Sequence Alignment

GBIO0002

Archana Bhardwaj
University of Liege

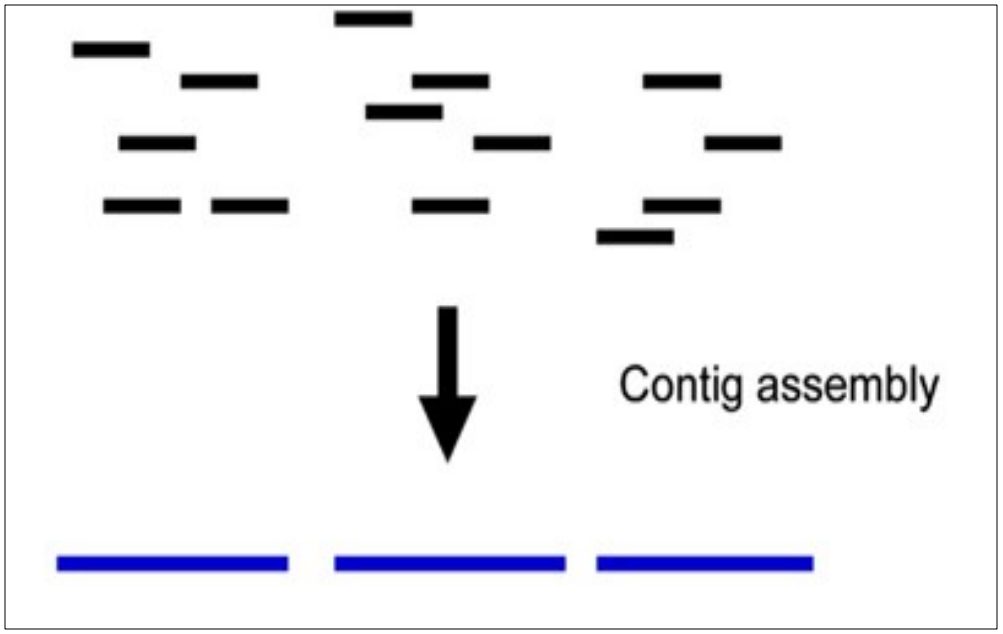
What is Sequence Alignment ?

A sequence alignment is a way of arranging the sequences of DNA , RNA, or protein to identify regions of similarity.

Comparable ?

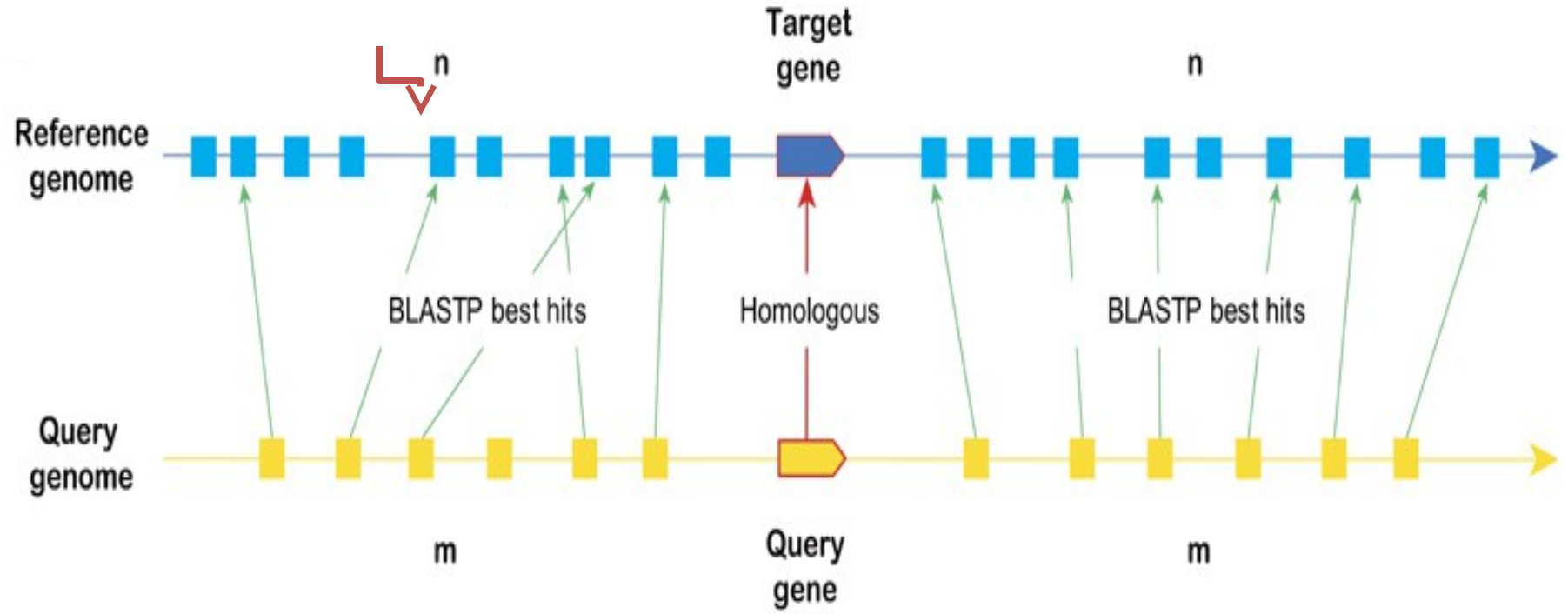
Sequence Alignment :Uses (1)

- **Sequence Assembly : Genome sequence are assembled by using the sequence alignment methods to find the overlap between many short pieces of DNA .**



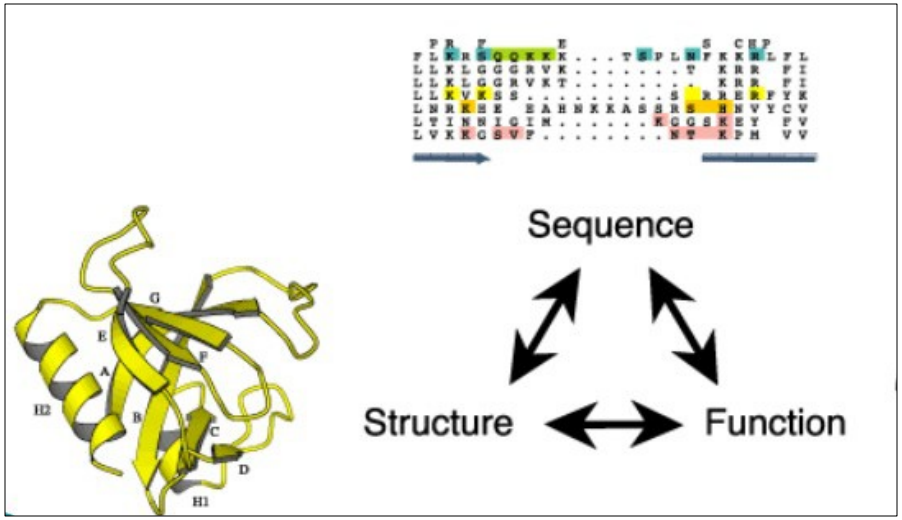
Sequence Alignment :Uses (2)

