Introduction to BLAST with Protein Sequences

Utah State University – Spring 2014
STAT 5570: Statistical Bioinformatics
Notes 6.2

References

- Chapter 2 of Biological Sequence Analysis (Durbin et al., 2001)
- http://blast.ncbi.nlm.nih.gov/Blast.cgi

Basic Local Alignment Search Tool

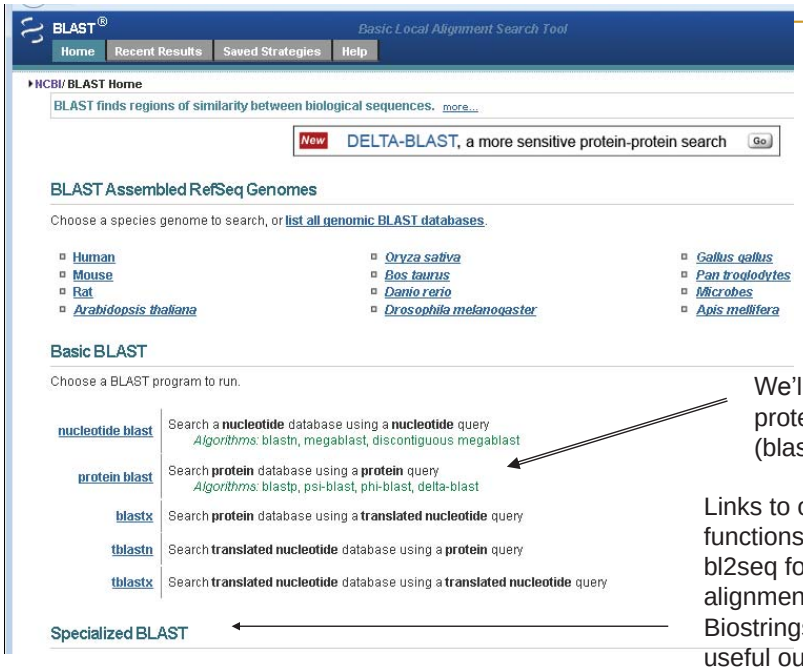
- Finds regions of similarity between a query sequence and a growing database
- Breaks query & database into fragments: (words)
- Starts out by aligning fragments, then extends alignment
- Not optimal alignment, but a good heuristic approach

simplified, rule of thumb

Recall FASTA file from pairseqsim library

```
library(Biostrings)
f1 <- "C://folder//ex.fasta"   # see slide 20 of Notes 6.1
ff <- readAAStringSet(f1, "fasta")
as.character(ff[1])
```

```
At1g01010 NAC domain protein, putative
MEDQVGFGFRPNDEELVGHYLRNKIEGNTSRDVEVAISEVNICSYDPWNLRFQSKYKSRDA
MWYFFSRRENNKGNRQSRTTVSGKWKLTGESVEVKDQWGFCSEGFRGKIGHKRVLVFLDGR
YPDKTKSDWVIHEFHYDLLPEHQRTYVICRLEYKGDDADILSAYAIDPTPAFVPNMTSSAG
SVVNQSRQRNSGSYNTYSEYDSANHGQQFNENSNIMQQQPLQGSFNPLLEYDFANHGGQWL
SDYIDLQQQVPYLAPYENESEMIWKHVIEENFEFLVDERTSMQQHYSDHRPKKPVSGVLPD
DSSDTETGSMIFEDTSSSTDSVGSSDEPGHTRIDDIPSLNIIEPLHNYKAQEQPKQQSKEK
VISSQKSECEWKMAEDSIKIPPSTNTVKQSWIVLENAQWNYLKNMIIGVLLFISVISWIIL
VG
```



We'll start here, with protein BLAST: (blastp)

