score:0.063
by Errami, Mounir; Sun, Zhaohui; George, Angela C; Long, Tara C; Skinner, Michael A; Wren, Jonathan D; Garner, Harold R
in Bioinformatics (Oxford, England) (2010)
Abstract: Document similarity metrics such as PubMed's 'Find related articles' feature, which have been primarily used to identify studies with similar topics, can now also be used to detect duplicated or potentially plagiarized papers within literature reference databases. However, the CPU-intensive nature of document comparison has limited MEDLINE text similarity studies to the comparison of abstracts, which constitute only a small ... More>>
Analysis Tools
A few tools to do more:
Find An Expert
Experts are potential reviewers, collaborators or competitors. Experts are identified from their publication history in this search.
1. Garner, Harold R ★★★
2. Lyu, Ping-Chiang ★★☆
3. Errami, Mounir ★☆☆
4. Hanauer, David A ★☆☆
5. Sternberg, Paul W ★☆☆
6. Lo, Wei-Cheng ★☆☆
7. Gibney, Gretchen ★☆☆
Implicit Keywords
Implicit Keywords help identify concepts that were not originally mentioned in the query. Words are extracted from the 50 best matches found by HelioBLAST.
View Implicit Keywords>>


Előadások | Élettani Intézet
Központi Könyvtár - Tudásbázis
UVA FASTA Downloads
NCBI BLAST < Sequence Si...
HelioText | Liberate your insigh...
helioblast.heliotext.com/search
Search
HelioText
Home
HelioBLAST
HelioBLAST resu
The 50 best matches found by He
Algorithms for recollection
structure.
by Vandamme, Stijn; De Turck, Filip
in TheScientificWorldJournal (2014)
Abstract: The common user inter
enter queries consisting of one o
the users have a clear vision wha
the same terms as indexed. For o
Medline PMID: 24616630
Evaluating Open-Source Full
by Jurcu, Daniel-Alexandru; Stoicu-Tivada
in Studies in health technology and inform
Abstract: This research presents
ICD-10 diagnostic codes via full-te
match when searching for a speci
the words that make up its text a
Medline PMID: 27350484
Identifying duplicate conten
by Errami, Mounir; Sun, Zhaohui; George,
Harold R
in Bioinformatics (Oxford, England) (2010)
Abstract: Document similarity me
primarily used to identify studies
potentially plagiarized papers wit
of document comparison has limi
which constitute only a small ...
Implicit Keyword | Average Frequency
information 1.0
data 0.9
system 0.8
searches 0.7
users 0.7
use 0.6
engines 0.6
web 0.6
used 0.6
user 0.6
queries 0.6
also 0.6
image 0.5
two 0.5
biomedical 0.5
one 0.5
articles 0.5
between 0.4
available 0.4
based 0.4
more 0.4
images 0.4
literature 0.4
all 0.4
medical 0.4
concepts 0.4
pubmed 0.4
Implicit Keywords
close
alysis Tools
tools to do more:
An Expert
rts are potential reviewers,
borators or competitors.
rts are identified from their
cation history in this search.
rner, Harold R ★★★
u, Ping-Chiang ★★☆
rami, Mounir ★☆☆
nauer, David A ★☆☆
ernberg, Paul W ★☆☆
, Wei-Cheng ★☆☆
oney, Gretchen ★☆☆
icit Keywords
cit Keywords help identify
epts that were not originally
tioned in the query. Words are
acted from the 50 best matches
d by HelioBLAST.
Implicit Keywords>>

Mit tanultunk ma?

- Az adatbázis keresés lényegében nagyléptékű szekvenciaillesztés.
- Nagyon kiforrott technika.
- Gyakran a bioinformatikai vizsgálódás kiindulópontja.

Feladat 5.

- Válassz ki egy érdekes cikket és a kivonatával keress hasonlókat az et-blast rendszerben. Mennyire hatékony a szolgáltatás?
- Esetleg fogalmazd meg egy néhány mondatos kivonatban a téged érdeklő problémát és azzal keress.