
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  
MWYFFSRRENNKGNRQSRTTVSGKWKLTTGESVEVKDQWGFCSEGFRGKIGHKRVLVFLDGR  
YDKTKSDWVIHEFHLDLLPEHQRTYVICRLEYKGDDADILSAYAIDPTPAFVPMNTSSAG  
SVVNQSRQRNSGSYNTYSEYDSANHGQQFNENSNIMQQQPLQGSFNPLLEYDFANHGGQWL  
SDYIDLQQQVPYLAPYENESEMIWKHVIEENFEFLVDERTSMQQHYSDHRPKKPVSGVLPD  
DSSDTETGSMIFEDTSSSTDSVGSSEDEPGHTRIDDIPSLNIEPLHNYKAQEOPKQOSKEK  
VISSQKSECEWKMAEDSIKIPSTNTVKQSWIVLENAQWNYLKNMIIGVLLFISVISWIIL  
VG
```

NCBI/BLAST Home

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

New DELTA-BLAST, a more sensitive protein-protein search

BLAST Assembled RefSeq Genomes

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

- [Human](#)
- [Oryza sativa](#)
- [Gallus gallus](#)
- [Mouse](#)
- [Bos taurus](#)
- [Pan troglodytes](#)
- [Rat](#)
- [Danio rerio](#)
- [Microbes](#)
- [Arabidopsis thaliana](#)
- [Drosophila melanogaster](#)
- [Apis mellifera](#)

Basic BLAST

Choose a BLAST program to run.

- [nucleotide blast](#) | Search a **nucleotide** database using a **nucleotide** query
Algorithms: blastn, megablast, discontinuous megablast
- [protein blast](#) | Search **protein** database using a **protein** query
Algorithms: blastp, psi-blast, phi-blast, delta-blast
- [blastx](#) | Search **protein** database using a **translated nucleotide** query
- [tblastn](#) | Search **translated nucleotide** database using a **protein** query
- [tblastx](#) | Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

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

Links to other functions, including b2seq for pairwise alignment (similar to Biostrings, but less useful output)

paste sequence here
(or accession / gi number)



Clear

Query subrange

From

To

for querying specific subsequences



nr database is most general



Choose Search Set

Database

Non-redundant protein sequences (nr)

Organism
Optional

Enter organism name or id--completions will be suggested Exclude +

Exclude
Optional

Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query
Optional

Enter an Entrez query to limit search

Program Selection

Algorithm

- blastp (protein-protein BLAST)
 - PSI-BLAST (Position-Specific Iterated BLAST)
 - PHI-BLAST (Pattern Hit Initiated BLAST)
 - DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
- Choose a BLAST algorithm

click here to start search

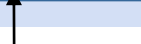


BLAST

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

+ Algorithm parameters

set parameters (next slide)



“significant”
matches
expected just
by chance in
database



select scoring
scheme here →

Algorithm parameters

General Parameters

Max target sequences	100	?
Short queries	<input checked="" type="checkbox"/> Automatically adjust parameters for short input sequences	?
Expect threshold	10	?
Word size	3	?
Max matches in a query range	0	?

fragment sizes to begin alignments ←

Scoring Parameters

Matrix	BLOSUM62	?
Gap Costs	Existence: 11 Extension: 1	?
Compositional adjustments	Conditional compositional score matrix adjustment	?

Filters and Masking

Filter	<input type="checkbox"/> Low complexity regions	?
Mask	<input type="checkbox"/> Mask for lookup table only	?
	<input type="checkbox"/> Mask lower case letters	?

