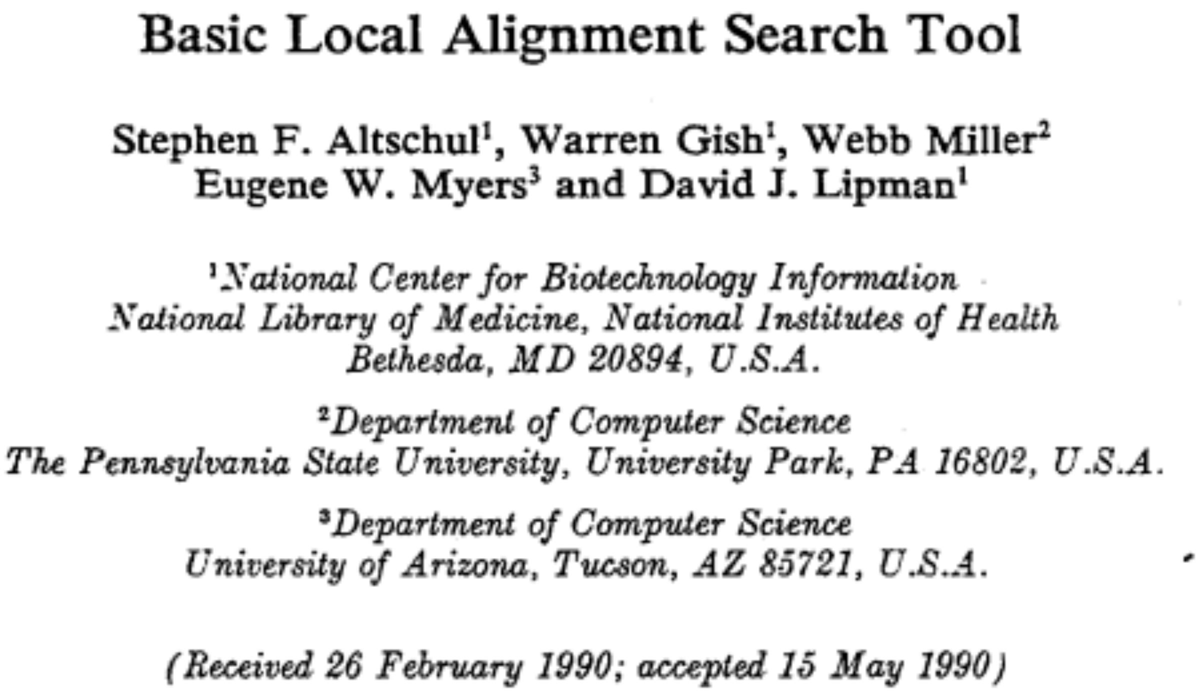
DATABASE SEARCHING USING BLAST

Hadrien Gourlé – Uganda 2014

BLAST

- Published in 1990



Basic Local Alignment Search Tool

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2]
Eugene W. Myers[3] and David J. Lipman[1]

[1]*National Center for Biotechnology Information*
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894, U.S.A.

[2]*Department of Computer Science*
The Pennsylvania State University, University Park, PA 16802, U.S.A.

[3]*Department of Computer Science*
University of Arizona, Tucson, AZ 85721, U.S.A.

(Received 26 February 1990; accepted 15 May 1990)

BLAST

- One of the most used Bioinformatics tool
- In the top 20 of most cited articles of all time

BLAST

- Find region of similarity between sequences
- Compare nucleotide or protein sequences

```
Query: 53   ttctggtccat 63
            |||||||*||||
Sbjct: 8848 ttctggaccat 8838
```