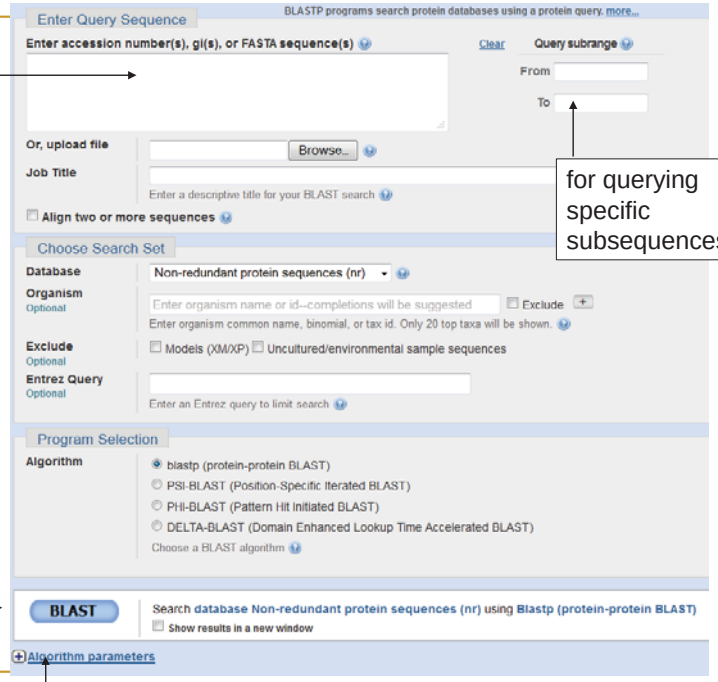
Links to other functions, including bl2seq for pairwise alignment (similar to Biostrings, but less useful output)

http://blast.ncbi.nlm.nih.gov/Blast.cgi

paste sequence here (or accession / gi number)

nr database is most general

click here to start search

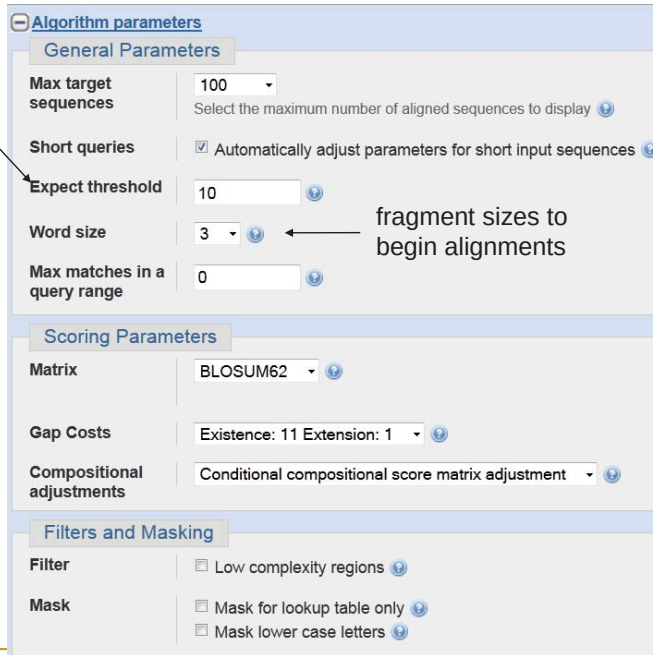


for querying specific subsequences

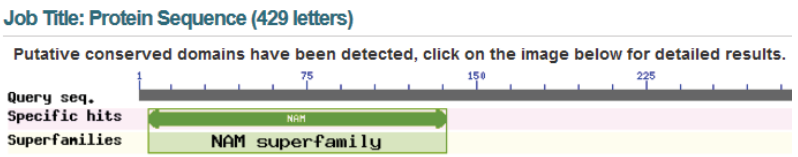
set parameters (next slide)

