Introduction to Bioinformatics

Course NR. 22111

BLAST3

Stefan Olevinskiy
s246026

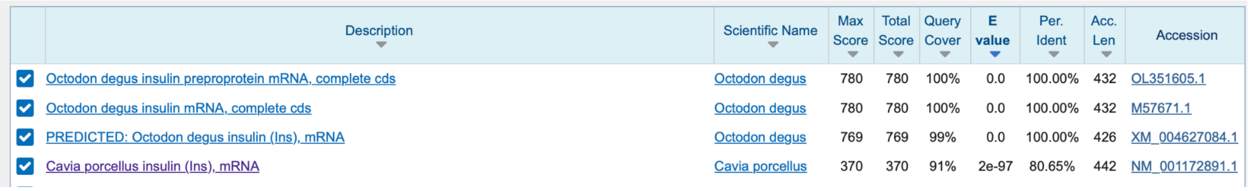
Polina Krasikova
s245850

Part 1 — Our first BLAST Search

For our first search, we ran the mRNA sequence for insulin from a South American rodent, the Degu (Octodon degus) through NCBI's BLASTN. We searched the "Nucleotide collection (nr/nt)" database. NR is the "Non Redundant" database, which contains all non-redundant (non-identical) sequences from GenBank and the full genome databases.

```
>gi|202471|gb|M57671.1|OCOINS Octodon degus insulin mRNA, complete cds
GCATTCTGAGGCATTCTCTAACAGGTTCTCGACCCTCCGCCATGGCCCCGTGGATGCATCTCCTCACCGT
GCTGGCCCTGCTGGCCCTCTGGGGACCCAACTCTGTTCAGGCCTATTCCAGCCAGCACCTGTGCGGCTCC
AACCTAGTGGAGGCACTGTACATGACATGTGGACGGAGTGGCTTCTATAGACCCCACGACCGCCGAGAGC
TGGAGGACCTCCAGGTGGAGCAGGCAGAACTGGGTCTGGAGGCAGGCGGCCTGCAGCCTTCGGCCCTGGA
GATGATTCTGCAGAAGCGCGGCATTGTGGATCAGTGCTGTAATAACATTTGCACATTTAACCAGCTGCAG
AACTACTGCAATGTCCCTTAGACACCTGCCTTGGGCCTGGCCTGCTGCTCTGCCCTGGCAACCAATAAAC
CCCTTGAATGAG
```

Then, we looked at the top hit, apart from the octodon degus insulin sequences. The top hit is Cavia porcellus insulin mRNA. The accession, alignment score, percent identity, query coverage and the E-value are all seen in the initial "Descriptions" section on the results page:



	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Octodon degus insulin preproprotein mRNA, complete cds	Octodon degus	780	780	100%	0.0	100.00%	432	OL351605.1
☑	Octodon degus insulin mRNA, complete cds	Octodon degus	780	780	100%	0.0	100.00%	432	M57671.1
☑	PREDICTED: Octodon degus insulin (Ins), mRNA	Octodon degus	769	769	99%	0.0	100.00%	426	XM_004627084.1
☑	Cavia porcellus insulin (Ins), mRNA	Cavia porcellus	370	370	91%	2e-97	80.65%	442	NM_001172891.1

To determine if there are any gaps in the alignment, we navigated to the "Alignments" section of the results page, and found the Cavia porcellus sequence. Or, you can also simply click on the hit to be redirected. As you can see below, there are 19 gaps, or 4% of the entire sequence.

Score	Expect	Identities	Gaps	Strand
370 bits(410)	2e-97	325/403(81%)	19/403(4%)	Plus/Plus

```
Query  41   CATGGCCCCGTGGATGCATCTCCTCACCGTGCTGGCCCTGCTGGCCCTCTGGGGACCCAA  100
            |||||| | ||||||||||||||||||||||||||||||||||||||||||||| |||||
Sbjct  48   CATGGCTCTGTGGATGCATCTCCTCACCGTGCTGGCCCTGCTGGCCCTCTGGGGGCCCAA  107
```

```
Query  101  CTCTGTTCAGGCCTATTCCAGCCAGCACCTGTGCGGCTCCAACCTAGTGGAGGCACTGTA  160
            | ||| |||||||| |  ||||| ||| ||||||||||||||| |||||||| || ||||
Sbjct  108  CACTGGTCAGGCCTTTGTCAGCCGGCATCTGTGCGGCTCCAACTTAGTGGAGACATTGTA  167

Query  161  --CA---TGACATGTGGACGGAGTGGCTTCTATAGACCCCACGACCGCCGAGAGCTGGAG  215
              ||   || || |  || ||  |  |||||||| |||| | ||||| || ||||| |||
Sbjct  168  TTCAGTGTGTCAGGATGATGGCTT--CTTCTATATACCCAAGGACCGTCGGGAGCTAGAG  225

Query  216  GACCTCCAGGTGGAGCAGGCAGAAC------TGGGTCTGGAGGCAGGCGGCCTGCAGCCT  269
            ||||  |||||||||||| ||||||      |||| |||| |||||| || || |||||
Sbjct  226  GACCCACAGGTGGAGCAGACAGAACTGGGCATGGGCCTGGGGGCAGGTGGACTACAGCCC  285

Query  270  TCGGCCCTGGAGATGATTCTGCAGAAGCGCGGCATTGTGGATCAGTGCTGTAATAACATT  329
            | ||| |||||||||   || |||||||| |||||||||||||||||||||| |  ||
Sbjct  286  TTGGCACTGGAGATGGCACTACAGAAGCGTGGCATTGTGGATCAGTGCTGTACTGGCACC  345

Query  330  TGCACATTTAACCAGCTGCAGAACTACTGCAATGTCCCTTAGACACCTGCCTTGGGCCTG  389
            ||||||    |||||||||||| |||||||||       |||||||||||||||  ||||
Sbjct  346  TGCACACGCCACCAGCTGCAGAGCTACTGCAA------CTAGACACCTGCCTTGAACCTG  399

Query  390  GCCTGCTGCTCTGCCCTGGCAACCAATAAACCCCTTGAATGAG  432
            |||| |  |||| ||||||||||||||||||||||||||||||
Sbjct  400  GCCTCCCACTCTCCCCTGGCAACCAATAAACCCCTTGAATGAG  442
```

Then, we found the best human hit that is not a synthetic construct. That was “Homo sapiens insulin isoform UC (INS) mRNA, complete cds, alternatively spliced” with accession MT335691.1. Quite interesting that the top human hit is an isoform!