Job Title: Protein Sequence (429 letters)

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



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

Protein Sequence (429 letters)

RID [JSS5YNJY01R](#) (Expires on 03-22 23:54 pm)

Query ID |cl|267992
Description None
Molecule type amino acid
Query Length 429

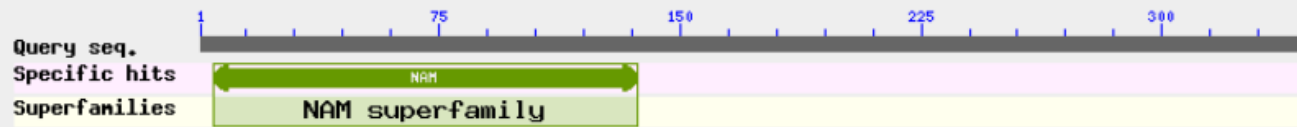
Database Name nr
Description All non-redundant GenBank translations+PDB+SwissProt environmental samples from
Program BLASTP 2.2.29+ [Citation](#)

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

Graphic Summary

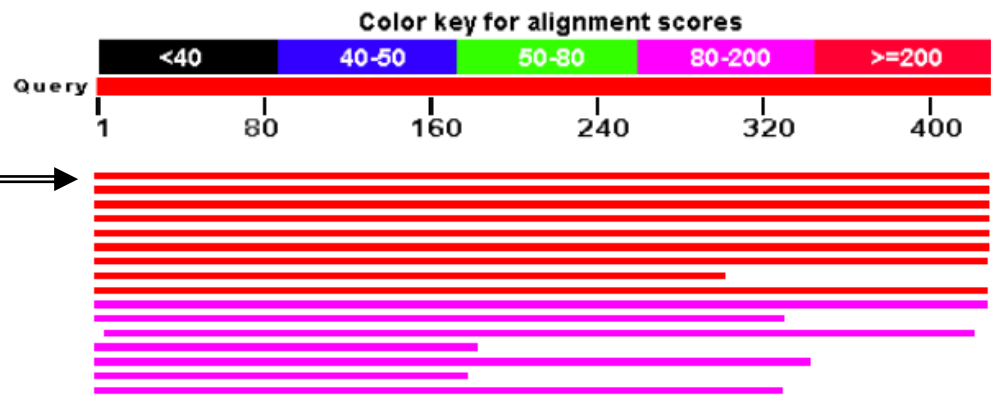
Show Conserved Domains

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



Distribution of 102 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments



most interested in top scoring alignments

Visualize alignments, colored by score

Scroll down or click on alignments to see more information

NAC domain-containing protein 1 [Arabidopsis thaliana]

Sequence ID: [ref|NP_171609.1](#) Length: 429 Number of Matches: 1[▶ See 4 more title\(s\)](#)Range 1: 1 to 429 [GenPept](#) [Graphics](#)

▾ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
897 bits(2318)	0.0	Compositional matrix adjust.	429/429(100%)	429/429(100%)	0/429(0%)
Query 28		MEDQVGFGRPNDEELVGHYLRNKIEGNTSRDVEVAISEVNICSYDPWNLRFAQSKYKSRD			87
Sbjct 1		MEDQVGFGRPNDEELVGHYLRNKIEGNTSRDVEVAISEVNICSYDPWNLRFAQSKYKSRD			60
Query 88		AMWYFFSRRENKGNRQSRRTTVSGKWKLTCESVEVKDQWGFCSEGFGRGKIGHKRVLVFLD			147
Sbjct 61		AMWYFFSRRENKGNRQSRRTTVSGKWKLTCESVEVKDQWGFCSEGFGRGKIGHKRVLVFLD			120
Query 148		GRYPDKTKSDWVIHFHYDLLPEHQRTYVICRLEYKGGDADILSAYAIDPTPAFVPMNTS			207
Sbjct 121		GRYPDKTKSDWVIHFHYDLLPEHQRTYVICRLEYKGGDADILSAYAIDPTPAFVPMNTS			180