"significant" matches expected just by chance in database

select scoring scheme here



fragment sizes to begin alignments

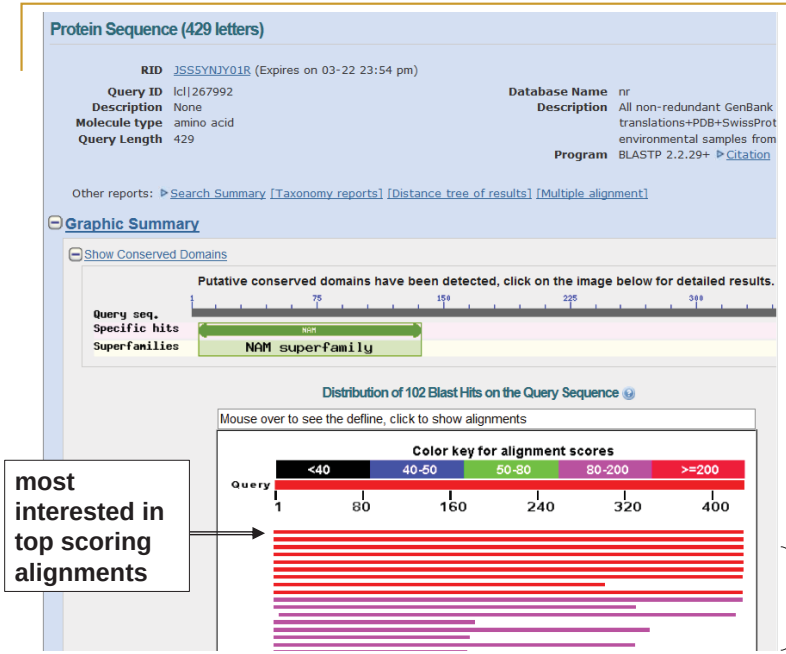


Request ID	JSS5YNJY01R
Status	Searching
Submitted at	Fri Mar 21 11:54:38 2014
Current time	Fri Mar 21 11:55:13 2014
Time since submission	00:00:35

This page will be automatically updated in **10** seconds

Wait for Results

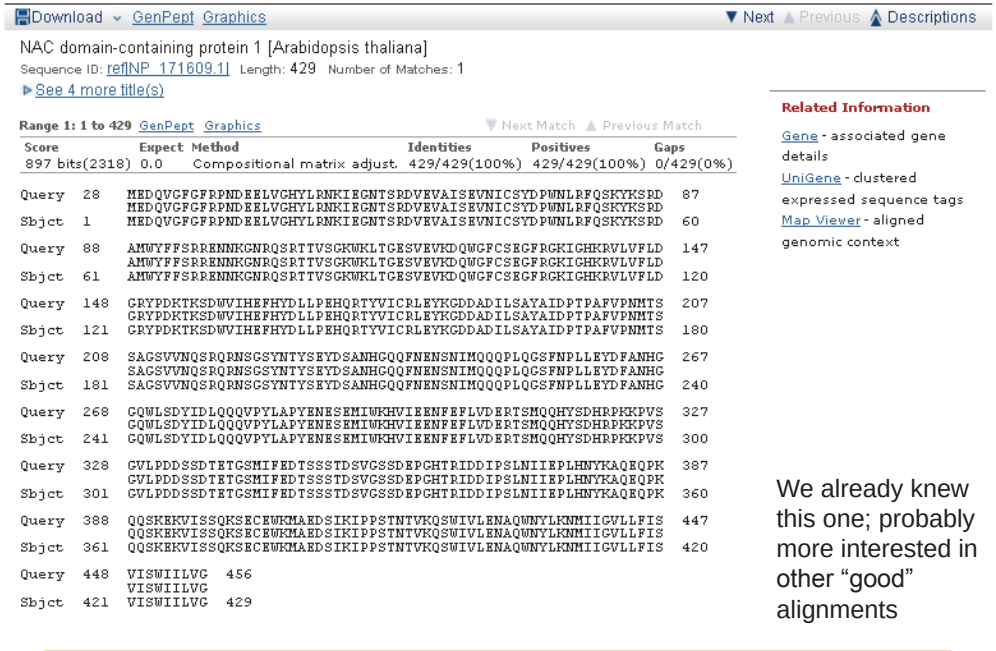
- Large database to search
  - time also: "traffic"-dependent
- This can be automated somewhat
  (write scripts to deliver & format results)
- Can arrange for e-mail to link of final results