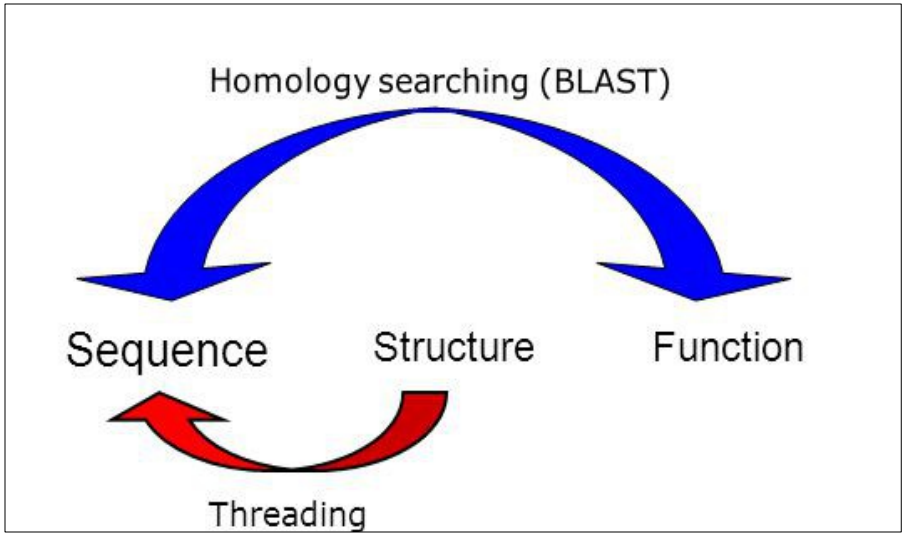
- **Gene Finding : Sequence similarity could help us to find the gene prediction just by doing comparison against the other set of sequences.**



Sequence Alignment :Uses (3)

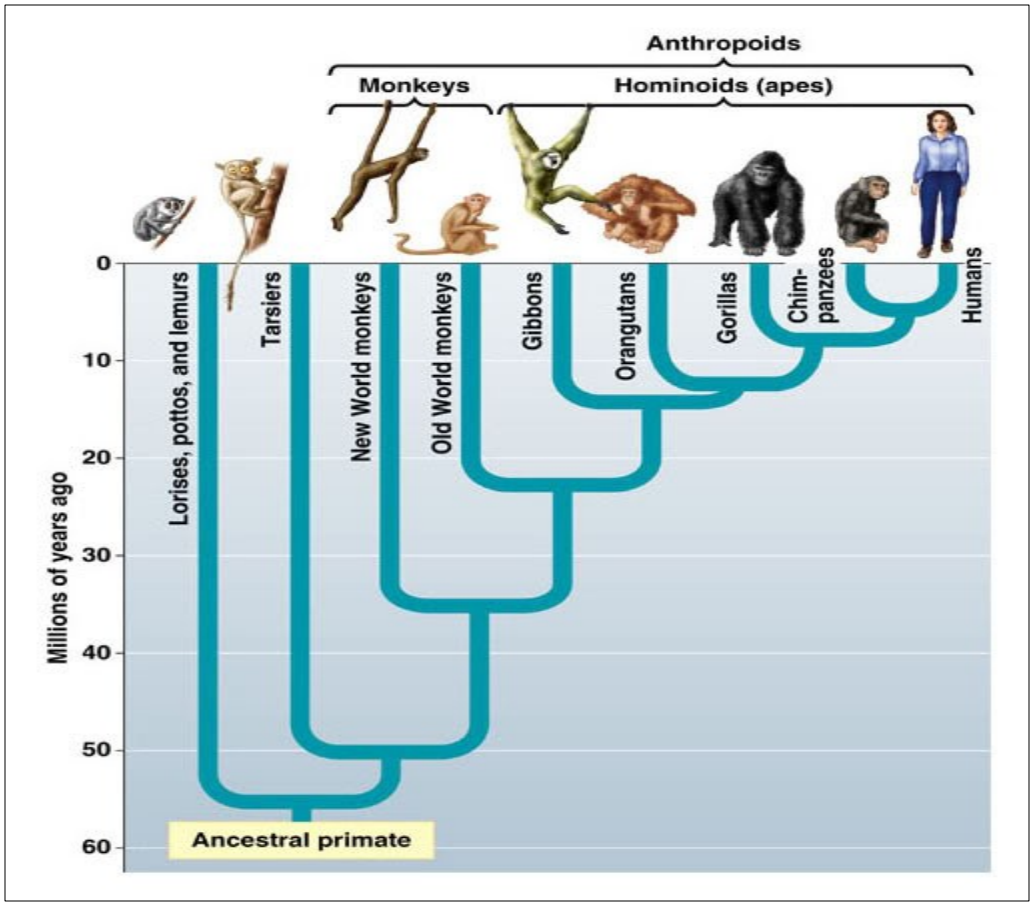
- **Function prediction : Function of any unknown sequence could be predicted by comparing with other known sequence .**





Sequence Alignment :Uses (4)

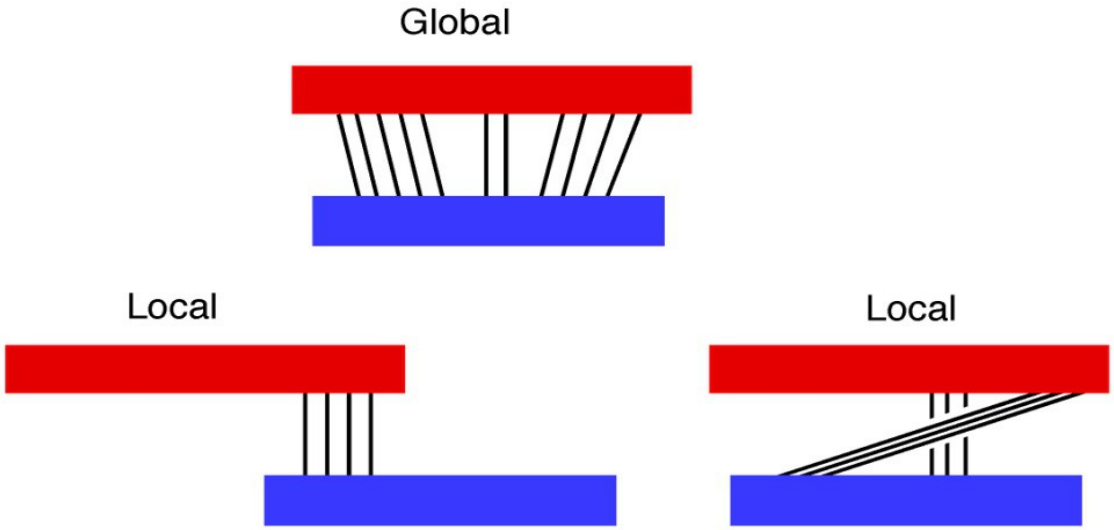
- **Sequence Divergence : Amount of sequence similarity (10%, 20%,30% ...sometimes 90 %) between sequences tell us how closely they are related**



Types of Alignments

- **Global : This attempt to align every residue in every sequence.**

- **Local: It is more useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context.**



Local Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| ||||||||||||||||
```

Query Sequence

