3 BLAST and FASTA

This lecture is based on the following papers, which are all recommended reading:

- D.J. Lipman and W.R. Pearson, Rapid and Sensitive Protein Similarity Searches. Science 227, 1435-1441 (1985).
- Pearson, W.R. and Lipman, D.J. Improved tools for biological sequence comparison. PNAS USA 85, 2444-2448 (1988).

宾州 据说在Dresden

- S.F. Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman. *Basic local alignment search tool*, J. Molecular Biology, 215:403-410 (1990).
- http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html, accessed April 2016.
- D. Gusfield, Algorithms on strings, trees and sequences, pg. 376–381, 1997.

3.1 BLAST

Scenario Queries 散列表

Pairwise alignment is used to detect homologies between different protein or DNA sequences, either as global or local alignments.

This can be solved using dynamic programming in time proportional to the product of the lengths of the two sequences being compared.

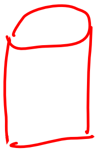
However, this is too slow for searching current databases and in practice algorithms are used that run much faster, at the expense of possibly missing some significant hits due to the heuristics employed.

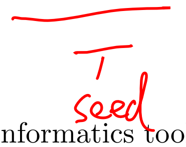
Such algorithms usually *seed and extend* approaches in which first small exact matches are found, which are then extended to obtain long inexact ones.

3.1.1 BLAST terminology

? Hash Tabel

NCBI -nr





BLAST, the Basic Local Alignment Search Tool, is perhaps the most widely used bioinformatics tool ever written. It is an alignment heuristic that determines local alignments between a *query* and a *database*.

启发式的

Let q be the query and d the database. A *segment* is simply a substring s of q or d.

A *segment-pair* (s,t) consists of two segments, one in q and one d, of the same length.

/

```
       x x x x x V A L L A R  x x x
  ...x x x x x x P A M M A R  x x x x x x...
```

We think of s and t as being aligned without gaps and *score* this alignment using a substitution score matrix, such as BLOSUM62.

The alignment score for (s,t) is denoted by $\sigma(s,t)$.

Definition 3.1.1 (HSP) *Let C be a given minimum score threshold. A segment pair (s,t) is called a* high-scoring segment pair (HSP)*, if it is* locally maximal *and $\sigma(s,t) \geq C$.*

Seed Extend

Locally maximal is defined by the X-drop algorithm, described below.

A *word* is simply a short substring. 若找到了Seed就延伸

....←LLLLL⇒DIT P....

....←LLLLLL⇒DIUR....

> The goal of BLAST is to compute all HSPs between two sequences (or a query and a database), for a given minimum score threshold C.

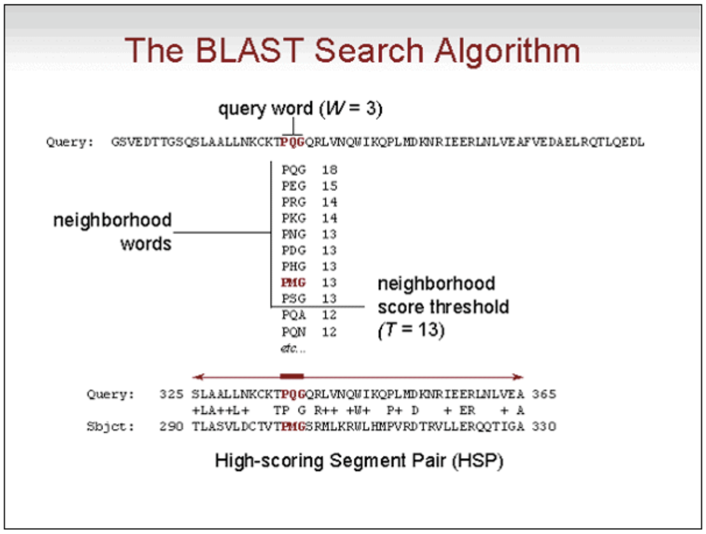
3.1.2 The BLAST algorithm

The BLAST algorithm has three parameters: The *word size* W, *word similarity threshold* T and *minimum match score* C.

For *protein* sequences, BLAST operates as follows:

1. The list ℓ of all words of length W that have similarity $\geq T$ to some word in the query sequence q is generated.
2. The database sequence d is scanned to detect each hit t of a word s from the list ℓ.
3. Each such pair (s, t) (called a *seed*) is extended in either direction until the running score drops X below the best score seen so far. (This is called the X*-drop algorithm*.)
4. For each seed match, the best extension is reported, if it has score $\geq C$.

In practice: W is $2 - 4$ for proteins.