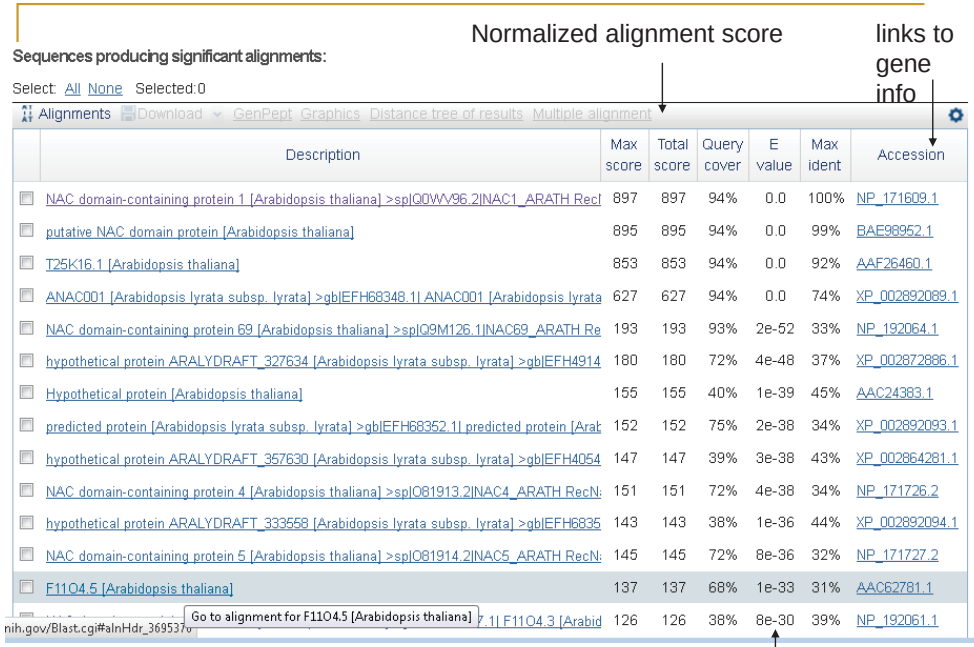
most interested in top scoring alignments

Visualize alignments, colored by score

Scroll down or click on alignments to see more information



We already knew this one; probably more interested in other “good” alignments



The expected # of sequences to score higher than this one in the database search, just by chance

Another “good” alignment

F11O4.5 [Arabidopsis thaliana]

Sequence ID: gb|AAC62781.1| Length: 310 Number of Matches: 1

Range 1: 2 to 281 GenPept Graphics

Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
137 bits(345)	1e-33	Compositional matrix adjust.	99/323(31%)	155/323(47%)	54/323(16%)

```
Query  28   MEDQVGFGFRPNDEELVGHYLRNKIEGNTSRDVEVAISEVNICSYDPWNLRFQSKYKSRD  87
            ++D VG+ F P  EEL+ HYL+NKI          +A+S++                KS D
Sbjct  2    VKDLVGYRFYPTGEELINHYLKNKI----------LALSKI----------------KSDD  36

Query  88   AMWYFFSRRENNKGNRQ--SRTTVSGKWKLTGESVEVKDQWGFCSEGFRGKIGHKRVLVF  145
             +WYFF  +E     ++    RTT SG WK TG   ++KD+       G RG+IG K+ LV+
Sbjct  37   PVWYFFCPKEYTSAKKKVTKRTTSSGYWKATGVDRKIKDK-----RGNRGEIGIKKTLVY  91

Query  146  LDGRYPDKTKSDWVIHEFHYDLLPEHQRTYVICRLEYKGDDADILSA--YAIDPTPAFVP  203
             +GR P    + WV+HE+H   LP+ QR YVIC++ YKG+D D+ S    + +P+ + V
Sbjct  92   YEGRVPKGVWTPWVMHEYHITCLPQDQRNYVICQVMYKGEDGDVPSGGNNSSEPSQSLVS  151

Query  204  NMTSSAGSVVNQSRQRNSGSYNTYSEYDSANHGQQFNENSNIMQQQPLQGSFNPLLEYDF  263
            +  +   +           G  N +       + G   NE  +       L    +P + +
Sbjct  152  DSNTVRATSPTALEFEKPGQENFFG-MSVDDLGTPKNEQEDFS----LWDVLDPDMLFSD  206

Query  264  ANHGGQWLSDYIDLQQQVPYLAPYENESEMIWKHVIEENFEFLVDER-------TSMQQH  316
             N+           +  Q P+L P ++E      +HV  E  E+L             SM ++
Sbjct  207  NNNP--------TVHPQAPHLTPNDDEFLGGLRHVNREQVEYLFANEDFISRPTLSMTEN  258

Query  317  YSDHRPKKPVSGVLPDDSSDTET  339
             +DHRPKK +SG++ D SSD+ +
Sbjct  259  RNDHRPKKALSGIIVDYSSDSNS  281
```