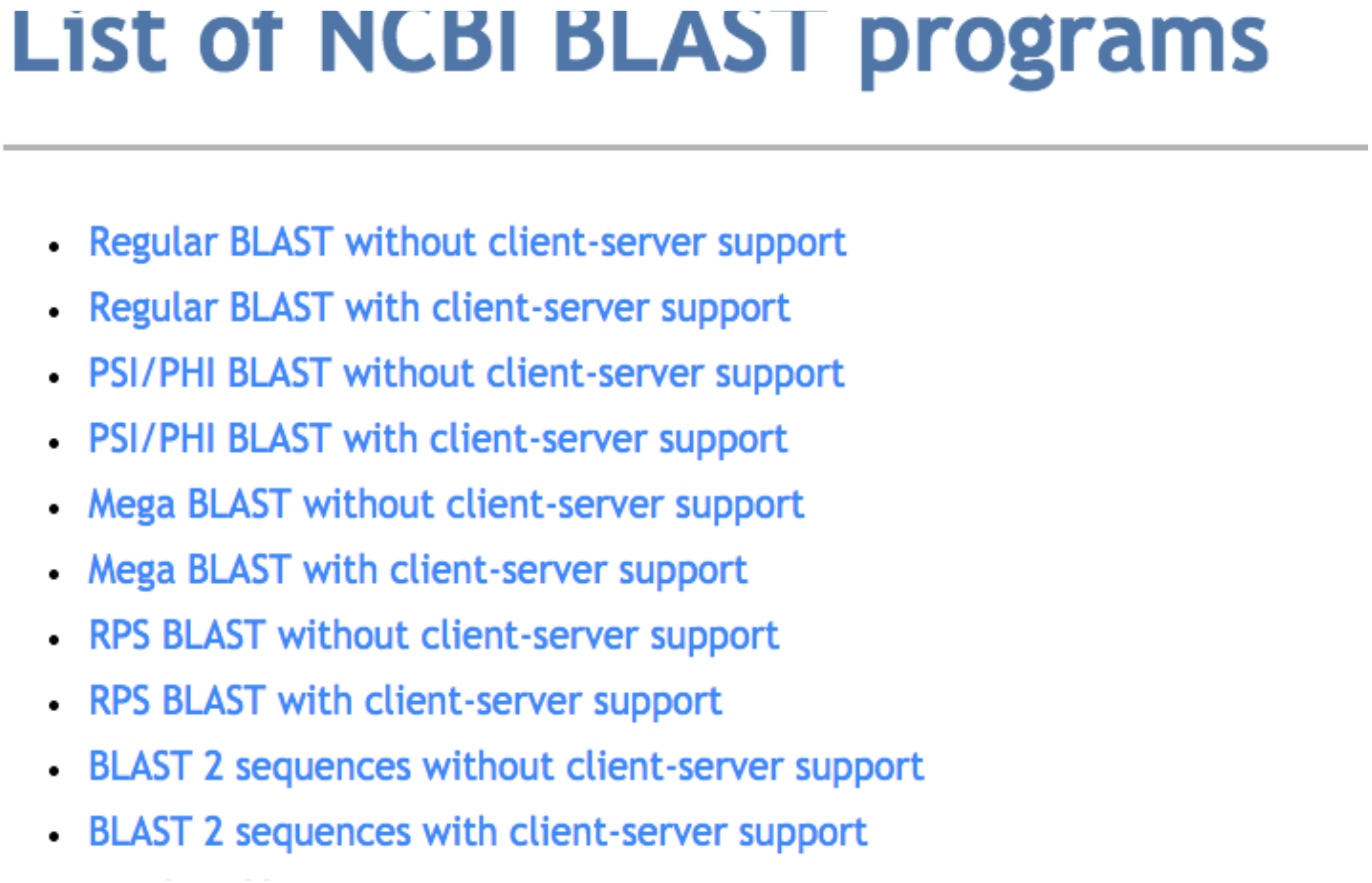
BLAST

- An improved version of the Smith-Waterman algorithm
- Weighted matrix

BLAST

- Blastn: nucleotide database / nucleotide query
- Blastp: protein database / protein query
- Blastx: protein database / translated nucleotide query
- tblastn: translated nucleotide database using a protein query
- tblastx: Search translated nucleotide database using a translated nucleotide query