Score	Expect	Identities	Gaps	Strand
205 bits(227)	2e-47	254/341(74%)	15/341(4%)	Plus/Plus

```
Query  33   CCCTCCGCCATGGCCCCGTGGATGCATCTCCTCACCGTGCTGGCCCTGCTGGCCCTCTGG  92
            || || |||||||||| ||||||||  |||||  || ||||||| |||||||||||||||
Sbjct  382  CCTTCTGCCATGGCCCTGTGGATGCGCCTCCTGCCCCTGCTGGCGCTGCTGGCCCTCTGG  441

Query  93   GGACCCAACTCTGTTCAGGCCTATTCCAGCCAGCACCTGTGCGGCTCCAACCTAGTGGAG  152
            |||||  || | |     |||| |   | ||| ||||||||||||||  |||| |||||
Sbjct  442  GGACCTGACCCAGCCGCAGCCTTTGTGAACCAACACCTGTGCGGCTCACACCTGGTGGAA  501

Query  153  GCACTGTACATGACATGTGGACGGA--GTGGCTTCTA-TAGAC-CCCACGACC-GCCGAG  207
```

```
            || || ||| |    || ||  |||  | |||||||  || || |||| |||| |||| |
Sbjct  502  GCTCTCTACCTAGTGTGCGG--GGAACGAGGCTTCTTCTACACACCCAAGACCCGCCGGG  559

Query  208  AGCTGGAGGACCTCCAGGTGGAGCAGGCAGAACTGGGT-------CTGGAGGCAGGCGGC  260
            ||   |||||||| ||||||| |||||  || |||||        |||| | ||||| ||
Sbjct  560  AGGCAGAGGACCTGCAGGTGGGGCAGGTGGAGCTGGGCGGGGGCCCTGGTG-CAGGCAGC  618

Query  261  CTGCAGCCTTCGGCCCTGGAGATGATTCTGCAGAAGCGCGGCATTGTGGATCAGTGCTGT  320
            |||||||| | ||||||||||  |   ||||||||||| ||||||||||| || ||||||
Sbjct  619  CTGCAGCCCTTGGCCCTGGAGGGGTCCCTGCAGAAGCGTGGCATTGTGGAACAATGCTGT  678

Query  321  AATAACATTTGCACATTTAACCAGCTGCAGAACTACTGCAA  361
            |  | ||| ||| |  |  |||||||| |||||||||||||
Sbjct  679  ACCAGCATCTGCTCCCTCTACCAGCTGGAGAACTACTGCAA  719
```

Then, we ran the same sequence through the human genomic + transcript database, getting Homo sapiens insulin (INS), transcript variant 3, mRNA with Sequence ID: NM_001185098.2 as the top hit.

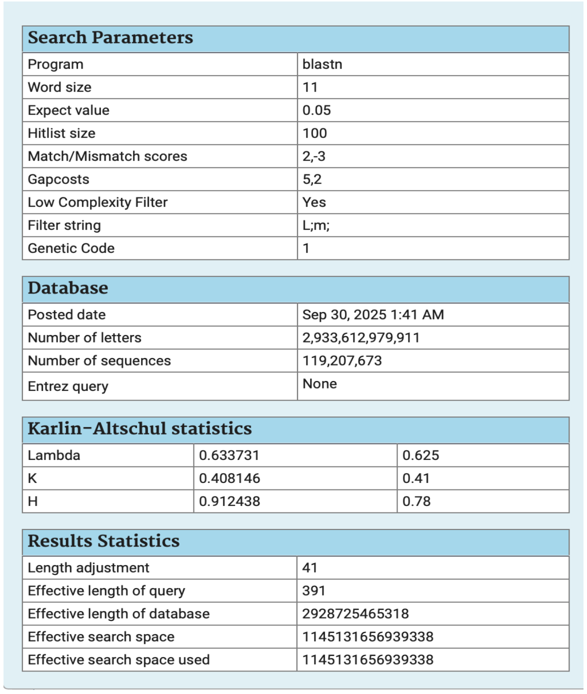
Score	Expect	Identities	Gaps	Strand
205 bits(227)	3e-50	254/341(74%)	15/341(4%)	Plus/Plus

```
Query  33   CCCTCCGCCATGGCCCCGTGGATGCATCTCCTCACCGTGCTGGCCCTGCTGGCCCTCTGG  92
            || || |||||||||| ||||||||  |||||  || ||||||| |||||||||||||||
Sbjct  230  CCTTCTGCCATGGCCCTGTGGATGCGCCTCCTGCCCCTGCTGGCGCTGCTGGCCCTCTGG  289

Query  93   GGACCCAACTCTGTTCAGGCCTATTCCAGCCAGCACCTGTGCGGCTCCAACCTAGTGGAG  152
            |||||  || | |     |||| |   | ||| ||||||||||||||  |||| |||||
Sbjct  290  GGACCTGACCCAGCCGCAGCCTTTGTGAACCAACACCTGTGCGGCTCACACCTGGTGGAA  349

Query  153  GCACTGTACATGACATGTGGACGGA--GTGGCTTCTA-TAGAC-CCCACGACC-GCCGAG  207
            || || ||| |    || ||  |||  | |||||||  || || |||| |||| |||| |
Sbjct  350  GCTCTCTACCTAGTGTGCGG--GGAACGAGGCTTCTTCTACACACCCAAGACCCGCCGGG  407

Query  208  AGCTGGAGGACCTCCAGGTGGAGCAGGCAGAACTGGGT-------CTGGAGGCAGGCGGC  260
            ||   |||||||| ||||||| |||||  || |||||        |||| | ||||| ||
Sbjct  408  AGGCAGAGGACCTGCAGGTGGGGCAGGTGGAGCTGGGCGGGGGCCCTGGTG-CAGGCAGC  466

Query  261  CTGCAGCCTTCGGCCCTGGAGATGATTCTGCAGAAGCGCGGCATTGTGGATCAGTGCTGT  320
```

```
            |||||||| | ||||||||||  |   ||||||||||| ||||||||||| || ||||||
Sbjct  467  CTGCAGCCCTTGGCCCTGGAGGGGTCCCTGCAGAAGCGTGGCATTGTGGAACAATGCTGT  526

Query  321  AATAACATTTGCACATTTAACCAGCTGCAGAACTACTGCAA  361
            |  | ||| ||| |  |  |||||||| |||||||||||||
Sbjct  527  ACCAGCATCTGCTCCCTCTACCAGCTGGAGAACTACTGCAA  567
```