```
5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Global Alignment

Target Sequence

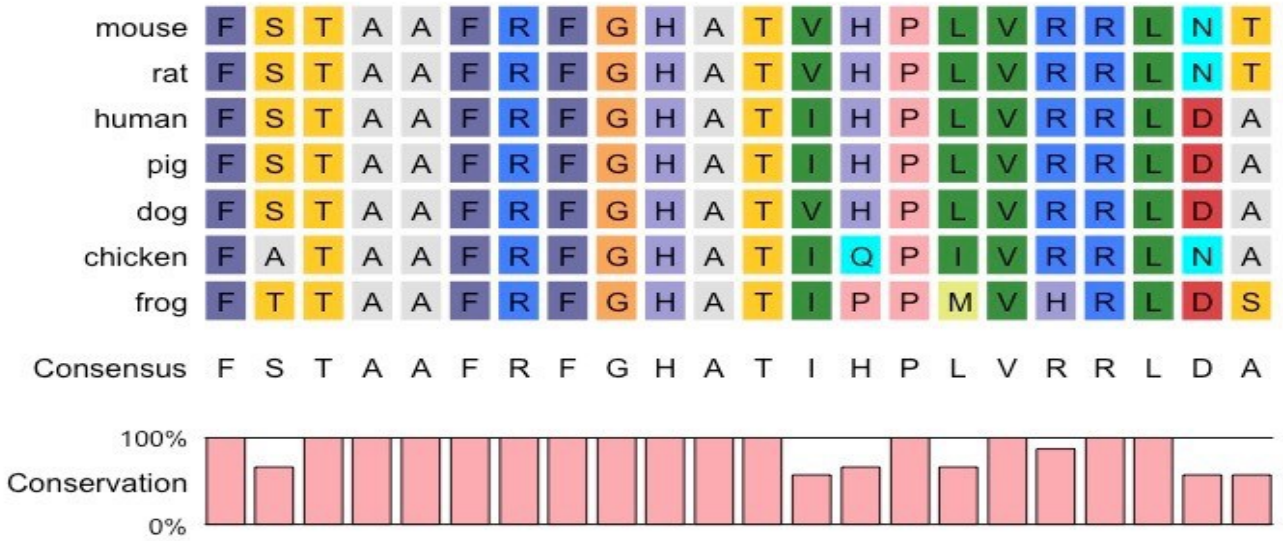
```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   ||||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

Query Sequence

Types of Alignments: Based on number of sequences

- **Pair wise Sequence Alignment** : This alignments can only be used between two sequences at a time.

- **Multiple Sequence Alignment** : This alignments can only be used between more than two sequences at a time.



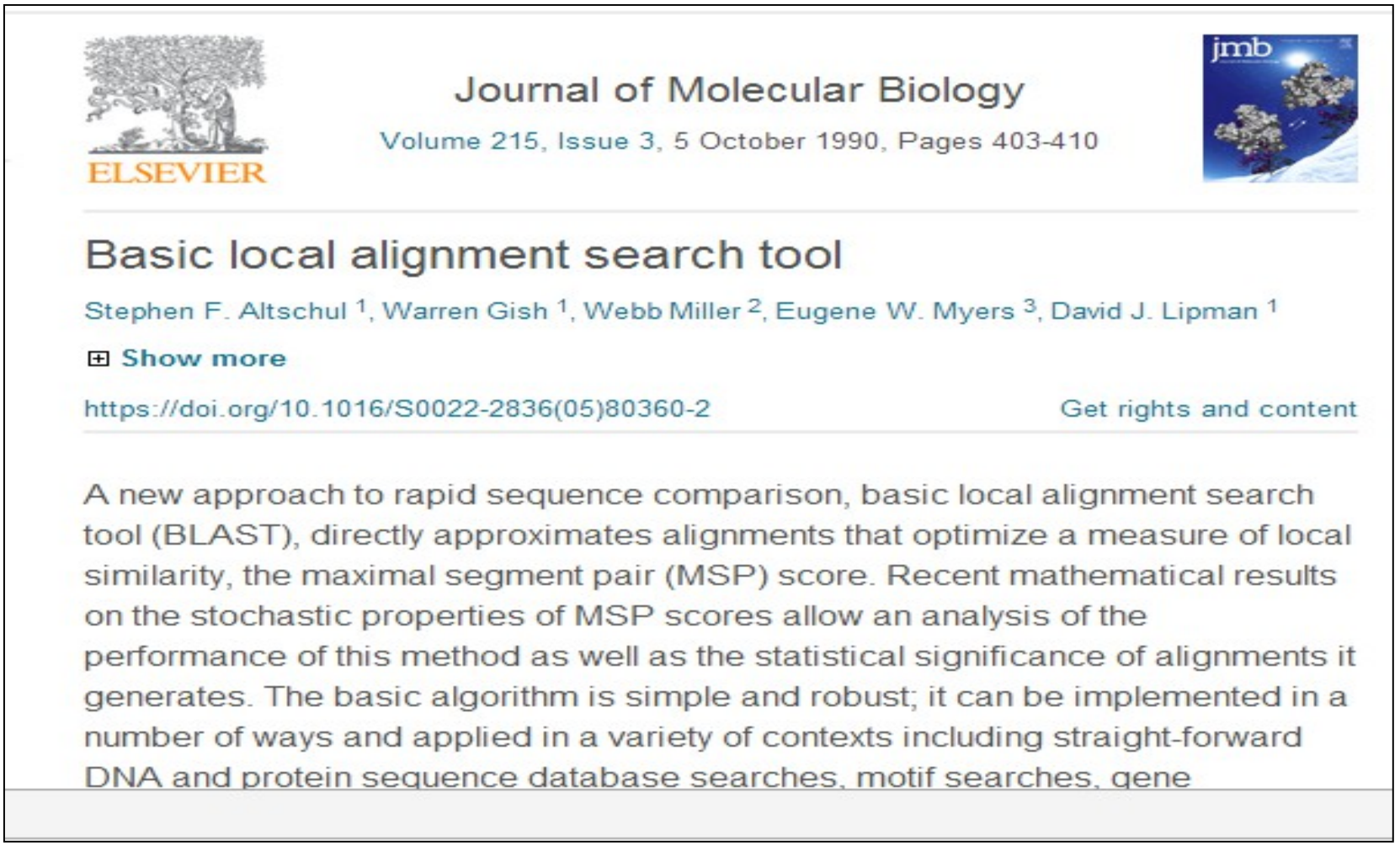
Tools for Sequence Alignments

There are many tools for sequence Alignment. In this session, we will discuss about

- **BLAST**
- **BLAT**
- **CLUSTALW**

Sequence Alignment : BLAST

- **BLAST** stands for **B**asic **L**ocal **A**lignment **S**earch **T**ool



ELSEVIER

Journal of Molecular Biology

Volume 215, Issue 3, 5 October 1990, Pages 403-410

Basic local alignment search tool

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

⊞ Show more

https://doi.org/10.1016/S0022-2836(05)80360-2 Get rights and content

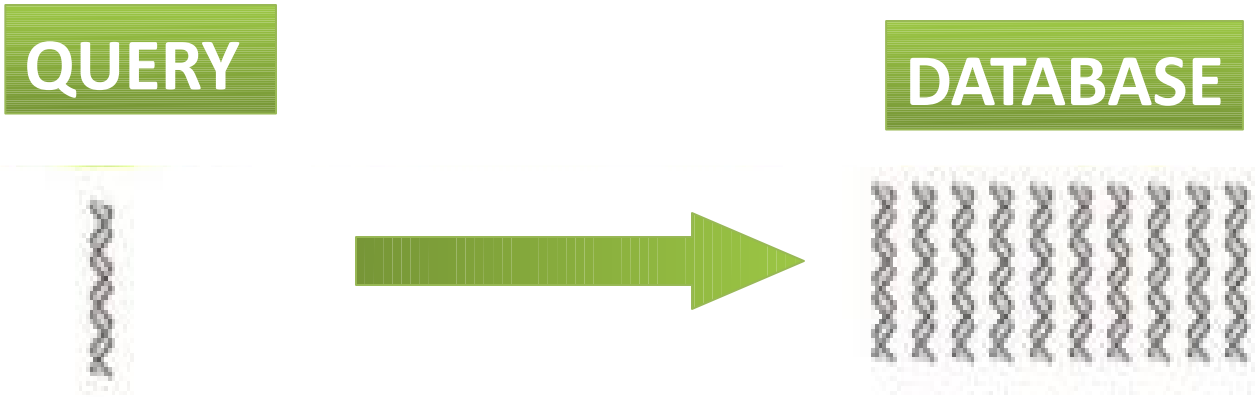
A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP scores allow an analysis of the performance of this method as well as the statistical significance of alignments it generates. The basic algorithm is simple and robust; it can be implemented in a number of ways and applied in a variety of contexts including straight-forward DNA and protein sequence database searches, motif searches, gene