The BLAST Search Algorithm

query word (*W* = 3)
neighborhood words
neighborhood score threshold (*T* = 13)
High-scoring Segment Pair (HSP)

```
Query:  GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL

PQG  18
PEG  15
PRG  14
PKG  14
PNG  13
PDG  13
PHG  13
PMG  13
PSG  13
PQA  12
PQN  12
etc...

Query:  325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
            +LA++L+   TP G R++ +W+  P+ D   + ER   + A
Sbjct:  290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
```

(https://www.cs.umd.edu/class/fall2011/cmsc858s/BLAST.pdf, accessed April 2016)

With care, the list of all words of length W that have similarity $\geq T$ to some word in the query sequence q can be produced in time proportional to the number of words in the list.

All seeds for a query are placed in a "keyword tree" and then, for each word in the tree, all exact locations of the word in the database d are detected in time proportional to the length of d.

The original version of BLAST did not allow "indels" (insertions or deletions), making hit extension very fast.

Note that the use of seeds of length W and the X-drop algorithm speed up the algorithm, but imply that BLAST is a heuristic that is not guaranteed to find all optimal local alignments.

For *DNA* sequences, BLAST operates as follows:

- The list L of all words of length W in the query sequence q is generated.
- The database d is scanned for all hits of words in L. Blast uses a two-bit encoding for DNA. This saves space and also search time, as four bases are encoded per byte.

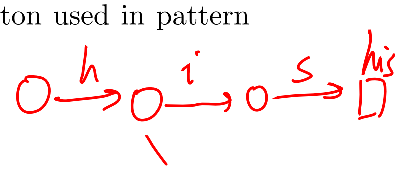
In practice, W is around 12 for DNA.

3.1.3 Keyword tree

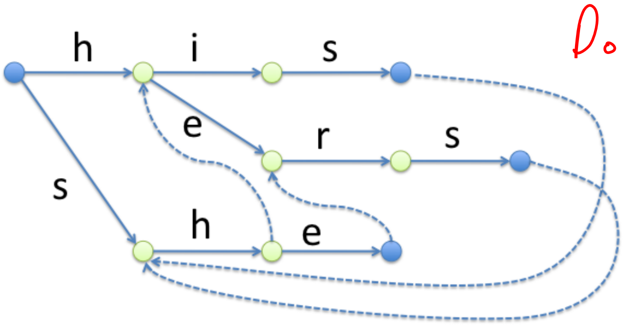
Assume that we want find occurrences of the words "`his`", "`hers`" and "`she`" in the text "`shishers`".

We can use a *keyword tree* to perform the search, which is a finite state automaton used in pattern matching:

Dot →



这不是唯一的方法



As we read the text from left to right, we move through the graph following arrows with matching labels, if we can, or jumping back to the root, if not. Whenever we complete a word in the graph, we print out its text location. The dashed arrows are followed to find overlapping matches.

Occurrences: "`his`" at position 2, "`hers`" at position 5, "`she`" at position 4.

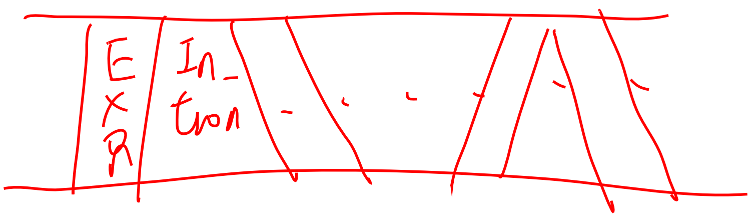
3.1.4 The BLAST family

There are a number of different variants of the BLAST program:

- BLASTN: compares a DNA query sequence to a DNA sequence database,
- BLASTP: compares a protein query sequence to a protein sequence database,
- TBLASTN: compares a protein query sequence to a DNA sequence database (6 frames translation),
- BLASTX: compares a DNA query sequence (6 frames translation) to a protein sequence database, and
- TBLASTX: compares a DNA query sequence (6 frames translation) to a DNA sequence database (6 frames translation).