So even though the database entry seems to be different, the alignment is actually exactly the same. This shows that you can do the job faster and easier by selecting the right search database for the purpose.

But the E-value is different this time! It's 3e-50 instead of 2e-47. Why? Let's take a look at the search summaries for both searches.



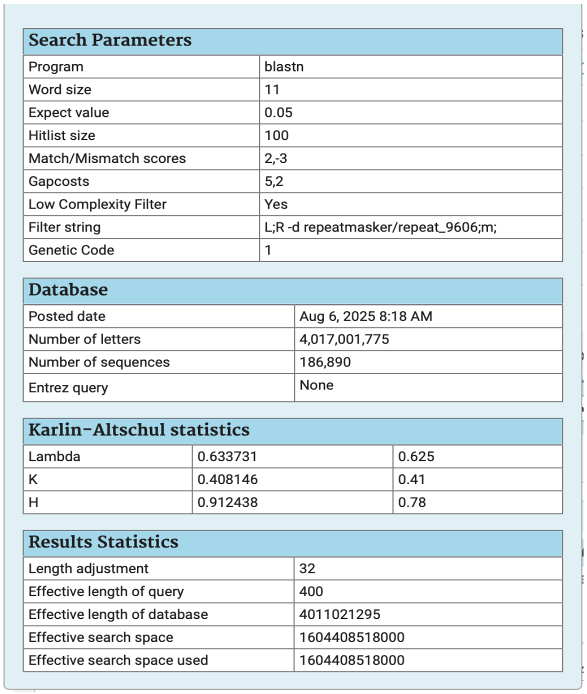
Search Parameters	
Program	blastn
Word size	11
Expect value	0.05
Hitlist size	100
Match/Mismatch scores	2,-3
Gapcosts	5,2
Low Complexity Filter	Yes
Filter string	L;m;
Genetic Code	1

Database	
Posted date	Sep 30, 2025 1:41 AM
Number of letters	2,933,612,979,911
Number of sequences	119,207,673
Entrez query	None

Karlin-Altschul statistics		
Lambda	0.633731	0.625
K	0.408146	0.41
H	0.912438	0.78

Results Statistics	
Length adjustment	41
Effective length of query	391
Effective length of database	2928725465318
Effective search space	1145131656939338
Effective search space used	1145131656939338

Nucleotide collection (nr/nt) database



Search Parameters	
Program	blastn
Word size	11
Expect value	0.05
Hitlist size	100
Match/Mismatch scores	2,-3
Gapcosts	5,2
Low Complexity Filter	Yes
Filter string	L;R -d repeatmasker/repeat_9606;m;
Genetic Code	1

Database	
Posted date	Aug 6, 2025 8:18 AM
Number of letters	4,017,001,775
Number of sequences	186,890
Entrez query	None

Karlin-Altschul statistics		
Lambda	0.633731	0.625
K	0.408146	0.41
H	0.912438	0.78

Results Statistics	
Length adjustment	32
Effective length of query	400
Effective length of database	4011021295
Effective search space	1604408518000
Effective search space used	1604408518000

human genomic + transcript database

We can see that the nucleotide collection (nr/nt) database is a lot larger (2,933,612,979,911 bp vs. 4,017,001,775bp)! The nucleotide collection (nr/nt) database is 730,3 larger than the human genomic + transcript database! And the E-value of the nucleotide collection (nr/nt) database is much larger too!

The E-value is measure of how likely it is to see an alignment with a given score just by chance, given the size of the database:

```
E-value ≈ K · (effective_query_len) · (effective_db_len) · e^(−λS)
```

*“For a fixed alignment score **S** (same HSP, same scoring), E scales **linearly** with the effective database length. Bigger DB → larger E (less “surprising” to see a high score by chance). Smaller DB → smaller E.” –ChatGPT*

So if it scales linearly, then with a database doubling we’d expect a doubling in the E-value too. Here, the database size is 730 larger for the nucleotide collection database, and its E-value is 667 times larger, so in the same ballpark.

Part 2: Assessing the statistical significance of BLAST hits

With BLAST, there is a risk of getting false positive results (hits to sequences that are not related to the input sequence) by purely stochastic means. So we will be examining what happens when we submit randomly generated sequences to BLAST searches.

The sequences were generated using provided code.

Random DNA sequences and BLASTN

```
Generating Sequence 1 for BLAST...
AATGCATGAGGTCCGTAAGGCTCCG

****Alignment**** 1
Title: gi|2874253773|emb|OY969722.1| MAG: uncultured Actinomycetota bacterium
isolate MFD10113.bin.2.32 genome assembly, chromosome: 1
Accession: OY969722
Length: 2813024
Max Score: 20.0
Bits: 40.14
Identities: 20
Align_length: 20
Gaps: 0
%Ident: 100.00 %
Query Cover: 80 %
E value: 3.13e+00
```