Distribution of maximum score

- Think of alignment score as sum of indep. random variables.
- Then the distribution of this score can be considered approximately normal. (why?)
- The asymptotic distribution of the maximum M_N of a series of N independent normal random variables is the extreme value distribution:

$$P\{M_N\} \approx \exp\left(-K \cdot N \cdot e^{\lambda(x-\mu)}\right)$$

- Use this to calculate the probability that the best match from a search of N unrelated sequences has score greater than our observed largest score S.

E-value of a local, ungapped alignment

Let HSP be a "high-scoring segment pair"
(one of the top-scoring ungapped local alignments).
Then the expected number of HSP's with scores
of at least S is the *E-value*:

$$E = K \cdot m \cdot n \cdot \exp(-\lambda \cdot S),$$

As an aside, the bit score is:

$$S' = \frac{\lambda S - \log K}{\log 2}$$

where

m = length of query sequence

n = length of database sequence (or length of database, in residues)

K = natural scale for search space size
λ = natural scale for scoring system
} depend on q_a and s(a,b)

Statistical Significance of an Alignment

X = # of HSP's with score at least S

$X \sim Poisson(E); \quad E[X] = E$

$$P\{X = a\} = \exp(-E)\frac{E^a}{a!}$$

$$P\{X = 0\} = \exp(-E)\frac{E^0}{0!} = e^{-E}$$

$$P\{X > 0\} = 1 - P\{X = 0\} = 1 - e^{-E}$$

Probability of observing a result at least as extreme as what was seen, just by chance, when the sequences are all unrelated (Null)
= P-value

Still need to look at biological similarity

Protein Sequence (429 letters)

RID JSS5YNJY01R (Expires on 03-22 23:5
Query ID lcl|267992
Description None
Molecule type amino acid
Query Length 429

Other reports: ▷Search Summary [Taxonomy reports]

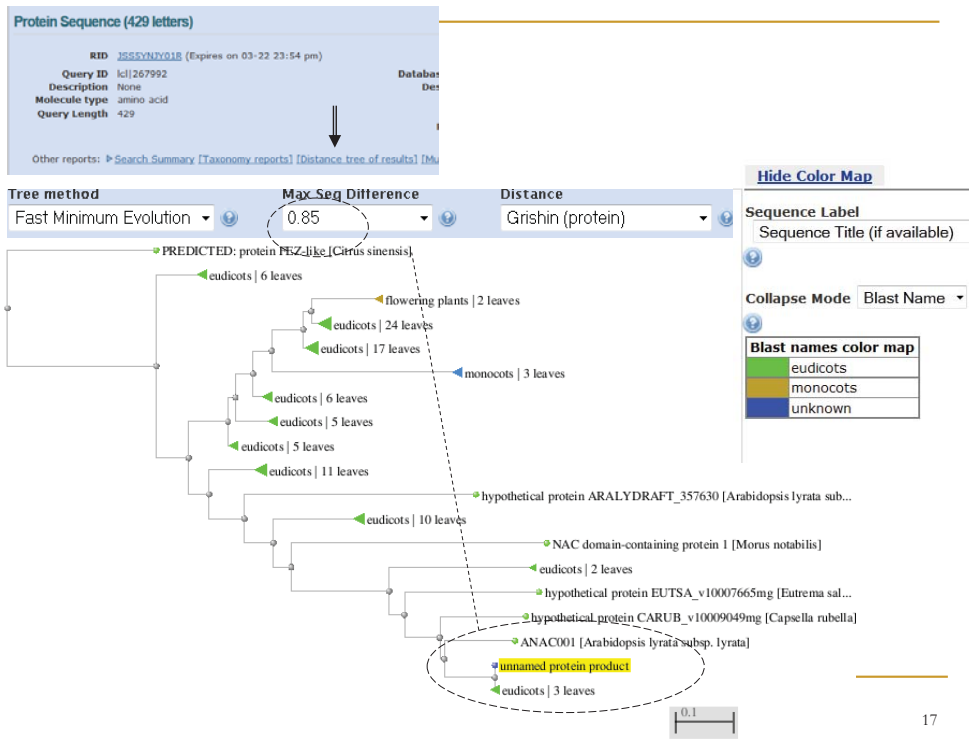
Search Parameters	
Program	blastp
Word size	3
Expect value	10
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Filter string	F
Genetic Code	1
Window Size	40
Threshold	11
Composition-based stats	2