BLAST

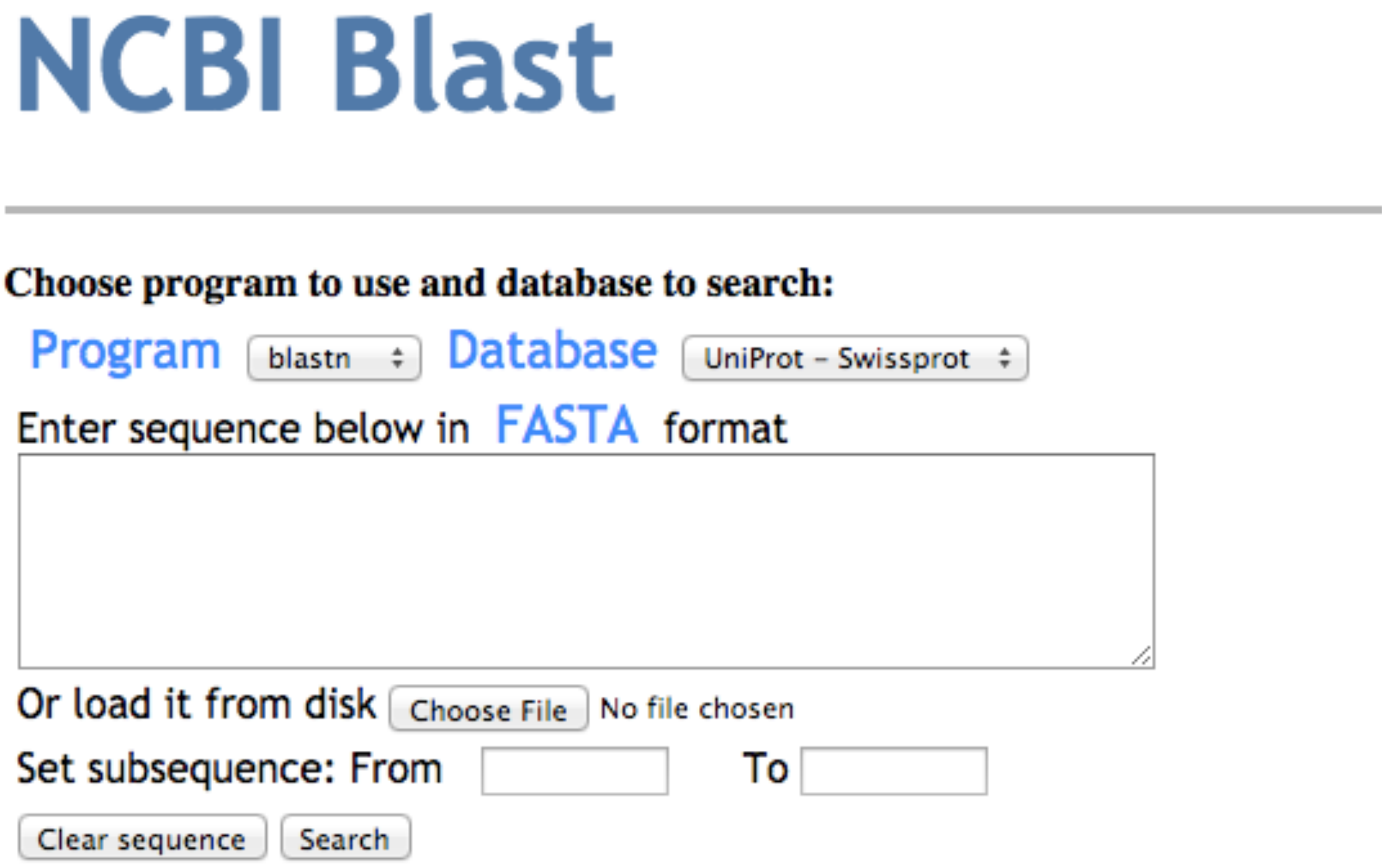


List of NCBI BLAST programs

- Regular BLAST without client-server support
- Regular BLAST with client-server support
- PSI/PHI BLAST without client-server support
- PSI/PHI BLAST with client-server support
- Mega BLAST without client-server support
- Mega BLAST with client-server support
- RPS BLAST without client-server support
- RPS BLAST with client-server support
- BLAST 2 sequences without client-server support
- BLAST 2 sequences with client-server support

BLAST

- User interface



NCBI Blast

Choose program to use and database to search:

Program: blastn Database: UniProt – Swissprot

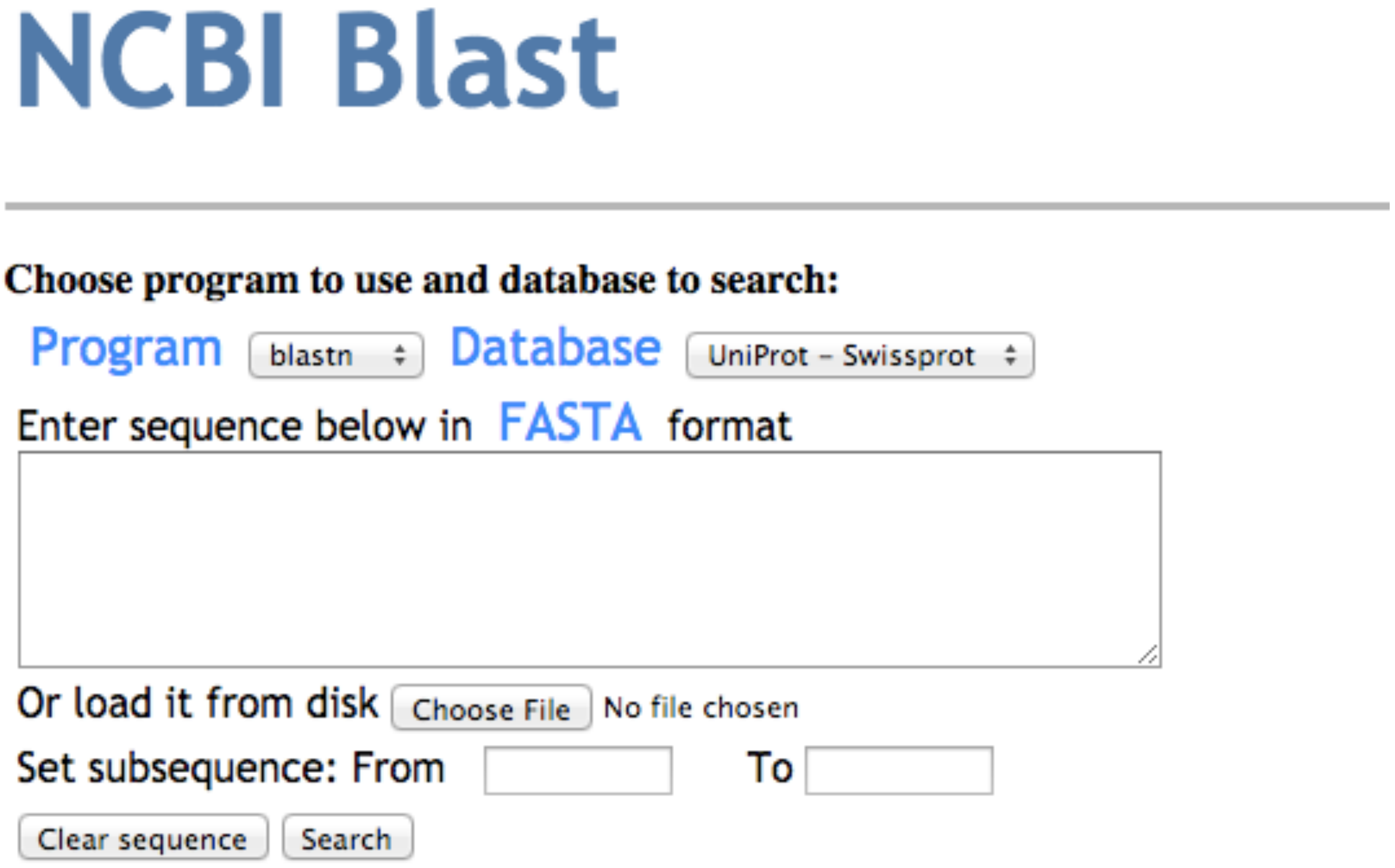
Enter sequence below in FASTA format

Or load it from disk: No file chosen

Set subsequence: From To

BLAST

- User interface



NCBI Blast

Choose program to use and database to search:

Program: blastn

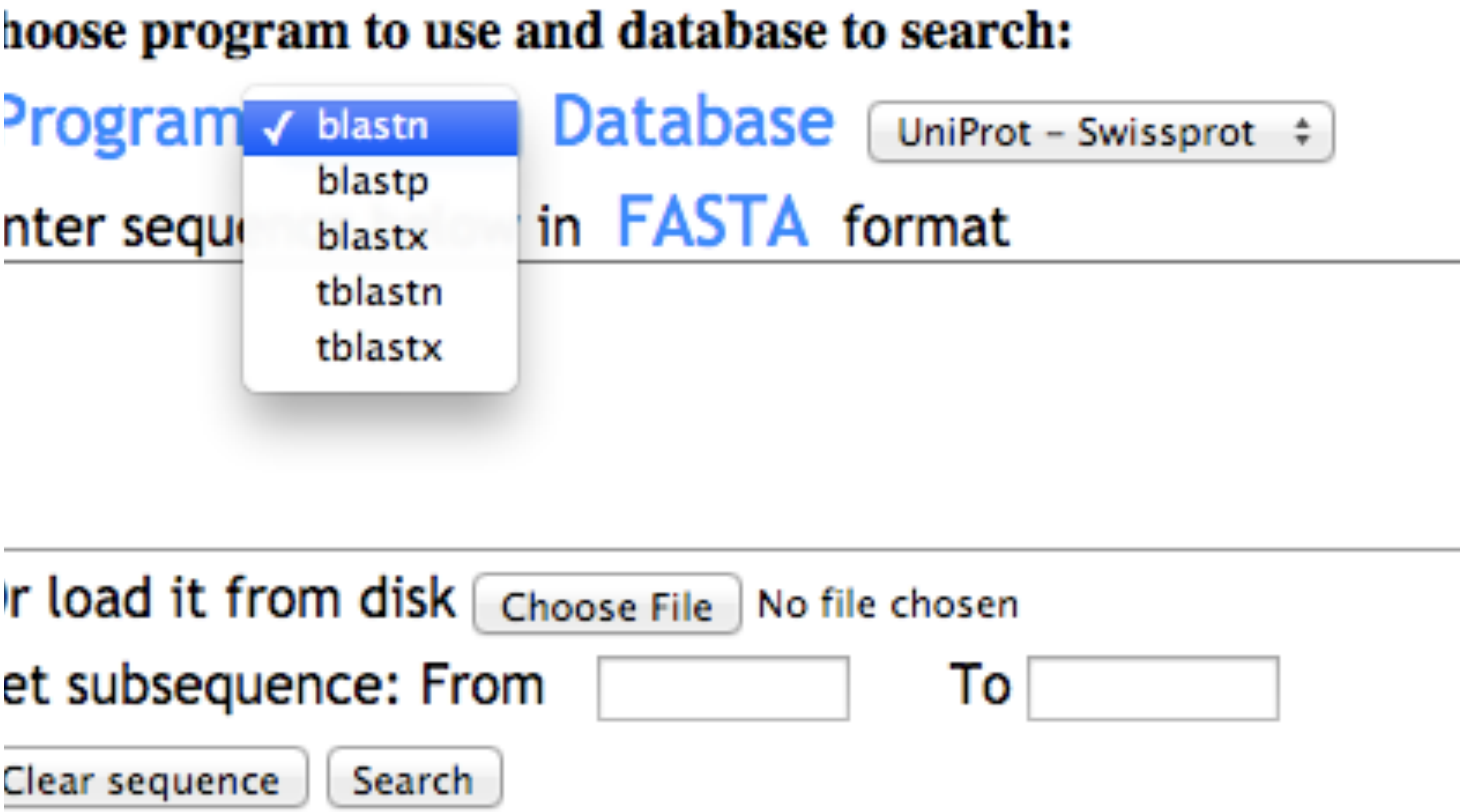
Database: UniProt - Swissprot

Enter sequence below in FASTA format

Or load it from disk

Set subsequence: From To

BLAST



hoose program to use and database to search:

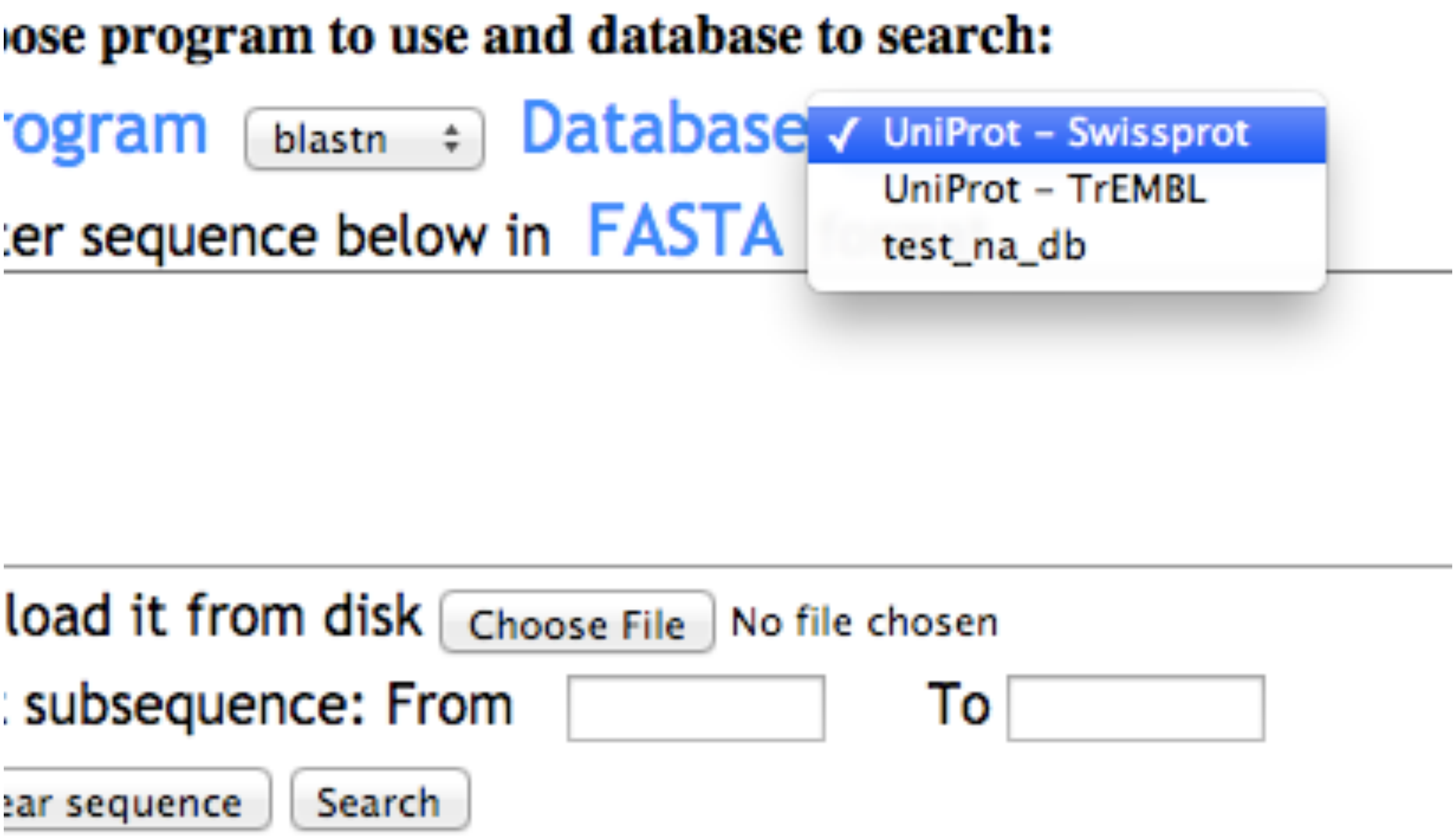
Program ✓ blastn Database UniProt – Swissprot

nter sequ in FASTA format

r load it from disk Choose File No file chosen

et subsequence: From To

BLAST



ose program to use and database to search:

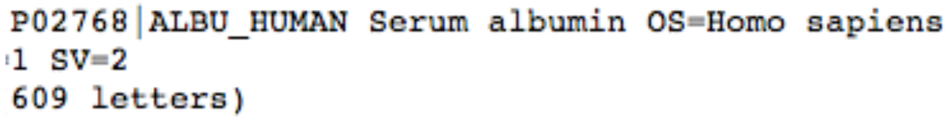
ogram blastn Database ✓ UniProt – Swissprot

er sequence below in FASTA

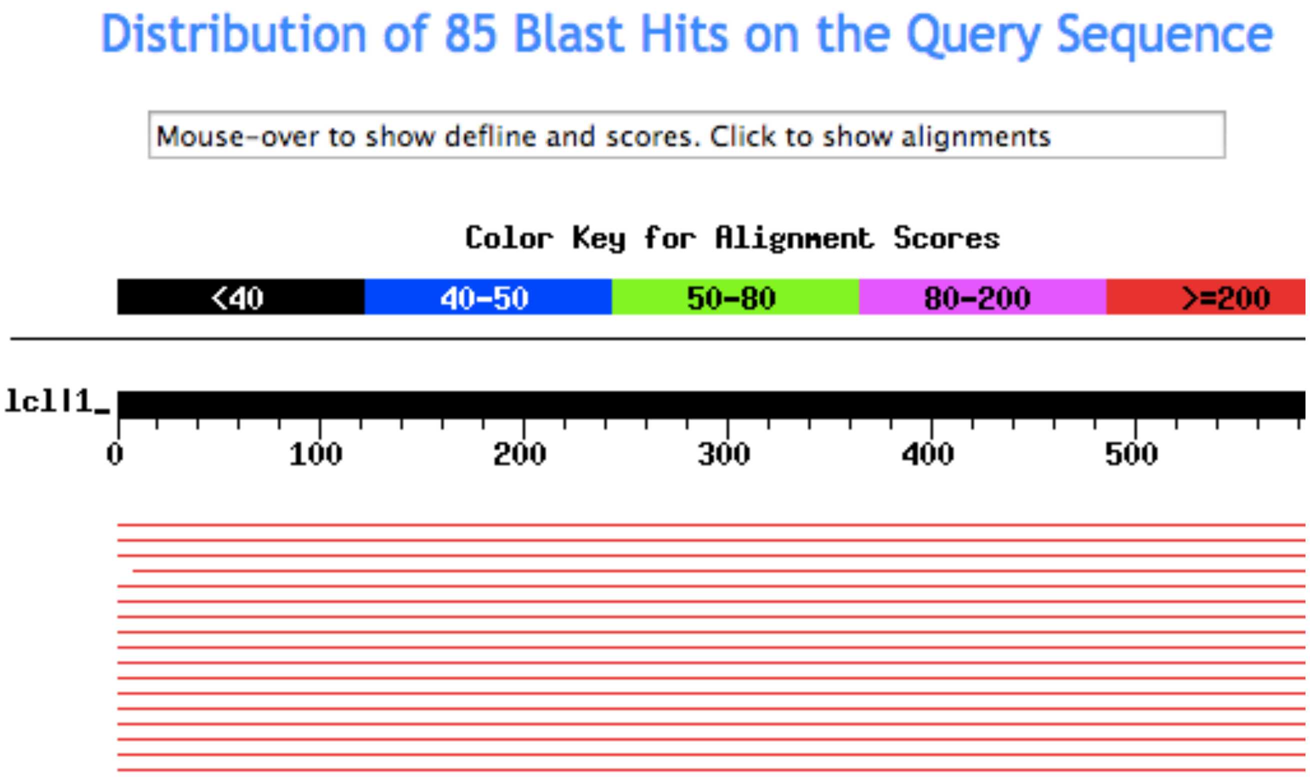
load it from disk Choose File No file chosen