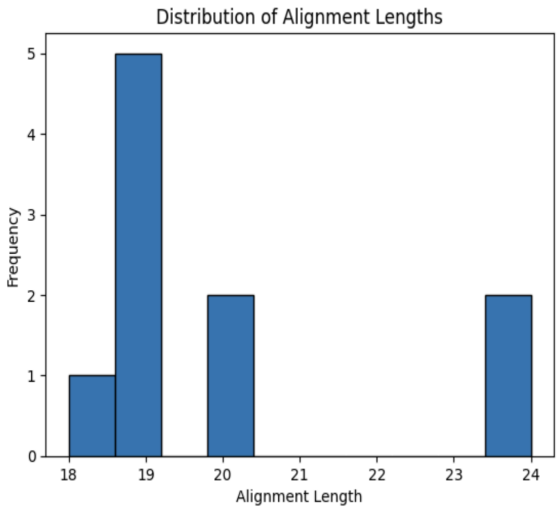
```
Query:   TGCATGAGGTCCGTAAGGCT
Match:   ||||||||||||||||||||
Subject: TGCATGAGGTCCGTAAGGCT


Generating Sequence 2 for BLAST...
TGCAGGCGCACACACCAGAGCGACA

****Alignment**** 1
Title: gi|2548755896|gb|CP128999.1| Rhodococcus opacus strain 3D chromosome 2,
complete sequence
Accession: CP128999
Length: 1906477
Max Score: 20.0
Bits: 40.14
Identities: 20
Align_length: 20
Gaps: 0
%Ident: 100.00 %
Query Cover: 80 %
E value: 3.13e+00
Query:   CAGGCGCACACACCAGAGCG
Match:   ||||||||||||||||||||
Subject: CAGGCGCACACACCAGAGCG


Generating Sequence 3 for BLAST...
TTCATTCAGTTCAGCCCCCTACTAA

****Alignment**** 1
Title: gi|3035640549|emb|OZ296835.1| Luscinia svecica genome assembly,
chromosome: 6
Accession: OZ296835
Length: 62524714
Max Score: 19.0
Bits: 38.1576
Identities: 19
Align_length: 19
Gaps: 0
%Ident: 100.00 %
Query Cover: 76 %
E value: 1.24e+01
Query:   ATTCAGTTCAGCCCCCTAC
Match:   |||||||||||||||||||
Subject: ATTCAGTTCAGCCCCCTAC
```

```
Generating Sequence 4 for BLAST...
AACTAATATTACAGGTACCCCCGAG

****Alignment**** 1
Title: gi|2131027698|ref|XM_045038943.1| PREDICTED: Felis catus uncharacterized
LOC123380507 (LOC123380507), mRNA
Accession: XM_045038943
Length: 4282
Max Score: 20.0
Bits: 40.14
Identities: 23
Align_length: 24
Gaps: 0
%Ident: 95.83 %
Query Cover: 96 %
E value: 3.13e+00
Query:   AACTAATATTACAGGTACCCCCGA
Match:   |||||| |||||||||||||||||
Subject: AACTAAGATTACAGGTACCCCCGA


Generating Sequence 5 for BLAST...
ACTGTGCCGGAGGCGCATCCCCGAG

****Alignment**** 1
Title: gi|2801843087|emb|OZ180145.1| Melanogrammus aeglefinus genome assembly,
chromosome: 13
Accession: OZ180145
Length: 24751165
Max Score: 20.0
Bits: 40.14
Identities: 23
Align_length: 24
Gaps: 0
%Ident: 95.83 %
Query Cover: 96 %
E value: 3.13e+00
Query:   CTGTGCCGGAGGCGCATCCCCGAG
Match:   |||||||||||||||||| |||||
Subject: CTGTGCCGGAGGCGCATCGCCGAG


Generating Sequence 6 for BLAST...
```

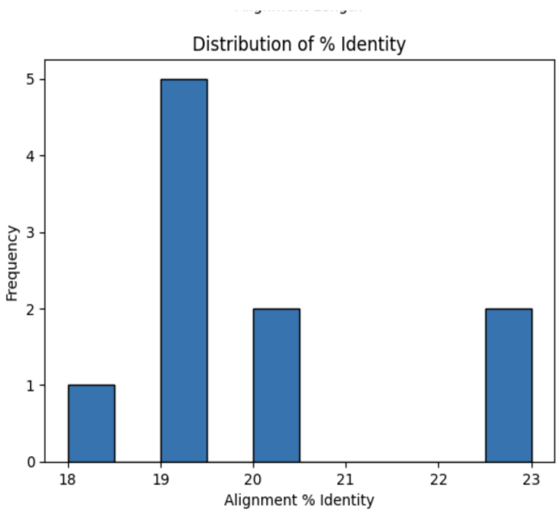
```
GCCGGTATTGGGTCTTCAGTCTGGA

****Alignment**** 1
Title: gi|3061149410|emb|OZ311093.1| Cydia inquinatana genome assembly,
chromosome: 10
Accession: OZ311093
Length: 43182711
Max Score: 19.0
Bits: 38.1576
Identities: 19
Align_length: 19
Gaps: 0
%Ident: 100.00 %
Query Cover: 76 %
E value: 1.24e+01
Query:   CCGGTATTGGGTCTTCAGT
Match:   |||||||||||||||||||
Subject: CCGGTATTGGGTCTTCAGT


Generating Sequence 7 for BLAST...
GTACTTGTTCACTACGGCCGGCTCT

****Alignment**** 1
Title: gi|2514243053|ref|XM_056810364.1| PREDICTED: Monodelphis domestica
transglutaminase 7 (TGM7), mRNA
Accession: XM_056810364
Length: 2410
Max Score: 19.0
Bits: 38.1576
Identities: 19
Align_length: 19
Gaps: 0
%Ident: 100.00 %
Query Cover: 76 %
E value: 1.24e+01
Query:   CTTGTTCACTACGGCCGGC
Match:   |||||||||||||||||||
Subject: CTTGTTCACTACGGCCGGC


Generating Sequence 8 for BLAST...
GGGACGGGTTCTTATGTTTGAAGAA

****Alignment**** 1
```

```
Title: gi|2814963160|emb|OZ078335.2| Lampetra planeri genome assembly,
chromosome: 12
Accession: OZ078335
Length: 15546478
Max Score: 19.0
Bits: 38.1576
Identities: 19
Align_length: 19
Gaps: 0
%Ident: 100.00 %
Query Cover: 76 %
E value: 1.24e+01
Query:   GGTTCTTATGTTTGAAGAA
Match:   |||||||||||||||||||
Subject: GGTTCTTATGTTTGAAGAA


Generating Sequence 9 for BLAST...
CTGCACTCCGGCGCACAGGAGACAC

****Alignment**** 1
Title: gi|3061161874|emb|OZ311359.1| Aethes rutilana genome assembly,
chromosome: 4
Accession: OZ311359
Length: 16498050
Max Score: 19.0
Bits: 38.1576
Identities: 19
Align_length: 19
Gaps: 0
%Ident: 100.00 %
Query Cover: 76 %
E value: 1.24e+01
Query:   CTGCACTCCGGCGCACAGG
Match:   |||||||||||||||||||
Subject: CTGCACTCCGGCGCACAGG


Generating Sequence 10 for BLAST...
TAAACTGTACTAGGATCGGAGCAAT

****Alignment**** 1
Title: gi|2948600938|ref|XM_072899951.1| PREDICTED: Anoplolepis gracilipes
histone lysine acetyltransferase CREBBP (LOC140669806), transcript variant X18,
mRNA
Accession: XM_072899951
Length: 12961
Max Score: 18.0
Bits: 36.1753
Identities: 18
Align_length: 18
Gaps: 0
%Ident: 100.00 %
Query Cover: 72 %
E value: 4.89e+01
Query:   AACTGTACTAGGATCGGA
Match:   ||||||||||||||||||
Subject: AACTGTACTAGGATCGGA
```