Input:

```
>000078_0077_0290 length=107
GGGTTGAGGCTGTCGGTAAACACCTGGGCGCGGGTGATATGGCCTTTTCAACGTCGAAAT
GCAGTTCCACGCGCCCAGGTAAAGCGTTCATCCAGCAGATGCGAGAA
```



Example of BLASTN result:

```
BLASTN 2.2.13 [Nov-27-2005]

Query= 000078_0077_0290 length=107 (107 letters)

Database: EcoliK12.fna 1 sequences; 4,639,675 total letters

>gi|49175990|ref|NC_000913.2| Escherichia coli K12, complete genome
          Length = 4639675

 Score =  172 bits (87), Expect = 4e-44
 Identities = 106/110 (96%), Gaps = 3/110 (2%)
 Strand = Plus / Plus


Query: 1       gggttgaggctgtcggtaaacacctgggcgcgggtgatatggcc-ttttcaacgtcgaaa 59
               |||||||||||||||||||||||||||||||||||||||||||| ||||||||||||||| 
Sbjct: 4621278 gggttgaggctgtcggtaaacacctgggcgcgggtgatatggcctttttcaacgtcgaaa 4621337


Query: 60      tgcagttccacgc--gcccaggtaaagcgttcatccagcagatgcgagaa 107
               |||||||||||||   ||||||||||||||||||||||||||||||||||
Sbjct: 4621338 tgcagttccacgccgccccaggtaaagcgttcatccagcagatgcgagaa 4621387
```

Example of BLASTX result:

```
BLASTX 2.2.13 [Nov-27-2005]

Query= 000078_0077_0290 length=107 (107 letters)

Database: nr 3,044,223 sequences; 1,047,289,308 total letters

>gi|16132203|ref|NP_418803.1| lipoate-protein ligase A [Escherichia coli K12]
 lipoate-protein ligase A [Escherichia coli K12]
 yjjF [Escherichia coli] Lipoate-protein ligase A lipoate-protein ligase A
          Length = 338

 Score = 50.4 bits (119), Expect = 2e-05
 Identities = 27/36 (75%), Positives = 27/36 (75%), Gaps = 1/36 (2%)
 Frame = -1

Query: 107 FSHLLDERFTWARGTA-FRR*KGHITRAQVFTDSLN 3
           FSHLLDERFTW      F   KGHITRAQVFTDSLN
Sbjct: 252 FSHLLDERFTWGGVELHFDVEKGHITRAQVFTDSLN 287
```

3.1.5 The BLAST E-value

Question: Suppose we have computed an HSP (s, t) with score S, how significant is this match?

In the following, we assume that the length m and n of the query and database are sufficiently large.

Definition 3.1.2 (E-value) *The number of HSPs with score $\geq S$ that one can expect to see by random chance is known as the E-*value*, which equals:*

$$E = Kmne^{-\lambda S}.$$

The E-value depend on two parameters, K and λ. These are based on the background probabilities of the symbols and on the employed scoring matrix. Essentially, they are scaling-factors for the search space and for the scoring scheme, respectively. (BLAST uses a built-in table of experimentally determined values of K and λ .)

BLAST 有一个 Hard coded Table

3.1.6 The BLAST bit score

In addition to the E-value, BLAST also reports a so-called *bit score*.

For a given HSP (s, t) we transform the *raw* score $S = \sigma(s, t)$ into a *bit score* thus:

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

Such bit scores make it easier to compare between different BLAST searches, because it hides the two parameters K and λ. Given the bit score of a match, one can easily compute the E-value that would arise for given query and database lengths. The E-value is obtain from a bit score S' as follows:

$$\text{Interpretation of} \rightarrow \boxed{E} = mn2^{-S'}.$$

To see this, first solve for S in the equation for S' above and then plug the result into the original E-value equation.

3.1.7 The BLAST P-value

The number of HSPs (s, t) with $\sigma(s, t) \geq S$ that one obtains when comparing against a database of *random sequence* is given by a Poisson distribution. So, the probability of finding exactly k HSPs with a score $\geq S$ is given by

$$P(k) = e^{-E} \times \frac{E^k}{k!},$$

泊松分布

where E is the E-value for S. [1]

Definition 3.1.3 (P-value) *The probability of finding at least one HSP with a score* $\geq S$ *"by chance" is*

$$P(k \geq 1) = 1 - P(0) = 1 - e^{-E},$$

called the P-value.

BLAST reports E-values rather than P-values because it is easier, for example, to interpret the difference between an E-value of 5 and 10, than to interpret the difference between a P-value of 0.993 and 0.99995.

3.2 FASTA

FASTA[2] (pronounced fast-ay) is a heuristic for finding significant matches between a query string q and a database string d.

FASTA's general strategy is to quickly find the most significant diagonals in the dynamic programming matrix.

The performance of the algorithm is influenced by a *word-size* parameter k, usually 6 for DNA and 2 for amino acids.

The first step of the algorithm is to determine all exact matches of length k between the two sequences, called *hot-spots*.

To find these exact matches quickly, a hash table is built that consists of all words of length k that are contained in the query sequence. Exact matches are then found by look-up of each word of length k contained in the database.