Query 208		SAGSVVNQSRQRNSGSYNTYSEYDSANHCQQFNENSNIMQQQLQGSFNPLLEYDFANHC			267
Sbjct 181		SAGSVVNQSRQRNSGSYNTYSEYDSANHCQQFNENSNIMQQQLQGSFNPLLEYDFANHC			240
Query 268		GQWLSDYIDLQQQVPYLAPYENESSEMIWKHVIEENFEFLVDERTSMQQHYSDHRPKKPV			327
Sbjct 241		GQWLSDYIDLQQQVPYLAPYENESSEMIWKHVIEENFEFLVDERTSMQQHYSDHRPKKPV			300
Query 328		GVL PDDSSDTETGSMIFEDTSSSTD SVGSSDEPGHTRIDDIPSLNII EPLHNYKAQE QPK			387
Sbjct 301		GVL PDDSSDTETGSMIFEDTSSSTD SVGSSDEPGHTRIDDIPSLNII EPLHNYKAQE QPK			360
Query 388		QQSKEKVISSQKSECEWKMAEDSIKIPPSNTNTVKQSWIVLENAQWNYLKNMIIIGVLLFIS			447
Sbjct 361		QQSKEKVISSQKSECEWKMAEDSIKIPPSNTNTVKQSWIVLENAQWNYLKNMIIIGVLLFIS			420
Query 448		VISWIIILVG 456			
Sbjct 421		VISWIIILVG 429			

Related Information[Gene](#) - associated gene details[UniGene](#) - clustered expressed sequence tags[Map Viewer](#) - aligned genomic context

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

Normalized alignment score

links to
gene
info

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

	Description	Max score	Total score	Query cover	E value	Max ident	Accession
<input type="checkbox"/>	NAC domain-containing protein 1 [Arabidopsis thaliana] >spIQ0WV96.2 NAC1_ARATH RecI	897	897	94%	0.0	100%	NP_171609.1
<input type="checkbox"/>	putative NAC domain protein [Arabidopsis thaliana]	895	895	94%	0.0	99%	BAE98952.1
<input type="checkbox"/>	T25K16.1 [Arabidopsis thaliana]	853	853	94%	0.0	92%	AAF26460.1
<input type="checkbox"/>	ANAC001 [Arabidopsis lyrata subsp. lyrata] >gbIEFH68348.1 ANAC001 [Arabidopsis lyrata]	627	627	94%	0.0	74%	XP_002892089.1
<input type="checkbox"/>	NAC domain-containing protein 69 [Arabidopsis thaliana] >spIQ9M126.1 NAC69_ARATH Re	193	193	93%	2e-52	33%	NP_192064.1
<input type="checkbox"/>	hypothetical protein ARALYDRAFT_327634 [Arabidopsis lyrata subsp. lyrata] >gbIEFH4914	180	180	72%	4e-48	37%	XP_002872886.1