- **Blast was developed by Stephan Altschul and colleagues at NCBI in 1990.**

- **BLAST is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA sequences.**

- **Blast is most used bioinformatics program (cited >60000 times).**

- **A BLAST search enables a researcher to compare a query sequence with a library or databases of sequences, and identify library sequences that resemble the query sequence above a certain threshold.**



Types of BLAST (1)

- **BLASTN : search nucleotide databases using a nucleotide query**
  - **(A)Query : ATGCATCGATC**
  - **(B) Database : ATCGATGATCGACATCGATCAGCTACG**

- **BLASTP : search protein databases using a protein query**
  - **(A)Query : VIVALASVEGAS**
  - **(B) DATABASE : TARDEFGGAVIVADAVISASTILHGGQWLC**

- **BLASTX : search protein databases using a translated nucleotide query**
  - **(A)Query : ATGCATCGATC**
  - **(B)DATABASE : TARDEFGGAVIVADAVISASTILHGGQWLC**

Types of BLAST (2)

- **TBLASTN : search translated nucleotide databases using a protein query**

(A)Query : TARDEFGGAVI
(B)DATABASE : ATCGATGATCGACATCGATCAGCTACG

- **TBLASTX : search translated nucleotide databases using a translated nucleotide query**

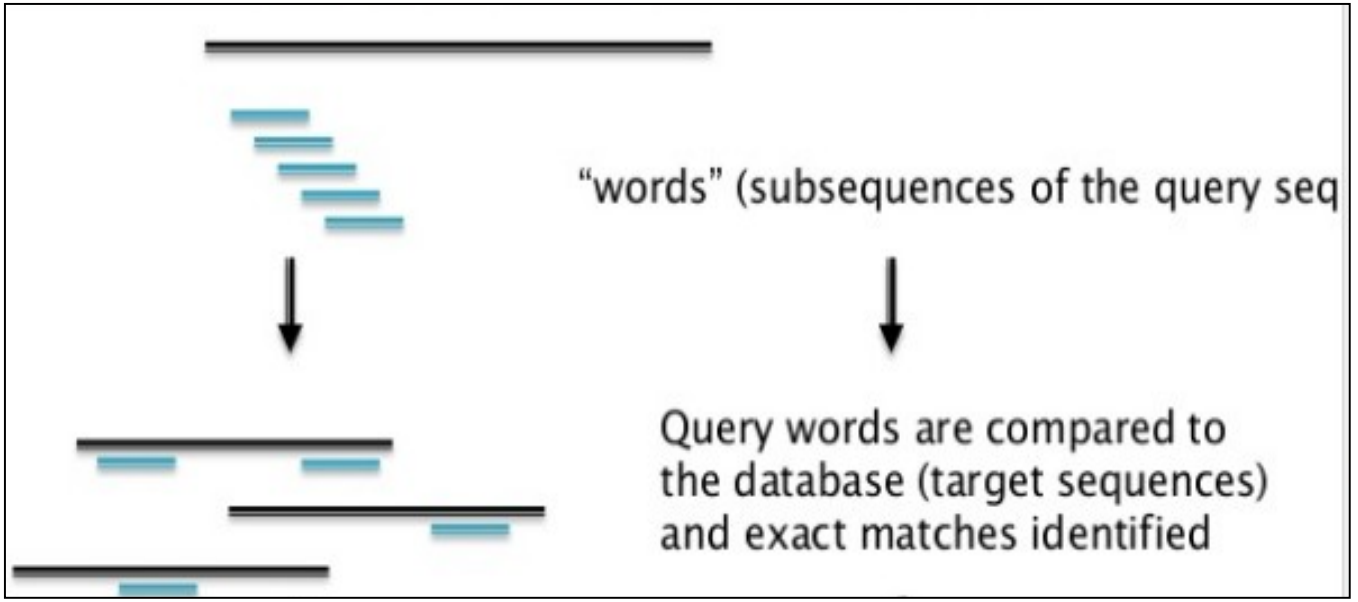
(A)Query : CGATGATCG
(B)DATABASE : ATCGATGATCGACATCGATCAGCTACG

Types of BLAST : ALL

Program	Database	Query
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Protein	Nt. → Protein