subsequence: From To

BLAST



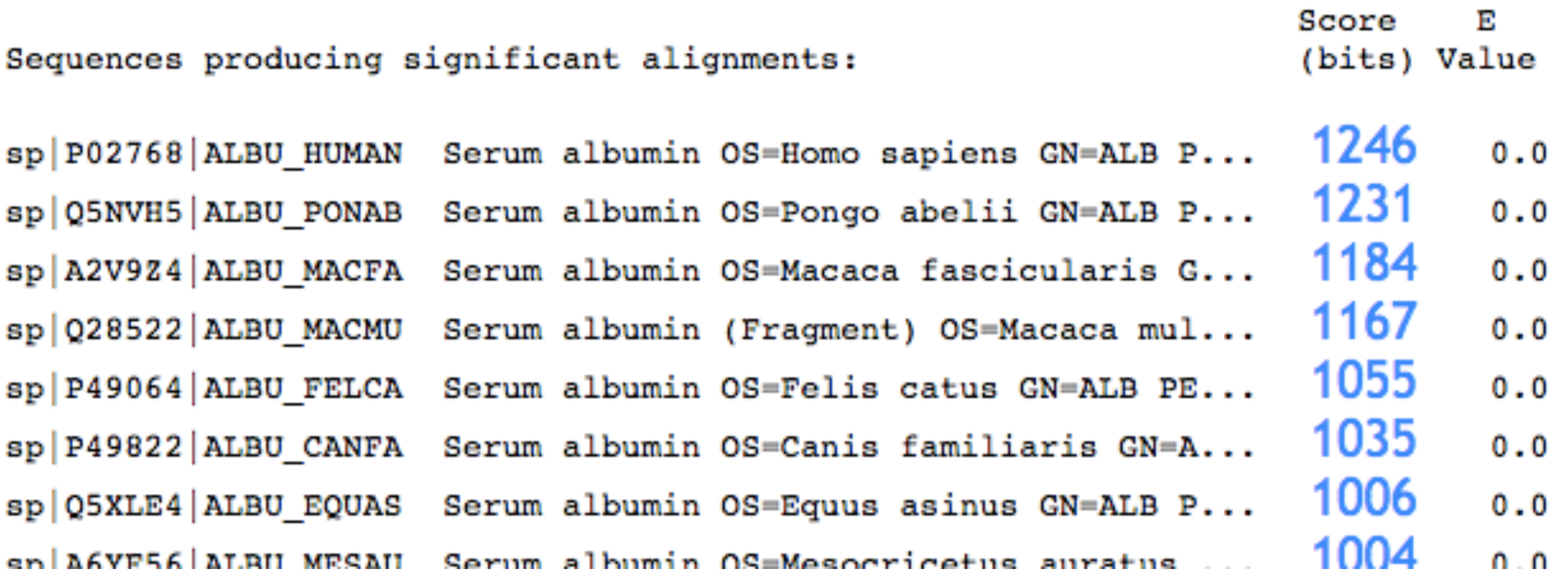
```
P02768|ALBU_HUMAN Serum albumin OS=Homo sapiens
1 SV=2
609 letters)
```