A *hot-spot* is given by (i, j), where i and j are the *locations* (i.e., start positions) of an exact match of length k.

Any such hot-spot (i, j) lies on the diagonal $(i - j)$ of the dynamic programming matrix.

Using this scheme, the main diagonal has number 0, whereas diagonals above the main one have positive numbers, the ones below negative.

A *diagonal run* is a set of hot-spots that lie in a consecutive sequence on the same diagonal. It corresponds to a gapless local alignment.

A score is assigned to each diagonal run. This is done by giving a positive score to each match (using e.g. the PAM250 match score matrix in the case of proteins) and a negative score for gaps in the run.

The algorithm then locates the ten best diagonal runs.

Each of the ten diagonal runs is re-scored using the match score matrix and the best-scoring sub-alignment of each is extracted.

The next step is to combine high scoring sub-alignments into a single larger alignment, allowing the introduction of gaps into the alignment.

Finally, a *banded* Smith-Waterman dynamic program is used to produce an optimal local alignment along the best matched regions.

In this way, FASTA determines a highest scoring region, not all high scoring alignments between two sequences. Hence, FASTA may miss instances of repeats or multiple domains shared by two proteins.

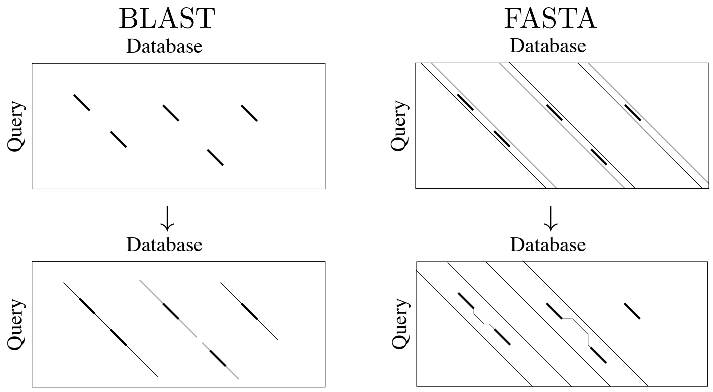
[1] Recall **Poisson distribution:** The probability that exactly k events occurs in a fixed time period, given that the expected number of events is λ in that time period, is given by:

$$P(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

Example: Assume that a call center gets 10 calls per hour, on average. The probability that it will get 100 calls in a given hour is $P(100, 10)$.

[2] Lipman, DJ; Pearson, WR (1985). Rapid and sensitive protein similarity searches. Science 227 (4693): 1435-41

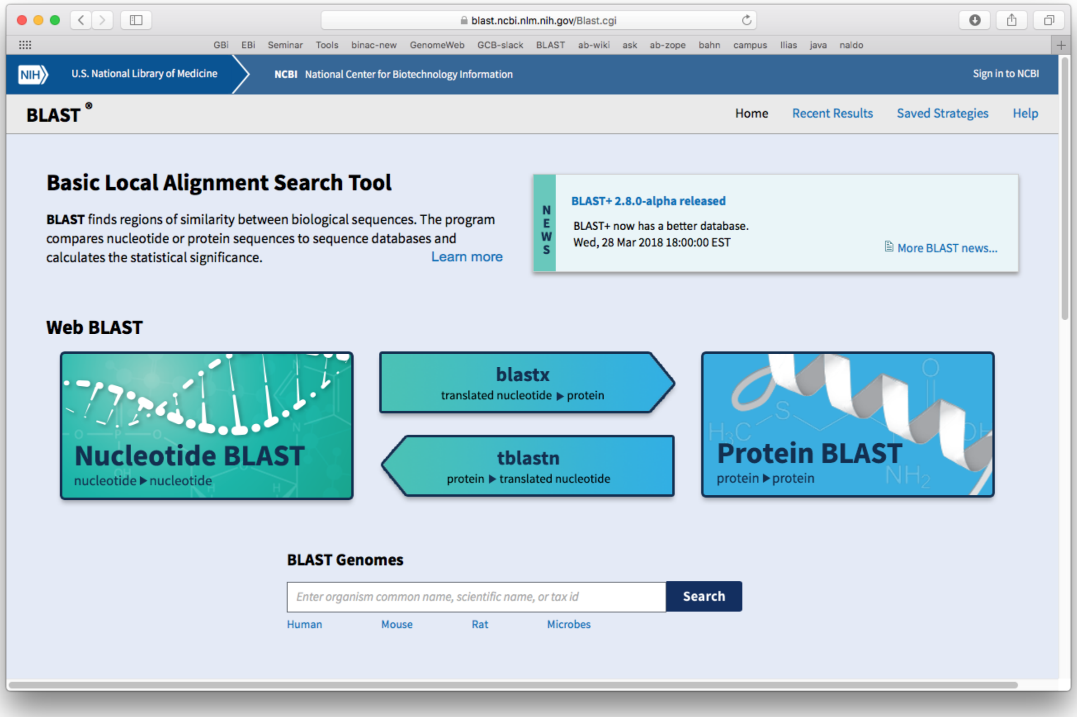
3.3 BLAST and FASTA



(a) (b)