TBLASTN	Nt. → Protein	Protein
TBLASTX	Nt. → Protein	Nt. → Protein

How BLAST Works?

- **Construct a dictionary of all words in the query**
- **Initiate a local alignment for each word match between query and DB**



BLAST: Global Alignment

- **It compares the whole sequence with another sequence.**
- **So, output of Global is one to one comparison of two sequences.**
- **This method is useful if you have small group of sequences.**

BLAST: Local Alignment

- **Local method uses the subset of sequence and attempts to align against the subset of another sequence.**

- **So, output of local alignment gives the subset of regions which are highly similar.**

- **Example : Compare two sequence A and B**

(A) GCATTACTAATATATTAGTAAATCAGAGTAGTA
|||||||||
(B) AAGCGAATAATATATTTATACTCAGATTATTGCGCG

BLAST: Input Format

Many program for sequence alignment expect sequences to be in FASTA format

Example 1 :

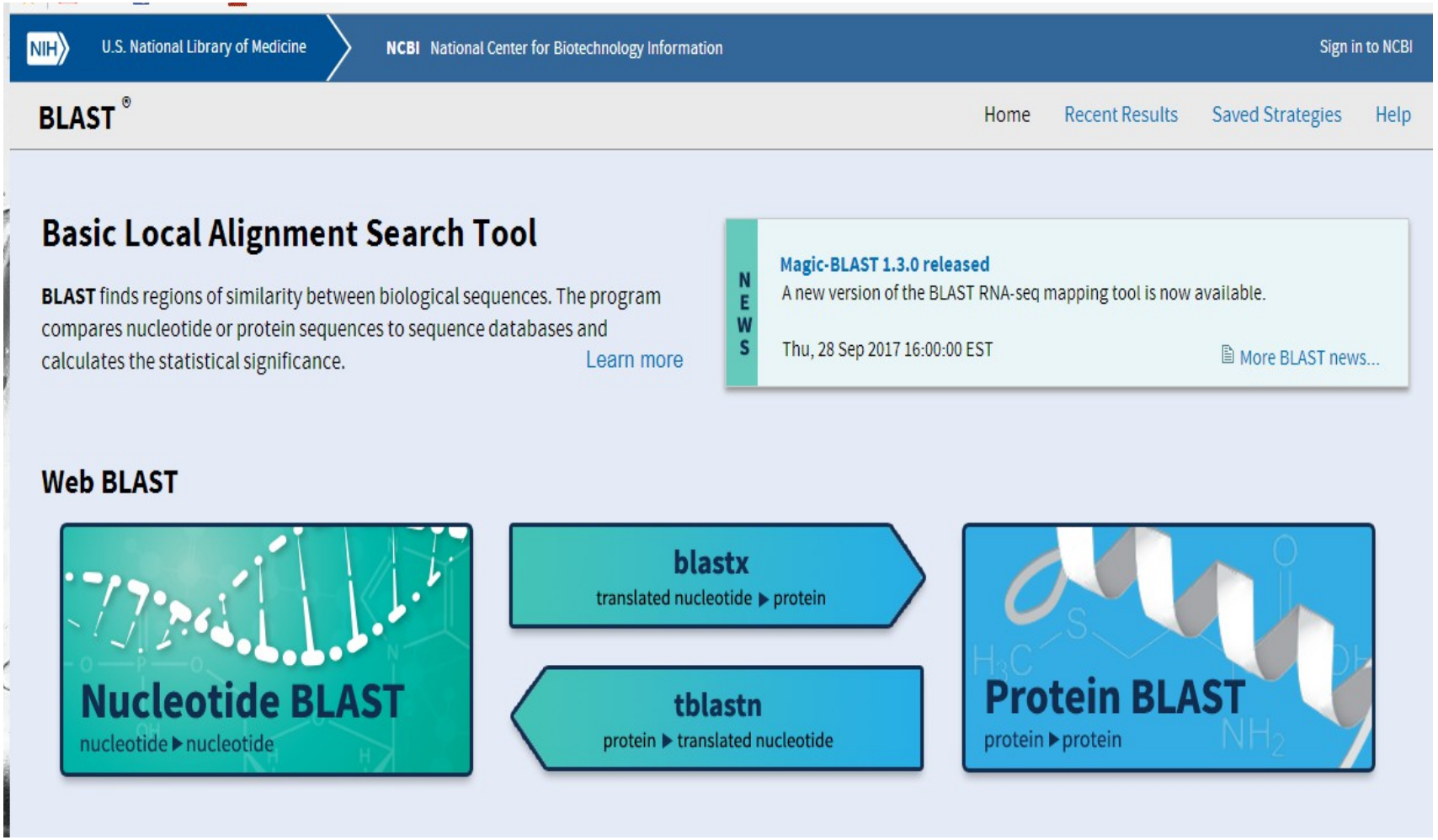
```
>L37107.1 Canis familiaris p53 mRNA, partial cds
GTTCCGTTTGGGGTTCCTGCATTCCGGGACAGCCAAGTCTGTTACTTGGACGTACTCCCCTCTCCTCAAC
AAGTTGTTTTGCCAGCTGGCGAAGACCTGCCCCGTGCAGCTGTGGGTCAGCTCCCCACCCCCACCCAATA
CCTGCGTCCGCGCTATGGCCATCTATAAGAAGTCGGAGTTCGTGACCGAGGTTGTGCGGCGCTGCCCCCA
CCATGAACGCTGCTCTGACAGTAGTGACGGTCTTGCCCCTCCTCAGCATCTCATCCGAGTGGAAGGAAAT
TTGCGGGCCAAGTACCTGGACGACAGAAACACTTTTCGACACAGTGTGGTGGTGCCTTATGAGCCACCCG
AGGTTGGCTCTGACTATACCACCATCCACTACAACTACATGTGTAACAGTTCCTGCATGGGAGGCATGAA