Distribution of 85 Blast Hits on the Query Sequence

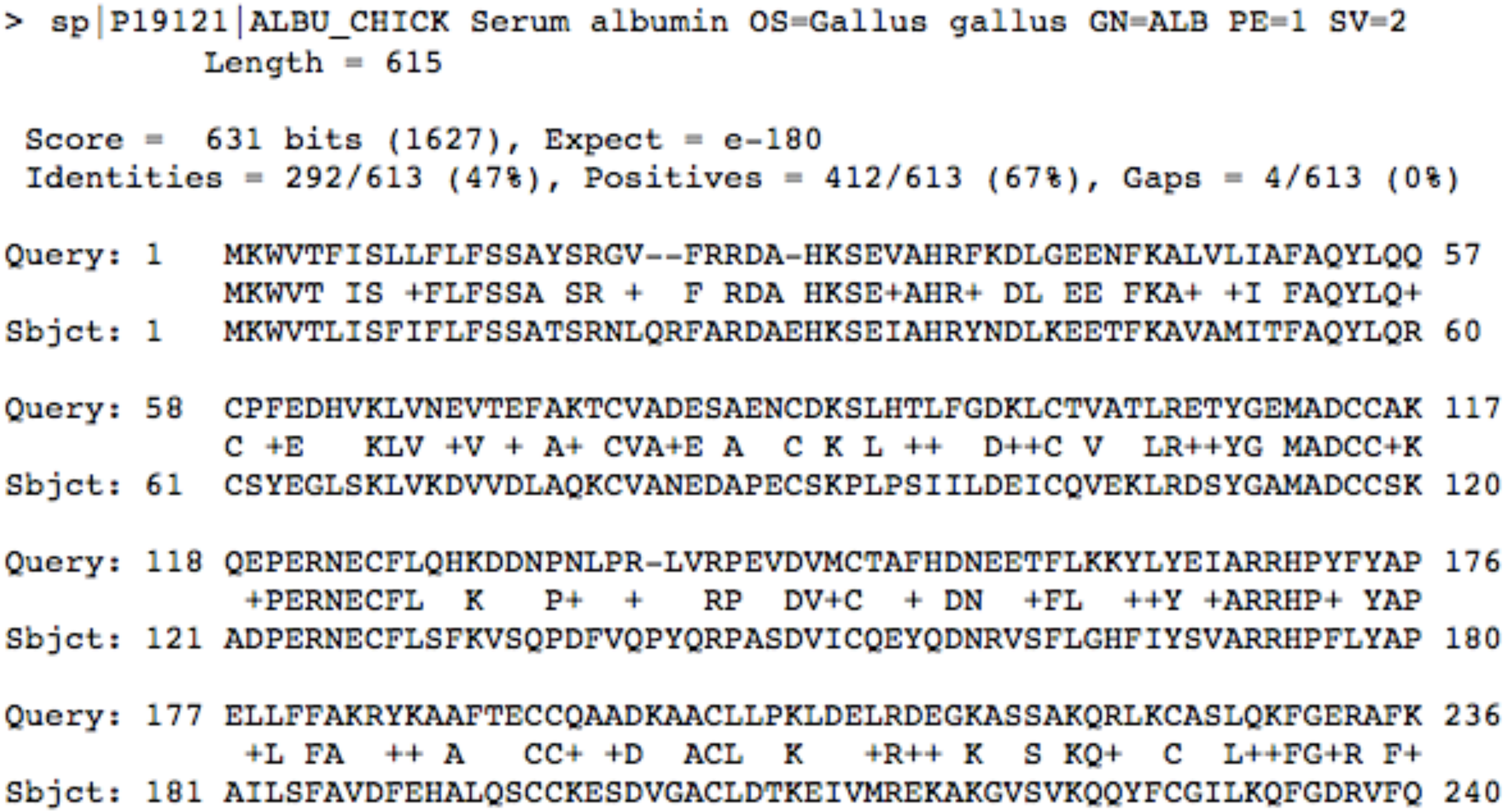
Mouse-over to show defline and scores. Click to show alignments

BLAST



```
                                                                     Score    E
Sequences producing significant alignments:                         (bits) Value

sp|P02768|ALBU_HUMAN  Serum albumin OS=Homo sapiens GN=ALB P...     1246    0.0
sp|Q5NVH5|ALBU_PONAB  Serum albumin OS=Pongo abelii GN=ALB P...     1231    0.0
sp|A2V9Z4|ALBU_MACFA  Serum albumin OS=Macaca fascicularis G...     1184    0.0
sp|Q28522|ALBU_MACMU  Serum albumin (Fragment) OS=Macaca mul...     1167    0.0
sp|P49064|ALBU_FELCA  Serum albumin OS=Felis catus GN=ALB PE...     1055    0.0
sp|P49822|ALBU_CANFA  Serum albumin OS=Canis familiaris GN=A...     1035    0.0
sp|Q5XLE4|ALBU_EQUAS  Serum albumin OS=Equus asinus GN=ALB P...     1006    0.0
sp|A6YF56|ALBU_MESAU  Serum albumin OS=Mesocricetus auratus ...     1004    0.0
```