(a) In BLAST, individual seeds are found and then extended without indels. (b) In FASTA, individual seeds contained in the same diagonal are merged and the resulting segments are then connected using a banded Smith-Waterman alignment.

3.4 BLAST as a web service


blast.ncbi.nlm.nih.gov/Blast.cgi
GBi EBi Seminar Tools binac-new GenomeWeb GCB-slack BLAST ab-wiki ask ab-zope bahn campus Ilias java naldo
NIH U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Sign in to NCBI
BLAST®
Home Recent Results Saved Strategies Help
NEWS
Learn more
More BLAST news...
Nucleotide BLAST
nucleotide ▶ nucleotide
blastx
translated nucleotide ▶ protein
tblastn
protein ▶ translated nucleotide
Protein BLAST
protein ▶ protein
Enter organism common name, scientific name, or tax id
Search
Human Mouse Rat Microbes

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.

BLAST+ 2.8.0-alpha released

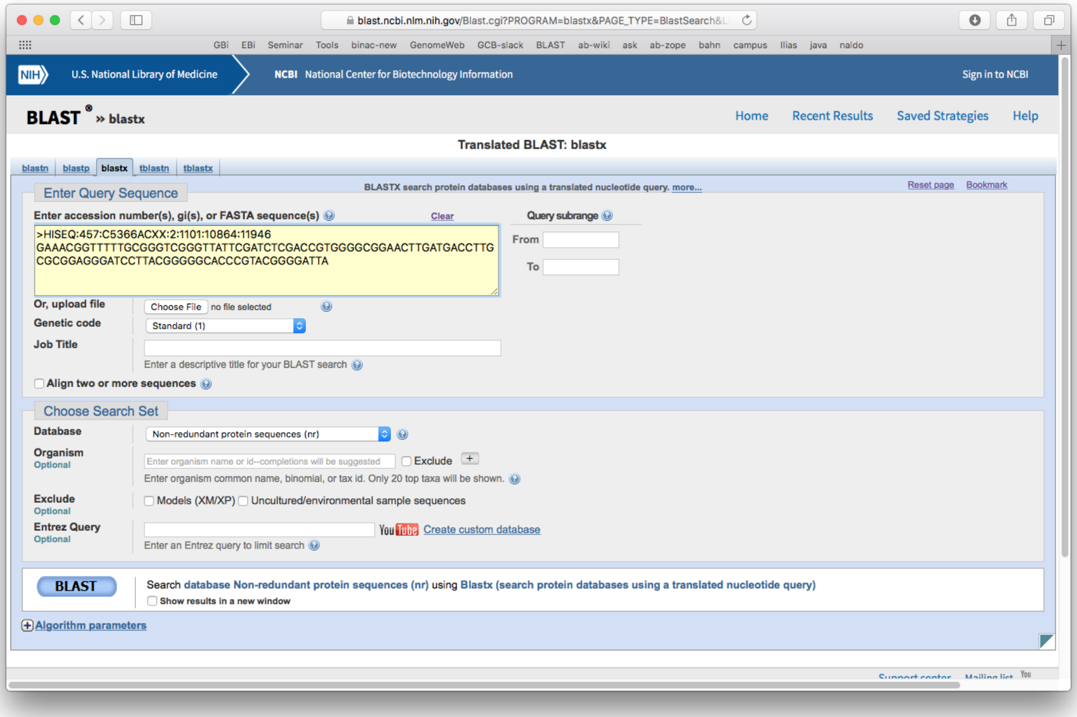
BLAST+ now has a better database.
Wed, 28 Mar 2018 18:00:00 EST

Web BLAST

BLAST Genomes

http://www.ncbi.nlm.nih.gov/BLAST/

3.4.1 BLAST run example


blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastx&PAGE_TYPE=BlastSearch&L
GBi EBi Seminar Tools binac-new GenomeWeb GCB-slack BLAST ab-wiki ask ab-zope bahn campus Ilias java naldo
NIH U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Sign in to NCBI
Home
Recent Results
Saved Strategies
Help
blastn
blastp
blastx
tblastn
tblastx
Reset page
Bookmark
Clear
Choose File
BLAST
Algorithm parameters