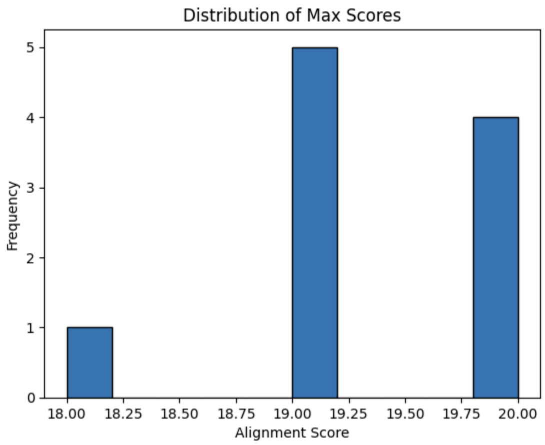
The typical length of the hits (the alignment length)



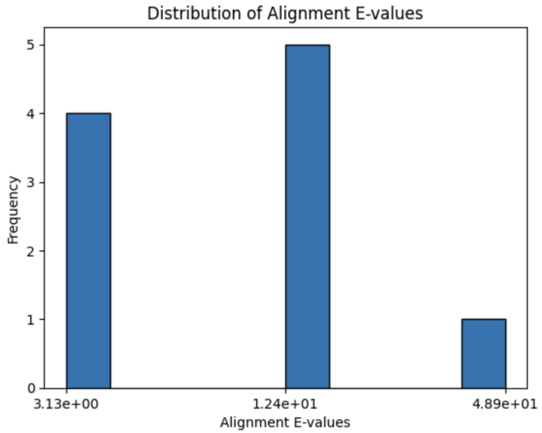
The typical % identity



The range of bit-scores ("max score")



The range of the E-values



There is no direct biological significance to any of these hits, of course, as the input sequences are made up. But given their short length of 25bp, we still have some decent hits!

Random protein sequences and BLASTP

For protein sequences, the typical length of the alignment is around 18–24 nucleotides. No gaps were found; all alignments were continuous. The range of E-values is 3–50.

Inspecting a few of the alignments in detail ("+" means similar sequences) we find that they look plausible at first glance because several show 100% identity over 19–24 bases. However, these sequences are far too short to be significant – the E-values indicate they occur by chance.

If we had used the default E-value cutoff of 10, we would still get a few hits with E ≈ 3, but the rest (E ≈ 12–50) would be excluded.

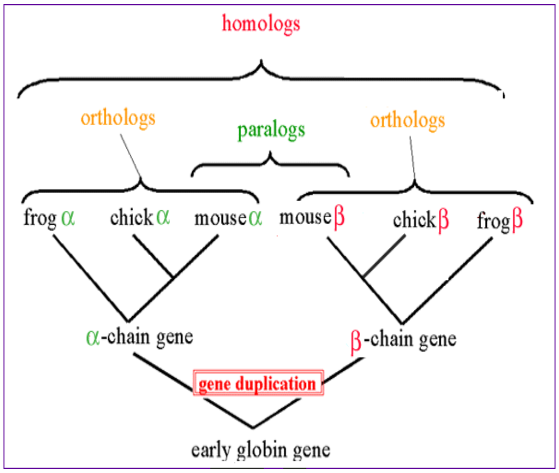
(Note that in contrast to protein BLAST (where the cutoff is usually 0.05), short random DNA sequences produce higher E-values (1–50) even when identical, because such matches are expected by chance in large nucleotide databases).

If we compare the result from BLAST'ing random DNA sequences to random peptide sequences, the risk of false positives is much higher for DNA (BLASTN) searches, because DNA uses only four bases, making short random matches common. Protein (BLASTP) searches, with 20 amino acids, have a much lower chance of random similarity. Therefore, random DNA sequences can appear to give “decent” E-values even when unrelated.

Part 3: Using BLAST to transfer functional information by finding homologs

One of the most common ways to use BLAST as a tool is in the situation where we have a sequence of unknown function, and want to find out which function it has. Since a large amount of sequence data has been gathered over the years, chances are that an evolutionarily related sequence with known function has already been identified. In general, such a related sequence is known as a "homolog".

Homo-, Ortho- and Paralogs:



A Homolog is a general term that describes a sequence that is related by any evolutionary means.

An Ortholog ("Ortho" = True) is a sequence that is "the same gene" in a different organism: The sequences shared a single common ancestor sequence, and have now diverged through speciation (e.g. the Alpha-globin gene in Human and Mouse).