<input type="checkbox"/>	Hypothetical protein [Arabidopsis thaliana]	155	155	40%	1e-39	45%	AAC24383.1
<input type="checkbox"/>	predicted protein [Arabidopsis lyrata subsp. lyrata] >gbIEFH68352.1 predicted protein [Ara	152	152	75%	2e-38	34%	XP_002892093.1
<input type="checkbox"/>	hypothetical protein ARALYDRAFT_357630 [Arabidopsis lyrata subsp. lyrata] >gbIEFH4054	147	147	39%	3e-38	43%	XP_002864281.1
<input type="checkbox"/>	NAC domain-containing protein 4 [Arabidopsis thaliana] >spIO81913.2 NAC4_ARATH RecN:	151	151	72%	4e-38	34%	NP_171726.2
<input type="checkbox"/>	hypothetical protein ARALYDRAFT_333558 [Arabidopsis lyrata subsp. lyrata] >gbIEFH6835	143	143	38%	1e-36	44%	XP_002892094.1
<input type="checkbox"/>	NAC domain-containing protein 5 [Arabidopsis thaliana] >spIO81914.2 NAC5_ARATH RecN:	145	145	72%	8e-36	32%	NP_171727.2
<input type="checkbox"/>	F11O4.5 [Arabidopsis thaliana]	137	137	68%	1e-33	31%	AAC62781.1
<input type="checkbox"/>	F11O4.3 [Arabid	126	126	38%	8e-30	39%	NP_192061.1

Go to alignment for F11O4.5 [Arabidopsis thaliana]

ni.gov/Blast.cgi#alnHdr_369537

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	MEDQVGFGERPNDEELVGHYLRNKIEGNTSRDVEVAISEVNICSYDPWNLRFQSKYKSRD				87
	++D VG+ F P EEL+ HYL+NKI		+A+S++		KS D
Sbjct 2	VKDLVGYRFYPTGEELINHYLKNKI-----LALSKI-----KSD				36
Query 88	AMWYFFSRRENNKGNRQ--SRTTVSGKWKLGTGESVEVKDQWGFCSSEGERGKIGHKRVLVF				145
	+WYFF +E ++ RTT SG WK TG ++KD+ G RG+IG K+ LV+				
Sbjct 37	PVWYFFFCPKEYTSAKKKVTKRRTSSSGYWKATGVDRKIKDK-----RGNRGEIGIKKTLVY				91
Query 146	LDGRYPDKTKSDWVIHEFHFDLLPEHQRTYVICRLEYKGGDDADILSA--YAIDPTPAFVP				203
	+GR P + WV+HE+H LP+ QR YVIC++ YKG+D D+ S + +P+ + V				
Sbjct 92	YEGRVPKGVWTPWVMHEYHITCLPQDQRNYVICQVMYKGEDGDVPSGGNNSSEPSQSLVS				151