BLAST



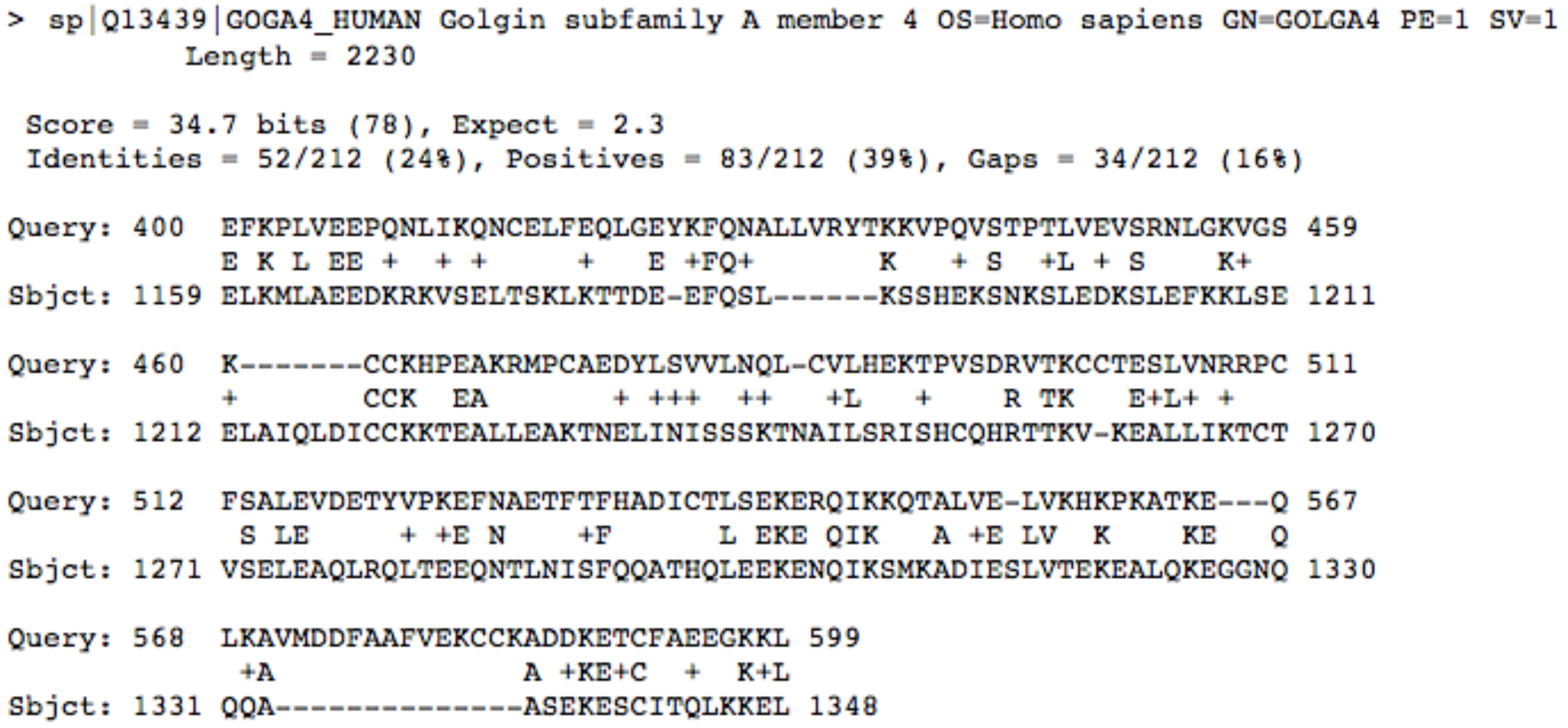
```
> sp|P19121|ALBU_CHICK Serum albumin OS=Gallus gallus GN=ALB PE=1 SV=2
          Length = 615

 Score =  631 bits (1627), Expect = e-180
 Identities = 292/613 (47%), Positives = 412/613 (67%), Gaps = 4/613 (0%)

Query: 1   MKWVTFISLLFLFSSAYSRGV--FRRDA-HKSEVAHRFKDLGEENFKALVLIAFAQYLQQ 57
            MKWVT IS +FLFSSA SR +  F RDA HKSE+AHR+ DL EE FKA+ +I FAQYLQ+
Sbjct: 1   MKWVTLISFIFLFSSATSRNLQRFARDAEHKSEIAHRYNDLKEETFKAVAMITFAQYLQR 60

Query: 58  CPFEDHVKLVNEVTEFAKTCVADESAENCDKSLHTLFGDKLCTVATLRETYGEMADCCAK 117
            C +E   KLV +V + A+ CVA+E A  C K L ++  D++C V  LR++YG MADCC+K
Sbjct: 61  CSYEGLSKLVKDVVDLAQKCVANEDAPECSKPLPSIILDEICQVEKLRDSYGAMADCCSK 120

Query: 118 QEPERNECFLQHKDDNPNLPR-LVRPEVDVMCTAFHDNEETFLKKYLYEIARRHPYFYAP 176
             +PERNECFL  K   P+  +   RP  DV+C  + DN  +FL  ++Y +ARRHP+ YAP
Sbjct: 121 ADPERNECFLSFKVSQPDFVQPYQRPASDVICQEYQDNRVSFLGHFIYSVARRHPFLYAP 180

Query: 177 ELLFFAKRYKAAFTECCQAADKAACLLPKLDELRDEGKASSAKQRLKCASLQKFGERAFK 236
             +L FA  ++ A   CC+ +D  ACL  K   +R++ K  S KQ+  C  L++FG+R F+
Sbjct: 181 AILSFAVDFEHALQSCCKESDVGACLDTKEIVMREKAKGVSVKQQYFCGILKQFGDRVFQ 240
```

BLAST

- E value ?

- Indicate the probability of finding random similarities

BLAST



```
> sp|Q13439|GOGA4_HUMAN Golgin subfamily A member 4 OS=Homo sapiens GN=GOLGA4 PE=1 SV=1
          Length = 2230

 Score = 34.7 bits (78), Expect = 2.3
 Identities = 52/212 (24%), Positives = 83/212 (39%), Gaps = 34/212 (16%)

Query: 400  EFKPLVEEPQNLIKQNCELFEQLGEYKFQNALLVRYTKKVPQVSTPTLVEVSRNLGKVGS 459
            E K L EE +  + +     +   E +FQ+       K   + S  +L + S    K+
Sbjct: 1159 ELKMLAEEDKRKVSELTSKLKTTDE-EFQSL------KSSHEKSNKSLEDKSLEFKKLSE 1211

Query: 460  K-------CCKHPEAKRMPCAEDYLSVVLNQL-CVLHEKTPVSDRVTKCCTESLVNRRPC 511
            +       CCK  EA       + +++  ++   +L   +    R TK   E+L+ +
Sbjct: 1212 ELAIQLDICCKKTEALLEAKTNELINISSSKTNAILSRISHCQHRTTKV-KEALLIKTCT 1270

Query: 512  FSALEVDETYVPKEFNAETFTFHADICTLSEKERQIKKQTALVE-LVKHKPKATKE---Q 567
             S LE     + +E N    +F      L EKE QIK   A +E LV  K    KE   Q
Sbjct: 1271 VSELEAQLRQLTEEQNTLNISFQQATHQLEEKENQIKSMKADIESLVTEKEALQKEGGNQ 1330

Query: 568  LKAVMDDFAAFVEKCCKADDKETCFAEEGKKL 599
             +A              A +KE+C  +  K+L
Sbjct: 1331 QQA--------------ASEKESCITQLKKEL 1348
```

BLAST

- Blast is useful to:

- Identify unknown sequences
- Find homologous sequences in other species
- Locating domains
- Find the position of a sequence in a genome

- Blast is not suited to:

- Dealing with HUGE Datasets! (ex: reads coming from sequencing)