A Paralog arises due to a gene duplication within a species. For example, Alpha- and Beta-globin are paralogs.

Notice that in both cases it's possible to transfer information, for example, about gene family / protein domains. We have already touched upon the comparison of (potentially) evolutionarily related sequences in the pairwise alignment exercise. However, this time we do not start with two sequences we assume are related, but instead, we start with a single sequence ("query sequence") which we will use to search the databases for homologs (we often informally speak of "BLAST hits", when discussing the sequences found).

```
LOCUS       CLONE12.DNA    609 BP DS-DNA             UPDATED    06/14/98

DEFINITION  UWGCG file capture

ACCESSION   -

KEYWORDS    -

SOURCE      -

COMMENT     Non-sequence data from original file:

BASE COUNT      174 A     116 C     162 G     157 T      0 OTHER

ORIGIN      ?

    clone12.dna Length: 609   Jun 13, 1998 - 03:39 PM   Check: 6014 ..

        1 AACGGGCACG GGACGCATGT AGCTGGAACA GTGGCAGCCG TAAATAATAA TGGTATCGGA

       61 GTTGCCGGGG TTGCAGGAGG AAACGGCTCT ACCAATAGTG GAGCAAGGTT AATGTCCACA

      121 CAAATTTTTA ATAGTGATGG GGATTATACA AATAGCGAAA CTCTTGTGTA CAGAGCCATT

      181 GTTTATGGTG CAGATAACGG AGCTGTGATC TCGCAAAATA GCTGGGGTAG TCAGTCTCTG
```

```
        241 ACTATTAAGG AGTTGCAGAA AGCTGCGATC GACTATTTCA TTGATTATGC AGGAATGGAC
        301 GAAACAGGAG AAATACAGAC AGGCCCTATG AGGGGAGGTA TATTTATAGC TGCCGCCGGA
        361 AACGATAACG TTTCCACTCC AAATATGCCT TCAGCTTATG AACGGGTTTT AGCTGTGGCC
        421 TCAATGGGAC CAGATTTTAC TAAGGCAAGC TATAGCACTT TTGGAACATG GACTGATATT
        481 ACTGCTCCTG GCGGAGATAT TGACAAATTT GATTTGTCAG AATACGGAGT TCTCAGCACT
        541 TATGCCGATA ATTATTATGC TTATGGAGAG GGAACATCCA TGGCTTGTCC ACATGTCGCC
        601 GGCGCCGCC
//
```

The sequence is a DNA fragment from an unknown non-cultivatable microorganism. It was cloned and sequenced directly from DNA extracted from a soil-sample, and it goes by the poetic name "CLONE12". It was amplified using degenerated PCR primers that target the middle ("core cloning") of the sequence of a group of known enzymes.

Now we will try to find the function of this sequence!

STEP 1 - cleaning up the sequence:

The sequence is (more or less) in GenBank format and the NCBI BLAST server expects the input to be in FASTA format, or to be "raw" unformatted sequence.

There are two solutions to this:

- Copy the sequence into a text-editor and manually create a FASTA file ("search and replace" and/or "rectangular selection" is useful for the reformatting).
  This is the most robust solution: it will always work. (Look at the Geany exercise for a reminder of how to do this).
- Hope the creators of the web-server you're using were kind enough to automatically remove non-DNA letters (paste in ONLY the DNA lines) - this turns out to be the case for both NCBI BLAST and VirtualRibosome, but it cannot be universally relied upon.

We will still convert the sequence to FASTA format manually. A good way to do this is by using LLMs, but a text editor will suffice here.

```
>Clone12
AACGGGCACGGGACGCATGTAGCTGGAACAGTGGCAGCCGTAAATAATAATGGTATCGGA
GTTGCCGGGGTTGCAGGAGGAAACGGCTCTACCAATAGTGGAGCAAGGTTAATGTCCACA
CAAATTTTTAATAGTGATGGGGATTATACAAATAGCGAAACTCTTGTGTACAGAGCCATT
GTTTATGGTGCAGATAACGGAGCTGTGATCTCGCAAAATAGCTGGGGTAGTCAGTCTCTG
ACTATTAAGGAGTTGCAGAAAGCTGCGATCGACTATTTCATTGATTATGCAGGAATGGAC
GAAACAGGAGAAATACAGACAGGCCCTATGAGGGGAGGTATATTTATAGCTGCCGCCGGA
AACGATAACGTTTCCACTCCAAATATGCCTTCAGCTTATGAACGGGTTTTAGCTGTGGCC
TCAATGGGACCAGATTTTACTAAGGCAAGCTATAGCACTTTTGGAACATGGACTGATATT
ACTGCTCCTGGCGGAGATATTGACAAATTTGATTTGTCAGAATACGGAGTTCTCAGCACT
TATGCCGATAATTATTATGCTTATGGAGAGGGAACATCCATGGCTTGTCCACATGTCGCC
GGCGCCGCC
```

STEP 2 - thinking about the task:

Based on the information given: is the sequence protein-coding?
Likely yes, but that's not guaranteed. It was amplified using “degenerated PCR primers that target the middle ("core cloning") of the sequence of a group of known enzymes.”

Those primers seem to target coding sequence (CDS) at the amino-acid level, so the amplicon is expected to fall inside a gene rather than UTR or intergenic DNA.

But: “environmental PCR can pick up paralogs, pseudogenes, or odd intron/exon structures (if eukaryotic)”.

Can we trust it will contain both a START and STOP codon?
Unlikely.

“Core cloning” almost always yields an internal fragment of a gene.

We should not expect an initiator ATG or a terminal stop codon. And just checking the sequence we can see an ATG immediately followed by a TAG, so we’re probably looking at an internal fragment.

Do we know if the sequence is sense or anti-sense?

No. Cloning preserves orientation, but without vector annotations you can’t assume which strand is coding. Consequently, you must translate and/or search both strands (six-frame translation and think which consequences the answers to these questions should have for your choice of methods and parameters.