CCGGCGGCCCATCCTCACTATCATCACCCTGGAAGACTCCAGTGGAAACGTGCTGGGACGCAACAGCTTT
GAGGTACGCGTTTGTGCCTGTCCCGGGAGAGACCGCCGGACTGAGGAGGAGAATTTCCACAAGAAGGGGG
AGCCTTGTCCTGAGCCACCCCCCGGGAGTACCAAGCGAGCACTGCCTCCCAGCACCAGCTCCTCTCCCCC
GCAAAAGAAGAAGCCACTAGATGGAGAATATTTCACCCTTCAGATCCGTGGGCGTGAACGCTATGAGATG
TTCAGGAATCTGAATGAAGCCTTGGAGCTGAAGGATGCCCAGAGTGGAAAGGAGCCAGGGGGAAGCAGGG
CTCACTCCAGCCACCTGAAGGCAAAGAAGGGGCAATCTACCTCTCGCCATAAAAAACTGATGTTCAAGAGAGAA
```

Example 2 :

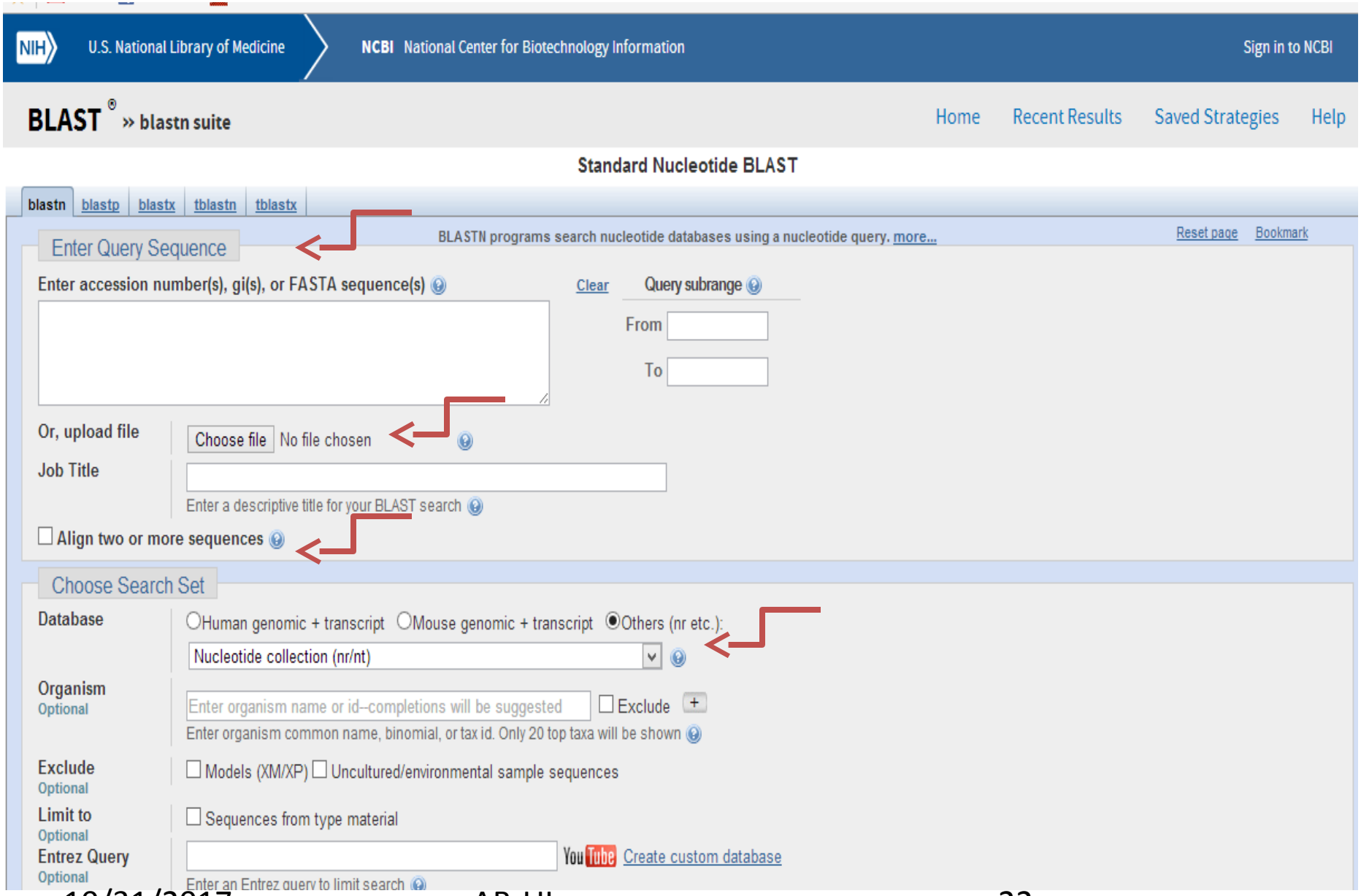
```
>NM_033360.3 Homo sapiens KRAS proto-oncogene, GTPase (KRAS), transcript
variant a, mRNA
TCCTAGGCGGCGGCCGCGGCGGCGGAGGCAGCAGCGGCGGCGGCAGTGGCGGCGGCGAAGGTGGCGGCGG
CTCGGCCAGTACTCCCGGCCCCCGCCATTTCGGACTGGGAGCGAGCGCGGCGCAGGCACTGAAGGCGGCG
GCGGGGCCAGAGGCTCAGCGGCTCCCAGGTGCGGGAGAGAGGCCTGCTGAAAATGACTGAATATAAACTT
GTGGTAGTTGGAGCTGGTGGCGTAGGCAAGAGTGCCTTGACGATACAGCTAATTCAGAATCATTTTGTGG
ACGAATATGATCCAACAATAGAGGATTCCTACAGGAAGCAAGTAGTAATTGATGGAGAAACCTGTCTCTT
GGATATTCTCGACACAGCAGGTCAAGAGGAGTACAGTGCAATGAGGGACCAGTACATGAGGACTGGGGAG
GGCTTTCTTTGTGTATTTGCCATAAATAATACTAAATCATTTGAAGATATTCACCATTATAGAGAACAAA
TTAAAGCAGTTAAGGACTCTGAAAATCTACCTATGGTCCTAGTAGGAAATAAATGTGATTTGCCTTCTAG
```

NCBI BLAST SERVER

Open the website : https://blast.ncbi.nlm.nih.gov/Blast.cgi



Window of BLASTN

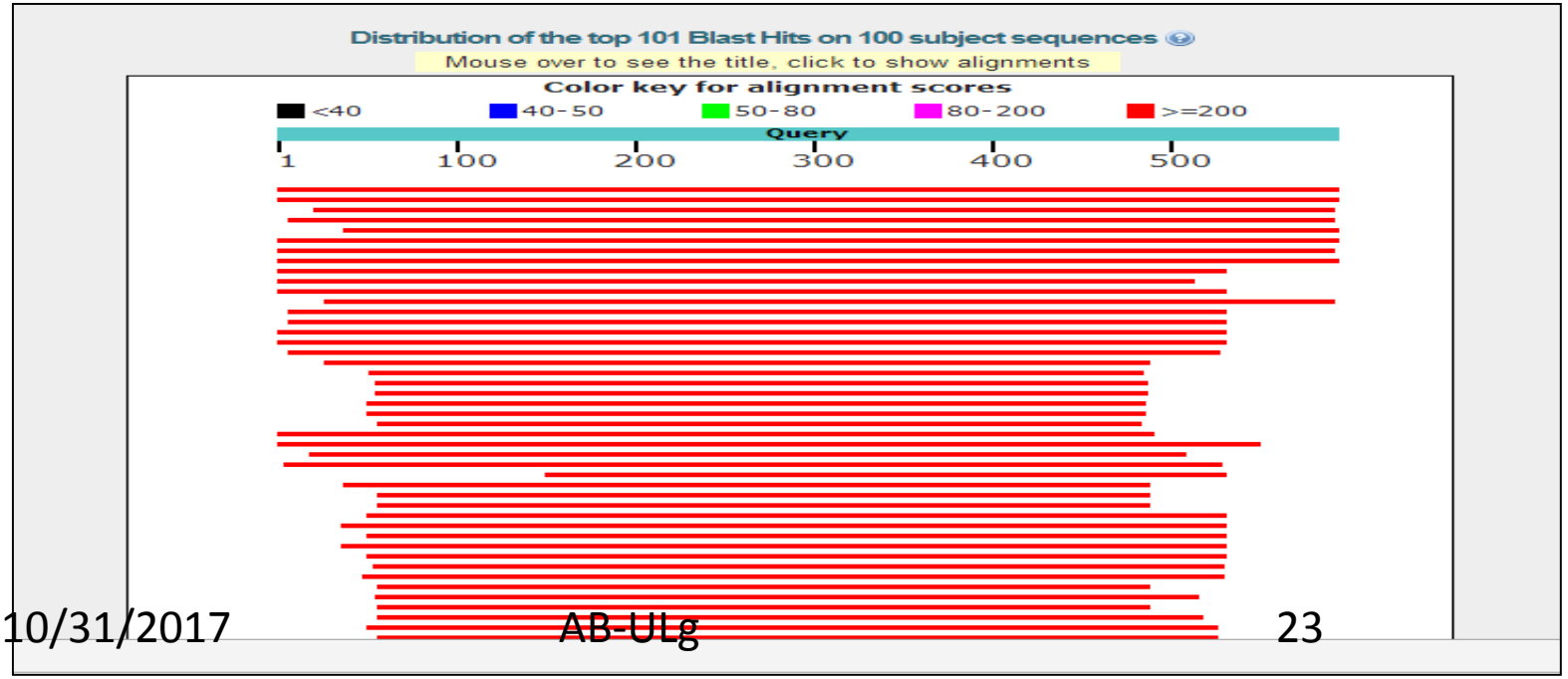


Let us work on BLASTN

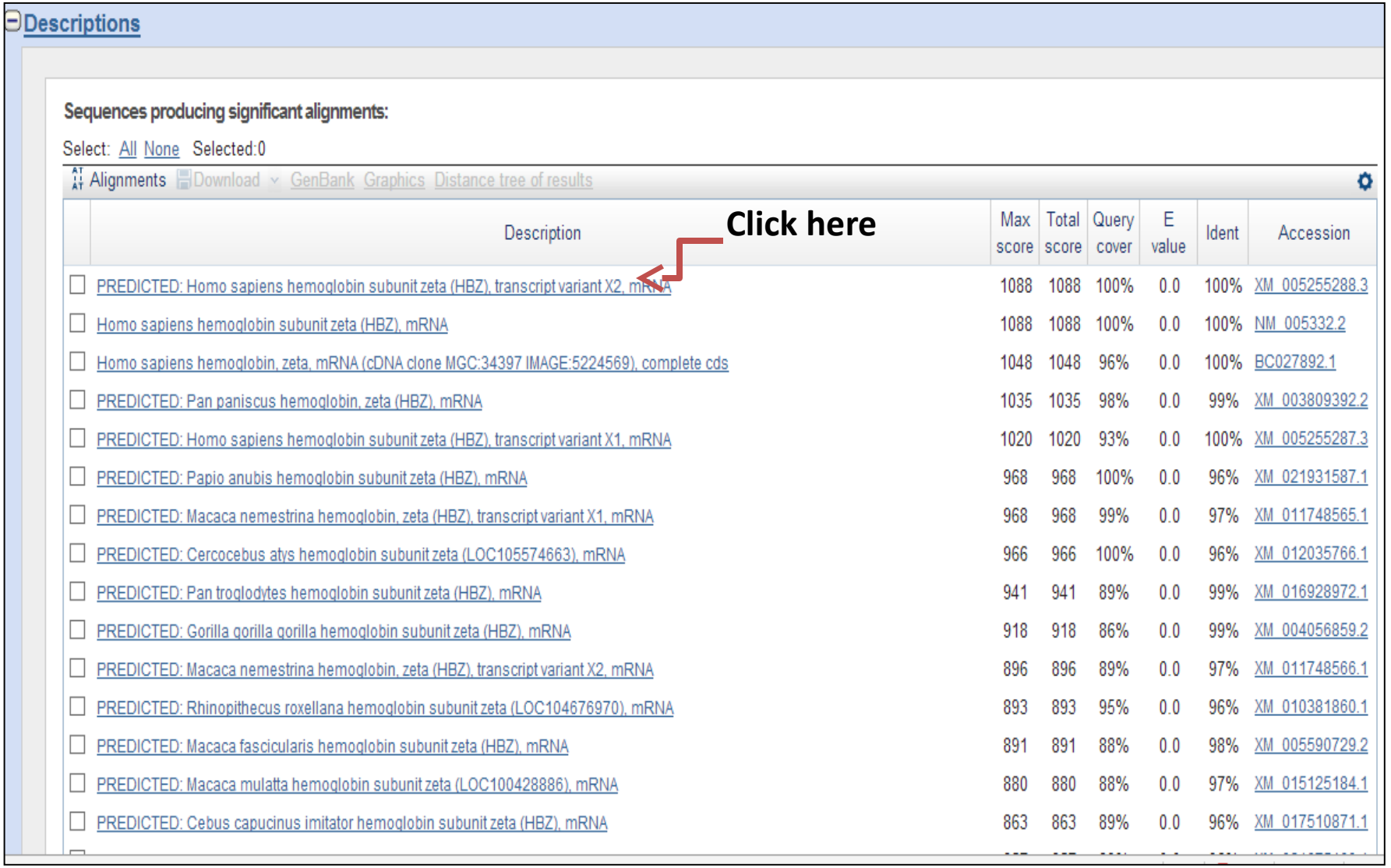
- **Select following sequence and give input into NCBI BLASTN query section**