BLAST® » blastx

Translated BLAST: blastx

BLASTX search protein databases using a translated nucleotide query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
>HISEQ:457:C5366ACXX:2:1101:10864:11946
GAAACGGTTTTTGCGGGTCGGGTTATTCGATCTCGACCGTGGGGCGGAACTTGATGACCTTG
CGCGGAGGGATCCTTACGGGGGCACCCGTACGGGGATTA
```

Query subrange

From

To

Or, upload file: no file selected

Genetic code: Standard (1)

Job Title:

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism *Optional*: Enter organism name or id--completions will be suggested ☐ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

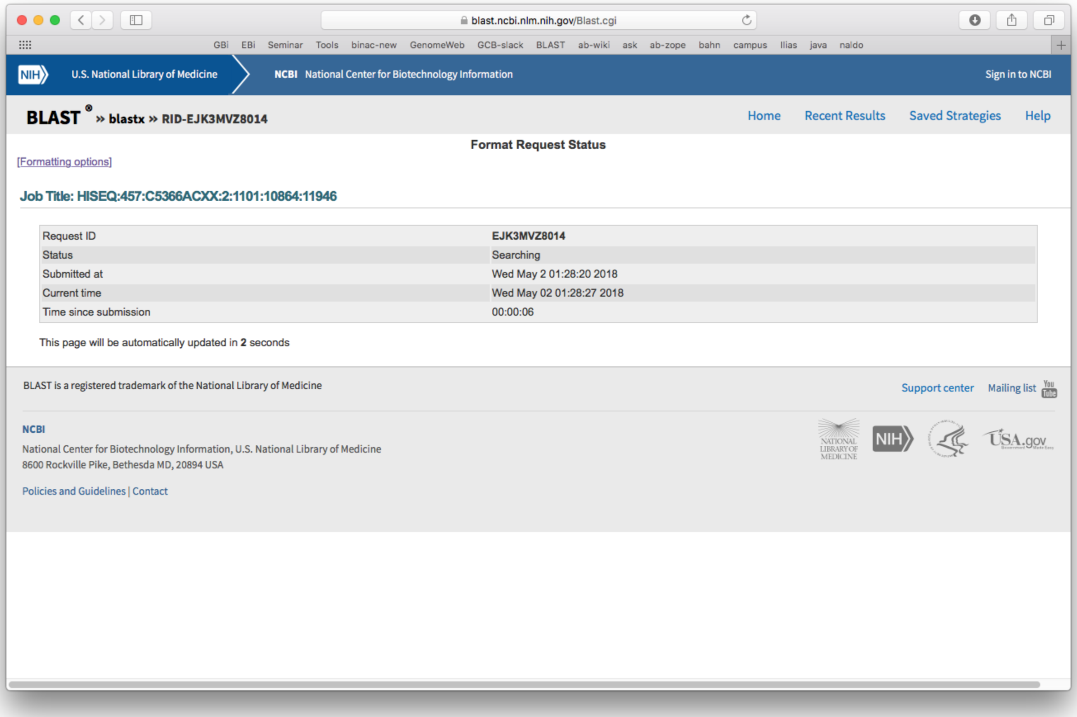
Exclude *Optional*: ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query *Optional*: Create custom database

Enter an Entrez query to limit search

Search database Non-redundant protein sequences (nr) using Blastx (search protein databases using a translated nucleotide query)

☐ Show results in a new window


blast.ncbi.nlm.nih.gov/Blast.cgi
GBi EBi Seminar Tools binac-new GenomeWeb GCB-slack BLAST ab-wiki ask ab-zope bahn campus Ilias java naldo
NIH U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Sign in to NCBI
Home
Recent Results
Saved Strategies
Help
[Formatting options]
Support center
Mailing list
Policies and Guidelines | Contact

BLAST® » blastx » RID-EJK3MVZ8014

Format Request Status

Job Title: HISEQ:457:C5366ACXX:2:1101:10864:11946

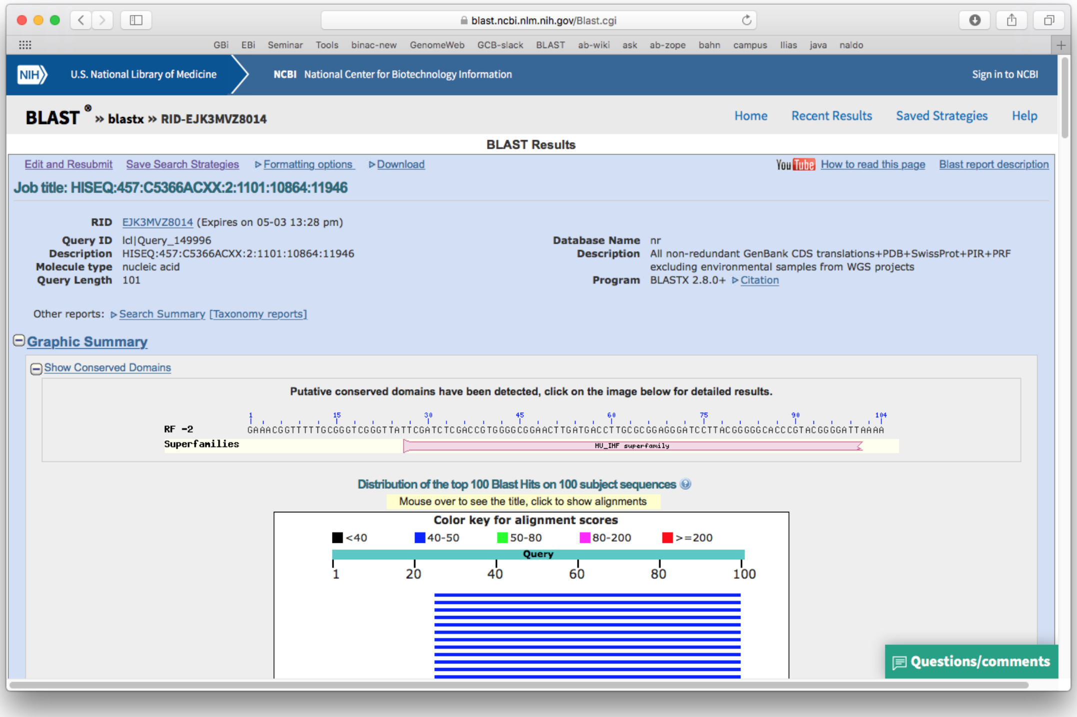
Request ID	EJK3MVZ8014
Status	Searching
Submitted at	Wed May 2 01:28:20 2018
Current time	Wed May 02 01:28:27 2018
Time since submission	00:00:06

This page will be automatically updated in **2** seconds

BLAST is a registered trademark of the National Library of Medicine

NCBI

National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA


blast.ncbi.nlm.nih.gov/Blast.cgi
GBi EBi Seminar Tools binac-new GenomeWeb GCB-slack BLAST ab-wiki ask ab-zope bahn campus Ilias java naldo
NIH U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Sign in to NCBI
Home
Recent Results
Saved Strategies
Help
Edit and Resubmit
Save Search Strategies
▷ Formatting options
▷ Download
YouTube How to read this page
Blast report description
RF -2
Superfamilies
HU_IHF superfamily
1
15
30
45
60
75
90
104
Query
1
20
40
60
80
100
Questions/comments

BLAST® » blastx » RID-EJK3MVZ8014

BLAST Results

Job title: HISEQ:457:C5366ACXX:2:1101:10864:11946

RID	EJK3MVZ8014 (Expires on 05-03 13:28 pm)		
Query ID	lcl|Query_149996	Database Name	nr
Description	HISEQ:457:C5366ACXX:2:1101:10864:11946	Description	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule type	nucleic acid		
Query Length	101	Program	BLASTX 2.8.0+ ▷ Citation

Other reports: ▷ Search Summary [Taxonomy reports]

Graphic Summary

Show Conserved Domains

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