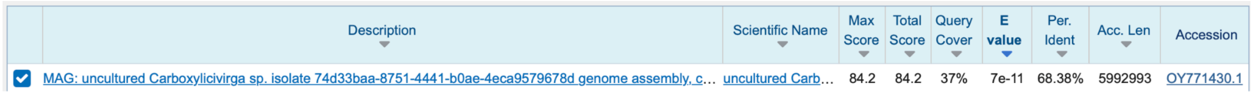
In summary, we treat the fragment as an internal piece of a coding gene of unknown orientation. This means we prefer protein-level searches that are more tolerant of incomplete or reversed coding regions.

STEP 3 - Performing the database search:

We considered hits with E-values below 1e-10 as significant. Hits above this threshold were ignored, as they likely represent random matches or distant/unreliable similarities

We used BLASTN to search the nucleotide sequence against the NR (non-redundant) database. This can detect very closely related DNA sequences, but it is less sensitive for evolutionary distant homologs.

We got a single significant hit, but the coverage is only 37% and percent identity is only 68%. So that suggests it's likely a related organism, especially considering the fact that it is also from an uncultured organism.

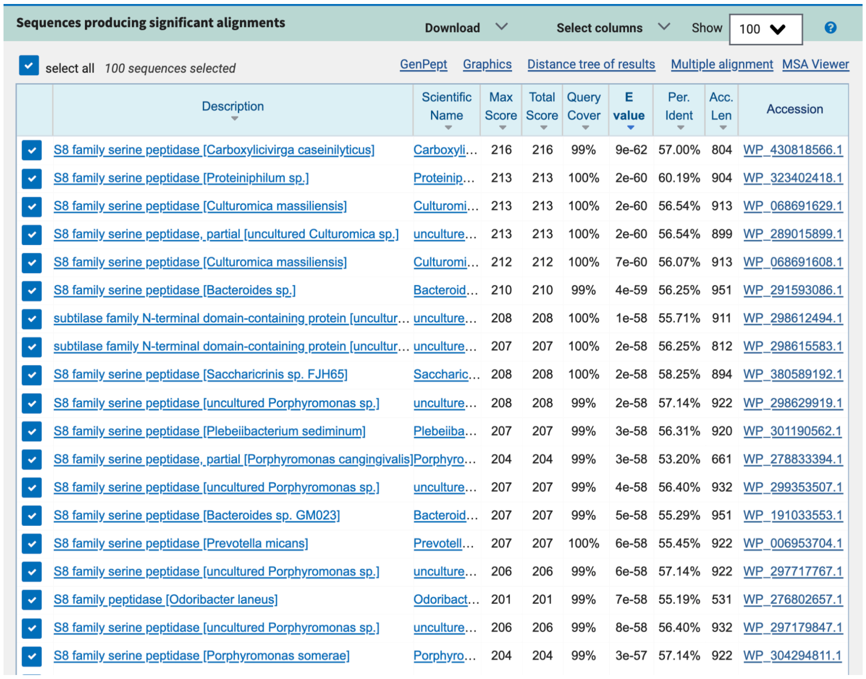


	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	MAG: uncultured Carboxylicivirga sp. isolate 74d33baa-8751-4441-b0ae-4eca9579678d genome assembly, c…	uncultured Carb…	84.2	84.2	37%	7e-11	68.38%	5992993	OY771430.1

Score	Expect	Identities	Gaps	Strand
84.2 bits(92)	7e-11	160/234(68%)	14/234(5%)	Plus/Minus

```
Query  2        ACGGGCACGGGACGCATGTAGCTGGAACAGTGGCAGCCGTAAATAATAATGGTATCGGAG  61

                ||| ||| || || ||||| ||||| ||  | | ||| ||||||||||||||||| || |

Sbjct  2245366  ACGAGCATGGAACACATGTGGCTGGTACGATAGGAGCAGTAAATAATAATGGTATAGGGG  2245307

Query  62       TTGCCGGGGTTGCAGGAGGAAACGGCTCTACCAATAGTGGAGCAAGGTTAATGTCCACAC  121

                ||   ||  | ||||| ||| | ||   ||| | |   ||||   ||||||||||    |

Sbjct  2245306  TTTGTGGAATAGCAGGTGGAGATGG---TACAACTCCCGGAGTTCGGTTAATGTCGTGCC  2245250

Query  122      AAATTTTTAATAGTGAT------GGGGATTATACAAATAGCGAAACTCTTGTGTACAG-A  174

                |  ||||| | ||||||      || ||| ||| || |   || | | |||    ||| |

Sbjct  2245249  AGGTTTTTGAAAGTGATGAAAACGGTGATGATATAAGTGCAGATAATTTTG----CAGCA  2245194

Query  175      GCCATTGTTTATGGTGCAGATAACGGAGCTGTGATCTCGCAAAATAGCTGGGGT  228

                || |||   |||||||| ||||| |||||  | || || |||||||| ||||||

Sbjct  2245193  GCTATTAAATATGGTGCGGATAATGGAGCCATAATTTCTCAAAATAGTTGGGGT  2245140
```

Then, we translated the DNA sequence into protein using Virtual Ribosome and ran BLASTP against the NR protein database, *because protein-level searches are more sensitive and can detect homologous enzymes across different species.*