```
>Seq1
ACCAAGGCCAGTCCTGAGCAGGCCCAACTCCAGTGCAGCTGCCCACCCTGCCGCCATGTCTCTGACCAAG
ACTGAGAGGACCATCATTGTGTCCATGTGGGCCAAGATCTCCACGCAGGCCGACACCATCGGCACCGAGA
CTCTGGAGAGGCTCTTCCTCAGCCACCCGCAGACCAAGACCTACTTCCCGCACTTCGACCTGCACCCGGG
GTCCGCGCAGTTGCGCGCGCACGGCTCCAAGGTGGTGGCCGCCGTGGGCGACGCGGTGAAGAGCATCGAC
GACATCGGCGGCGCCCTGTCCAAGCTGAGCGAGCTGCACGCCTACATCCTGCGCGTGGACCCGGTCAACT
TCAAGCTCCTGTCCCACTGCCTGCTGGTCACCCTGGCCGCGCGCTTCCCCGCCGACTTCACGGCCGAGGC
CCACGCCGCCTGGGACAAGTTCCTATCGGTCGTATCCTCTGTCCTGACCGAGAAGTACCGCTGAGCGCCG
CCTCCGGGACCCCCAGGACAGGCTGCGGCCCCTCCCCCGTCCTGGAGGTTCCCCAGCCCCACTTACCGCG TAATGCGCCAATAAACCAATGAACGAAGC
```

- **You will get list of Hits**



- **You will see statistic of alignments (Identity, E value)**



Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Click here

Description	Max score	Total score	Query cover	E value	Ident	Accession
PREDICTED: Homo sapiens hemoglobin subunit zeta (HBZ), transcript variant X2, mRNA	1088	1088	100%	0.0	100%	XM_005255288.3
Homo sapiens hemoglobin subunit zeta (HBZ), mRNA	1088	1088	100%	0.0	100%	NM_005332.2
Homo sapiens hemoglobin, zeta, mRNA (cDNA clone MGC:34397 IMAGE:5224569), complete cds	1048	1048	96%	0.0	100%	BC027892.1
PREDICTED: Pan paniscus hemoglobin, zeta (HBZ), mRNA	1035	1035	98%	0.0	99%	XM_003809392.2
PREDICTED: Homo sapiens hemoglobin subunit zeta (HBZ), transcript variant X1, mRNA	1020	1020	93%	0.0	100%	XM_005255287.3
PREDICTED: Papio anubis hemoglobin subunit zeta (HBZ), mRNA	968	968	100%	0.0	96%	XM_021931587.1
PREDICTED: Macaca nemestrina hemoglobin, zeta (HBZ), transcript variant X1, mRNA	968	968	99%	0.0	97%	XM_011748565.1
PREDICTED: Cercocebus atys hemoglobin subunit zeta (LOC105574663), mRNA	966	966	100%	0.0	96%	XM_012035766.1
PREDICTED: Pan troglodytes hemoglobin subunit zeta (HBZ), mRNA	941	941	89%	0.0	99%	XM_016928972.1
PREDICTED: Gorilla gorilla gorilla hemoglobin subunit zeta (HBZ), mRNA	918	918	86%	0.0	99%	XM_004056859.2
PREDICTED: Macaca nemestrina hemoglobin, zeta (HBZ), transcript variant X2, mRNA	896	896	89%	0.0	97%	XM_011748566.1
PREDICTED: Rhinopithecus roxellana hemoglobin subunit zeta (LOC104676970), mRNA	893	893	95%	0.0	96%	XM_010381860.1
PREDICTED: Macaca fascicularis hemoglobin subunit zeta (HBZ), mRNA	891	891	88%	0.0	98%	XM_005590729.2
PREDICTED: Macaca mulatta hemoglobin subunit zeta (LOC100428886), mRNA	880	880	88%	0.0	97%	XM_015125184.1
PREDICTED: Cebus capucinus imitator hemoglobin subunit zeta (HBZ), mRNA	863	863	89%	0.0	96%	XM_017510871.1

- **How well alignment is ? : Bad, Good, Very Good?**

PREDICTED: Homo sapiens hemoglobin subunit zeta (HBZ), transcript variant X2, mRNA

Sequence ID: XM_005255288.3 **Length:** 1342 **Number of Matches:** 1

Range 1: 748 to 1336 GenBank Graphics ▼ Next Match ▲ Previous Match

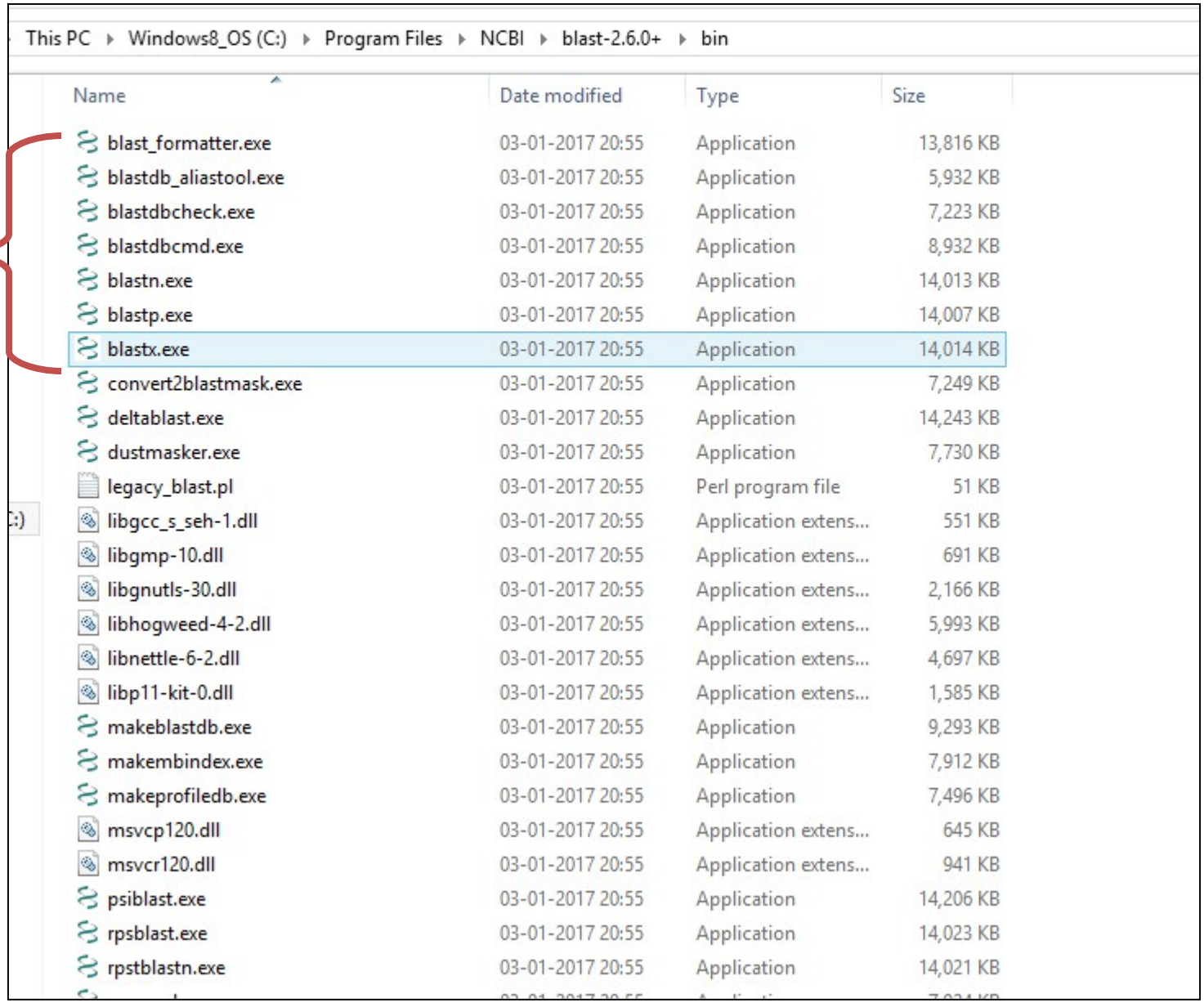
Score	Expect	Identities	Gaps	Strand
1088 bits(589)	0.0	589/589(100%)	0/589(0%)	Plus/Plus