```
GAAACGGTTTTTGCGGGTCGGGTTATTCGATCTCGACCGTGGGGCGGAACTTGATGACCTTGCGCGGAGGGATCCTTACGGGGCACCCGTACGGGGATTAAAA
```

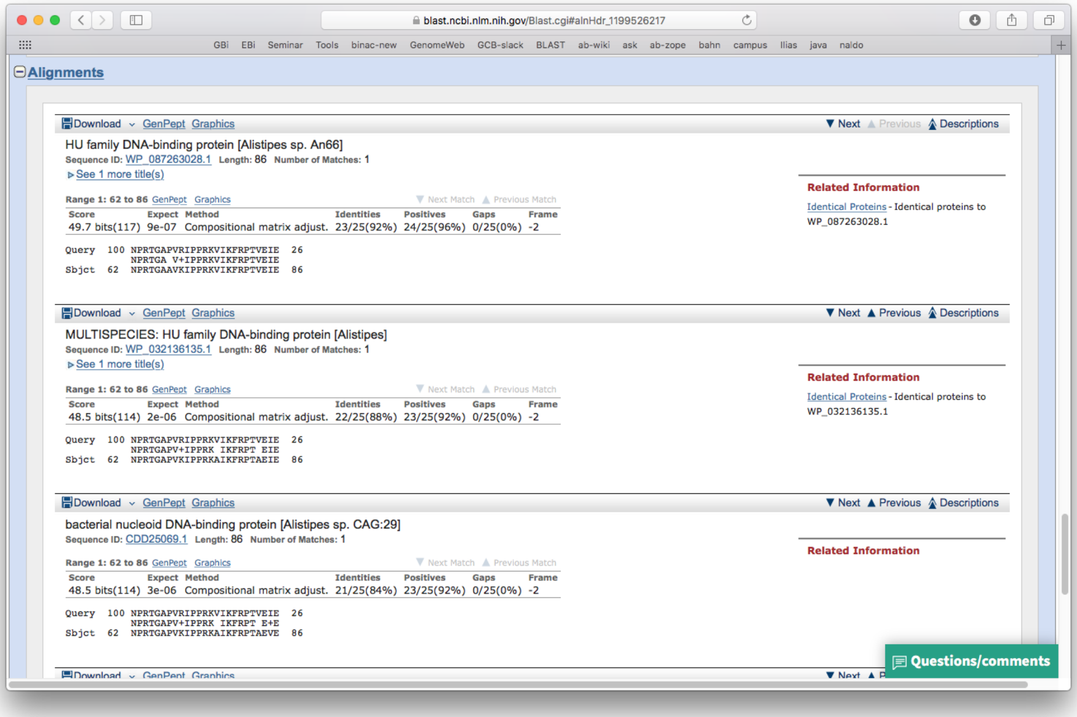
Distribution of the top 100 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments

Color key for alignment scores

■ <40 ■ 40-50 ■ 50-80 ■ 80-200 ■ >=200

3.4.2 BLAST run output


blast.ncbi.nlm.nih.gov/Blast.cgi#alnHdr_1199526217
GBi EBi Seminar Tools binac-new GenomeWeb GCB-slack BLAST ab-wiki ask ab-zope bahn campus Ilias java naldo
Download ⌄ GenPept Graphics
▼ Next ▲ Previous ▲ Descriptions
▼ Next Match ▲ Previous Match
Questions/comments

Alignments

HU family DNA-binding protein [Alistipes sp. An66]
Sequence ID: WP_087263028.1 Length: 86 Number of Matches: 1
▷ See 1 more title(s)

Related Information
Identical Proteins - Identical proteins to WP_087263028.1

Range 1: 62 to 86 GenPept Graphics