Database	
Posted date	Mar 17, 2014 4:14 PM
Number of letters	13,475,590,452
Number of sequences	37,891,398
Entrez query	none

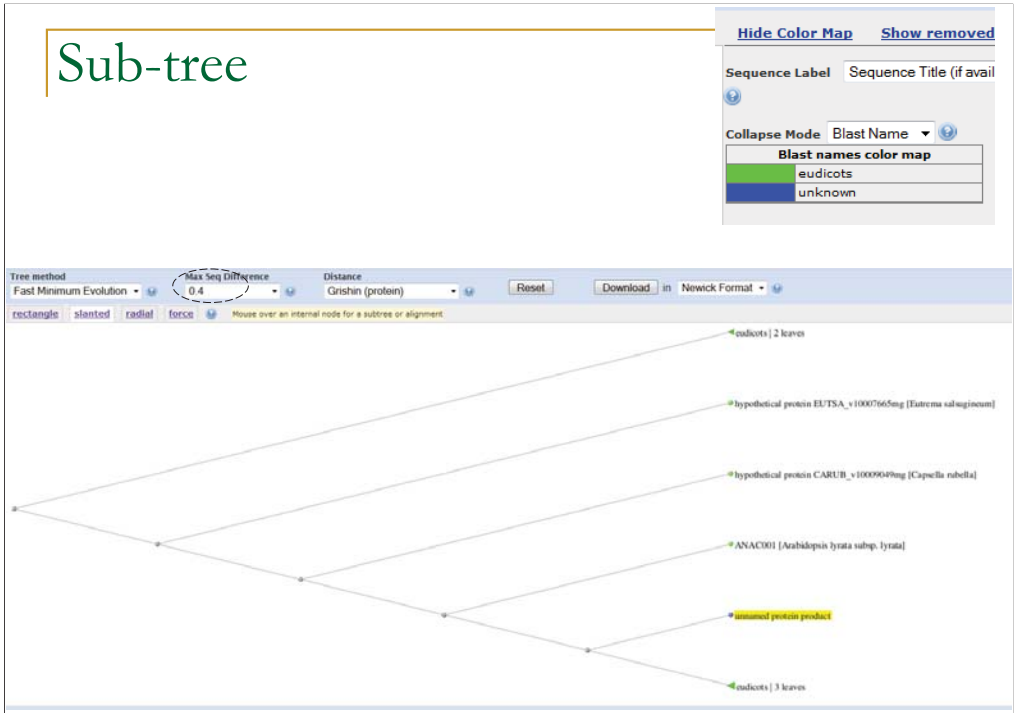
Karlin-Altschul statistics		
Lambda	0.314361	0.267
K	0.131705	0.041
H	0.399222	0.14
Alpha	0.7916	1.9
Alpha_v	4.96466	42.6028
Sigma		43.6362

Estimated values for λ & K

H = relative "entropy" of target and background residue frequencies
(like the average information available per position to distinguish an alignment from chance)



Sub-tree



Other resources to consider

- Altschul et al. (1990) “Basic Local Alignment Search Tool.” Journal of Molecular Biology 215:403-410.
- Mitrophanov & Borodovsky (2006) “Statistical significance in biological sequence analysis.” Briefings in Bioinformatics 7(1):2-24

Summary

- BLAST
  look at finding global alignment
  from smaller, local alignments
- E-value
  expected number of better scores
  just by chance,
  when sequences are not related
- Statistical significance ≠ biological relevance