Sequences producing significant alignments Download Select columns Show 100

☑ select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	S8 family serine peptidase [Carboxylicivirga caseinilyticus]	Carboxyli…	216	216	99%	9e-62	57.00%	804	WP_430818566.1
☑	S8 family serine peptidase [Proteiniphilum sp.]	Proteinip…	213	213	100%	2e-60	60.19%	904	WP_323402418.1
☑	S8 family serine peptidase [Culturomica massiliensis]	Culturomi…	213	213	100%	2e-60	56.54%	913	WP_068691629.1
☑	S8 family serine peptidase, partial [uncultured Culturomica sp.]	unculture…	213	213	100%	2e-60	56.54%	899	WP_289015899.1
☑	S8 family serine peptidase [Culturomica massiliensis]	Culturomi…	212	212	100%	7e-60	56.07%	913	WP_068691608.1
☑	S8 family serine peptidase [Bacteroides sp.]	Bacteroid…	210	210	99%	4e-59	56.25%	951	WP_291593086.1
☑	subtilase family N-terminal domain-containing protein [uncultur…	unculture…	208	208	100%	1e-58	55.71%	911	WP_298612494.1
☑	subtilase family N-terminal domain-containing protein [uncultur…	unculture…	207	207	100%	2e-58	56.25%	812	WP_298615583.1
☑	S8 family serine peptidase [Saccharicrinis sp. FJH65]	Saccharic…	208	208	100%	2e-58	58.25%	894	WP_380589192.1
☑	S8 family serine peptidase [uncultured Porphyromonas sp.]	unculture…	208	208	99%	2e-58	57.14%	922	WP_298629919.1
☑	S8 family serine peptidase [Plebeiibacterium sediminum]	Plebeiiba…	207	207	99%	3e-58	56.31%	920	WP_301190562.1
☑	S8 family serine peptidase, partial [Porphyromonas cangingivalis]	Porphyro…	204	204	99%	3e-58	53.20%	661	WP_278833394.1
☑	S8 family serine peptidase [uncultured Porphyromonas sp.]	unculture…	207	207	99%	4e-58	56.40%	932	WP_299353507.1
☑	S8 family serine peptidase [Bacteroides sp. GM023]	Bacteroid…	207	207	99%	5e-58	55.29%	951	WP_191033553.1
☑	S8 family serine peptidase [Prevotella micans]	Prevotell…	207	207	100%	6e-58	55.45%	922	WP_006953704.1
☑	S8 family serine peptidase [uncultured Porphyromonas sp.]	unculture…	206	206	99%	6e-58	57.14%	922	WP_297717767.1
☑	S8 family peptidase [Odoribacter laneus]	Odoribact…	201	201	99%	7e-58	55.19%	531	WP_276802657.1
☑	S8 family serine peptidase [uncultured Porphyromonas sp.]	unculture…	206	206	99%	8e-58	56.40%	932	WP_297179847.1
☑	S8 family serine peptidase [Porphyromonas somerae]	Porphyro…	204	204	99%	3e-57	57.14%	922	WP_304294811.1

Based on the BLAST results, CLONE12 is most likely a serine protease, belonging to the S8 family. BLASTP using the translated ORF gave several significant hits with high query coverage and identity.

Therefore, we have strong evidence that CLONE12 encodes a peptidase/protease enzyme.

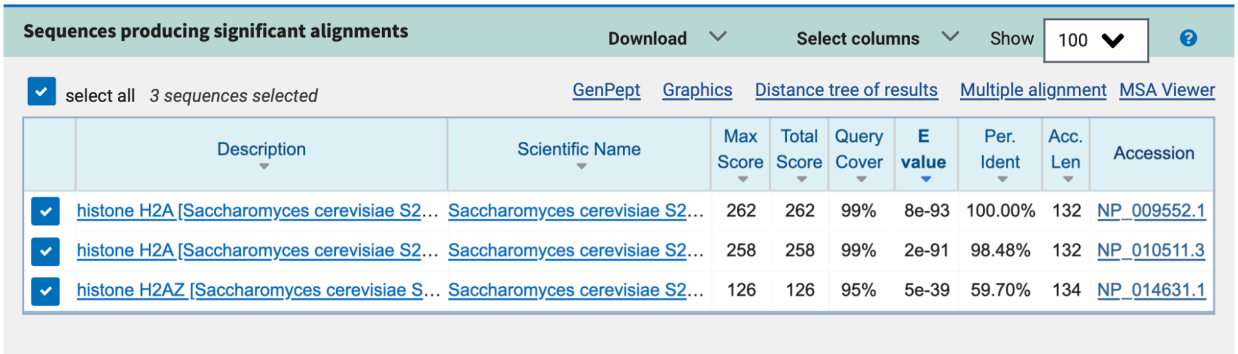
Part 4: BLAST'ing Genomes

We looked up the HTA2 gene in SGD (http://www.yeastgenome.org- use the search box at the top of the page)

HTA2 and HTA1 are paralogous genes encoding nearly identical histone H2A proteins. They are functionally redundant, and deletion of one can be compensated by the other.

How many high-confidence hits do we get?

three :



Sequences producing significant alignments Download Select columns Show 100

☑ select all *3 sequences selected* GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	histone H2A [Saccharomyces cerevisiae S2…	Saccharomyces cerevisiae S2…	262	262	99%	8e-93	100.00%	132	NP_009552.1
☑	histone H2A [Saccharomyces cerevisiae S2…	Saccharomyces cerevisiae S2…	258	258	99%	2e-91	98.48%	132	NP_010511.3
☑	histone H2AZ [Saccharomyces cerevisiae S…	Saccharomyces cerevisiae S2…	126	126	95%	5e-39	59.70%	134	NP_014631.1

Do the hits make sense, from what you have read about HTA2 at the SGD webpage?

The hits make sense, as HTA1 and HTA2 encode nearly identical histone *H2A proteins*.

Then, we searched the translated version of the human genome with the database set to "Reference proteins (refseq_protein)" and, of course, “Human” entered in the Organism field.

We found approximately 29-32 high-confidence hits with E-value better than 10^{-10}

First five:

1. NP_003503.1 — histone H2A type 1-C
2. NP_001035807.1 — histone H2A type 2-A
3. NP_003508.1 — histone H2A type 2-C
4. NP_003500.1 — histone H2A type 1
5. NP_542163.1 — histone H2A type 1-H

All the high-confidence hits were histone H2A proteins once again.

And that will be all for today. To recap, we used BLAST to identify homologous genes and their protein products. To explore these homologs further, the next logical step would be to collect the full-length sequences of the best hits, rather than just the partial regions found by BLAST. Then, with those sequences, we could perform pairwise alignments to compare specific differences or run a multiple sequence alignment to study their evolutionary relationships.

BLAST can also be used to build a dataset starting from a known “seed” sequence. Instead of trying to locate variants through keyword searches in GenBank, we can simply BLAST the known sequence, such as a reference insulin gene, and select the top hits as related variants for deeper analysis.