```
 Score          Expect  Method                        Identities   Positives    Gaps      Frame
 49.7 bits(117) 9e-07   Compositional matrix adjust.  23/25(92%)   24/25(96%)   0/25(0%)  -2

Query  100  NPRTGAPVRIPPRKVIKFRPTVEIE  26
            NPRTGA V+IPPRKVIKFRPTVEIE
Sbjct  62   NPRTGAAVKIPPRKVIKFRPTVEIE  86
```

MULTISPECIES: HU family DNA-binding protein [Alistipes]
Sequence ID: WP_032136135.1 Length: 86 Number of Matches: 1
▷ See 1 more title(s)

Related Information
Identical Proteins - Identical proteins to WP_032136135.1

Range 1: 62 to 86 GenPept Graphics

```
 Score          Expect  Method                        Identities   Positives    Gaps      Frame
 48.5 bits(114) 2e-06   Compositional matrix adjust.  22/25(88%)   23/25(92%)   0/25(0%)  -2

Query  100  NPRTGAPVRIPPRKVIKFRPTVEIE  26
            NPRTGAPV+IPPRK IKFRPT EIE
Sbjct  62   NPRTGAPVKIPPRKAIKFRPTAEIE  86
```

bacterial nucleoid DNA-binding protein [Alistipes sp. CAG:29]
Sequence ID: CDD25069.1 Length: 86 Number of Matches: 1

Related Information

Range 1: 62 to 86 GenPept Graphics

```
 Score          Expect  Method                        Identities   Positives    Gaps      Frame
 48.5 bits(114) 3e-06   Compositional matrix adjust.  21/25(84%)   23/25(92%)   0/25(0%)  -2

Query  100  NPRTGAPVRIPPRKVIKFRPTVEIE  26
            NPRTGAPV+IPPRK IKFRPT E+E
Sbjct  62   NPRTGAPVKIPPRKAIKFRPTAEVE  86
```