Query 204	NMTSSAGSVVNQSRQRNSGSYNTYSEYDSANHGQQFNENSNIMQQQPLOGSFNPLLEYDF				263
	+ + + G N + + G NE + L +P + +				
Sbjct 152	DSNTVRATSPTALEFEKPGQENFFG-MSVDDLGT PKNEQEDEF S----LWDVLDPDMLEFSD				206
Query 264	ANHGGQWLSDYIDLQQQVPYLAPYENESEMIWKHVIEENFEFLVDER-----TSMQQH				316
	N+ + Q P+L P ++E +HV E E+L SM ++				
Sbjct 207	NNNP-----TVHPQAPHLT PNDDEF LGGLRHVNREQVEYLFANEDFISRPTLSMTEN				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

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

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

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

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

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](#)]



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)

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

Protein Sequence (429 letters)

RID [JSS5YNJY01R](#) (Expires on 03-22 23:54 pm)

Query ID |cl|267992

Description |None

Molecule type |amino acid

Query Length |429

Databases

Des



Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#) [Mu](#)

Tree method

Fast Minimum Evolution

Max Seq Difference

0.85

Distance

Grishin (protein)

[Hide Color Map](#)

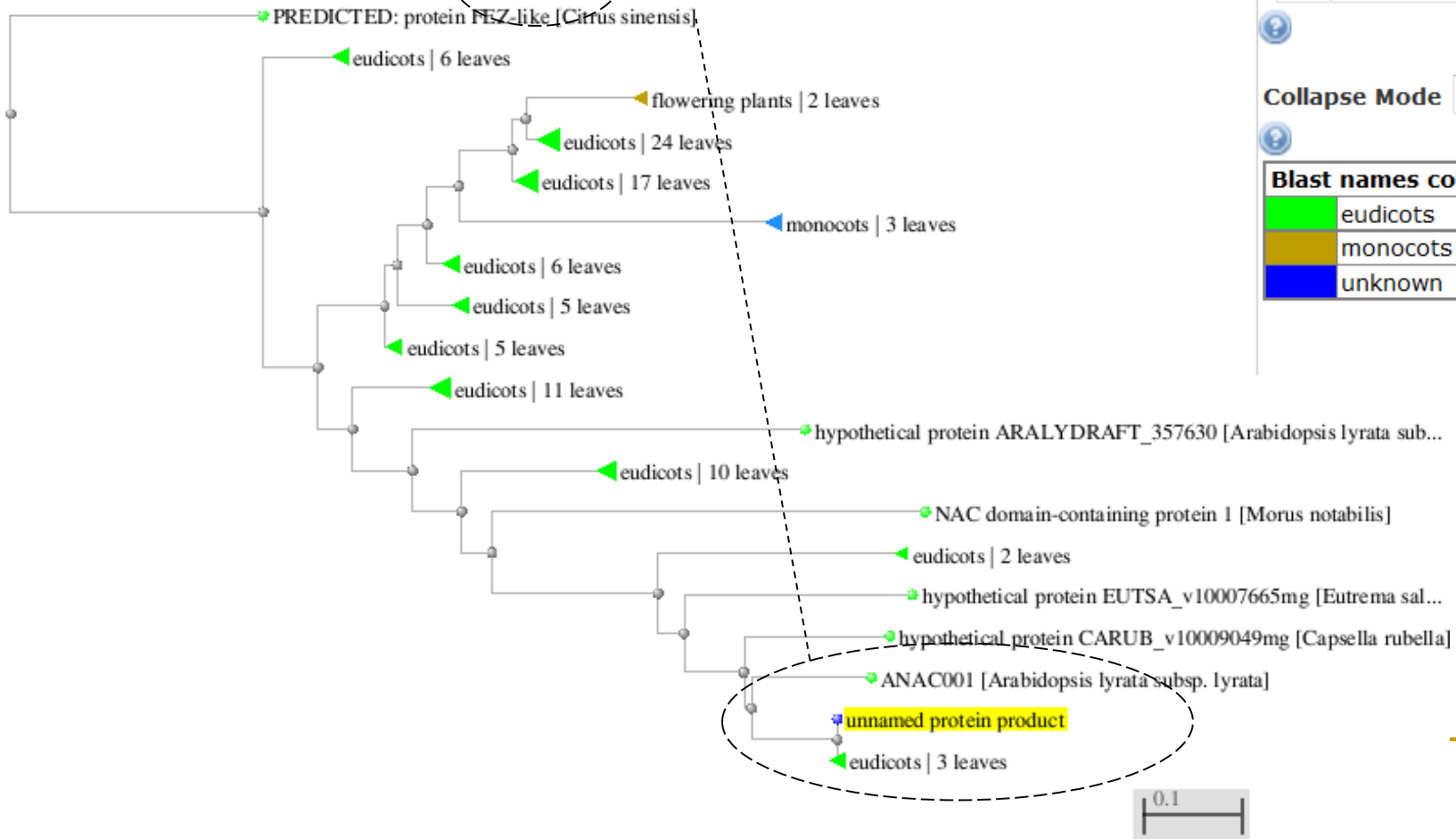
Sequence Label

Sequence Title (if available)

Collapse Mode [Blast Name](#)

Blast names color map

	eudicots
	monocots
	unknown



Sub-tree

[Hide Color Map](#) [Show removed](#)

Sequence Label

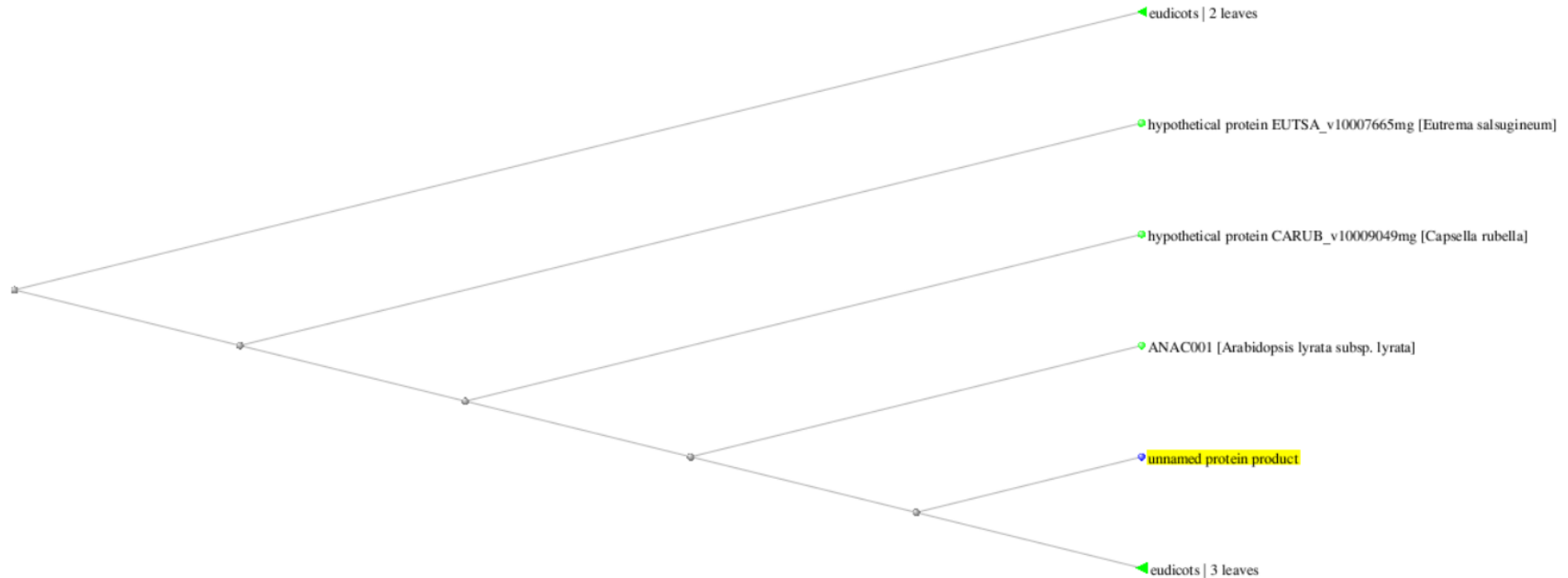
Collapse Mode

Blast names color map

	eudicots
	unknown

Tree method: Max Seq Difference: Distance: in

Mouse over an internal node for a subtree or alignment



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 \neq biological relevance