```
Query  1     ACCAAGGCCAGTCCTGAGCAGGCCCAACTCCAGTGCAGCTGCCCACCCTGCCGCCATGTC  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  748   ACCAAGGCCAGTCCTGAGCAGGCCCAACTCCAGTGCAGCTGCCCACCCTGCCGCCATGTC  807

Query  61    TCTGACCAAGACTGAGAGGACCATCATTGTGTCCATGTGGGCCAAGATCTCCACGCAGGC  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  808   TCTGACCAAGACTGAGAGGACCATCATTGTGTCCATGTGGGCCAAGATCTCCACGCAGGC  867

Query  121   CGACACCATCGGCACCGAGACTCTGGAGAGGCTCTTCCTCAGCCACCCGCAGACCAAGAC  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  868   CGACACCATCGGCACCGAGACTCTGGAGAGGCTCTTCCTCAGCCACCCGCAGACCAAGAC  927

Query  181   CTACTTCCCGCACTTCGACCTGCACCCGGGGTCCGCGCAGTTGCGCGCGCACGGCTCCAA  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  928   CTACTTCCCGCACTTCGACCTGCACCCGGGGTCCGCGCAGTTGCGCGCGCACGGCTCCAA  987

Query  241   GGTGGTGGCCGCCGTGGGCGACGCGGTGAAGAGCATCGACGACATCGGCGGCGCCCTGTC  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  988   GGTGGTGGCCGCCGTGGGCGACGCGGTGAAGAGCATCGACGACATCGGCGGCGCCCTGTC  1047

Query  301   CAAGCTGAGCGAGCTGCACGCCTACATCCTGCGCGTGGACCCGGTCAACTTCAAGCTCCT  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1048  CAAGCTGAGCGAGCTGCACGCCTACATCCTGCGCGTGGACCCGGTCAACTTCAAGCTCCT  1107

Query  361   GTCCCACTGCCTGCTGGTCACCCTGGCCGCGCGCTTCCCCGCCGACTTCACGGCCGAGGC  420
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1108  GTCCCACTGCCTGCTGGTCACCCTGGCCGCGCGCTTCCCCGCCGACTTCACGGCCGAGGC  1167

Query  421   CCACGCCGCCTGGGACAAGTTCCTATCGGTCGTATCCTCTGTCCTGACCGAGAAGTACCG  480
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1168  CCACGCCGCCTGGGACAAGTTCCTATCGGTCGTATCCTCTGTCCTGACCGAGAAGTACCG  1227

Query  481   CTGAGCGCCGCCTCCGGGACCCCCAGGACAGGCTGCGGCCCCTCCCCCGTCCTGGAGGTT  540
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1228  CTGAGCGCCGCCTCCGGGACCCCCAGGACAGGCTGCGGCCCCTCCCCCGTCCTGGAGGTT  1287

Query  541   CCCCAGCCCCACTTACCGCGTAATGCGCCAATAAACCAATGAACGAAGC  589
             ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1288  CCCCAGCCCCACTTACCGCGTAATGCGCCAATAAACCAATGAACGAAGC  1336
```

QUESTIONS

Offline BLAST

- **Download NCBI blast from following link.**
- ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/ncbi-bl
- (Wait for 5-10 minutes)
- **Double click on .exe and follow the instruction for the installation.**
- **step A: create database**
- **step B: chose program (blastp or blastn)**


This PC ▸ Windows8_OS (C:) ▸ Program Files ▸ NCBI ▸ blast-2.6.0+ ▸ bin
Name
Date modified
Type
Size
blast_formatter.exe 03-01-2017 20:55 Application 13,816 KB
blastdb_aliastool.exe 03-01-2017 20:55 Application 5,932 KB
blastdbcheck.exe 03-01-2017 20:55 Application 7,223 KB
blastdbcmd.exe 03-01-2017 20:55 Application 8,932 KB
blastn.exe 03-01-2017 20:55 Application 14,013 KB
blastp.exe 03-01-2017 20:55 Application 14,007 KB
blastx.exe 03-01-2017 20:55 Application 14,014 KB
convert2blastmask.exe 03-01-2017 20:55 Application 7,249 KB
deltablast.exe 03-01-2017 20:55 Application 14,243 KB
dustmasker.exe 03-01-2017 20:55 Application 7,730 KB
legacy_blast.pl 03-01-2017 20:55 Perl program file 51 KB
libgcc_s_seh-1.dll 03-01-2017 20:55 Application extens... 551 KB
libgmp-10.dll 03-01-2017 20:55 Application extens... 691 KB
libgnutls-30.dll 03-01-2017 20:55 Application extens... 2,166 KB
libhogweed-4-2.dll 03-01-2017 20:55 Application extens... 5,993 KB
libnettle-6-2.dll 03-01-2017 20:55 Application extens... 4,697 KB
libp11-kit-0.dll 03-01-2017 20:55 Application extens... 1,585 KB
makeblastdb.exe 03-01-2017 20:55 Application 9,293 KB
makembindex.exe 03-01-2017 20:55 Application 7,912 KB
makeprofiledb.exe 03-01-2017 20:55 Application 7,496 KB
msvcp120.dll 03-01-2017 20:55 Application extens... 645 KB
msvcr120.dll 03-01-2017 20:55 Application extens... 941 KB
psiblast.exe 03-01-2017 20:55 Application 14,206 KB
rpsblast.exe 03-01-2017 20:55 Application 14,023 KB
rpstblastn.exe 03-01-2017 20:55 Application 14,021 KB

Let us run BLASTN (1)

- **Download sample data from course website.**

- **Create the database (nucleotide) of Sequence.fasta by command as follows:**

makeblastdb -in Sequence.fasta -input_type fasta -dbtype nucl –out /path/nucle_db

```
New DB name:   C:\Users\RIL\Desktop\blast-run\db
New DB title:  Sequence.fasta
Sequence type: Nucleotide
Keep MBits: T
Maximum file size: 10000000000B
Adding sequences from FASTA; added 12 sequences in 0.0123802 seconds.

C:\Program Files\NCBI\blast-2.6.0+\bin>makeblastdb.exe -in Sequence.fasta -input
_type fasta -dbtype nucl -out C:\Users\RIL\Desktop\blast-run\nucle_db


Building a new DB, current time: 10/27/2017 14:59:03
New DB name:   C:\Users\RIL\Desktop\blast-run\nucle_db
New DB title:  Sequence.fasta
Sequence type: Nucleotide
Keep MBits: T
Maximum file size: 10000000000B
Adding sequences from FASTA; added 12 sequences in 0.0109341 seconds.

C:\Program Files\NCBI\blast-2.6.0+\bin>_
```

Let us run BLASTN (2)

- **Three output files will be created with extension of .nin, nhr, nsq**

  nucle_db.nin

  nucle_db.nhr

  nucle_db.nsq

- **Now Run query file against the database file by following command**

Blastn.exe –query sequence_query.fasta –db nucle_db –out blastn-output

BLASTN OUTPUT

```
BLASTN 2.6.0+


Reference: Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb
Miller (2000), "A greedy algorithm for aligning DNA sequences", J
Comput Biol 2000; 7(1-2):203-14.


Database: Sequence.fasta
           12 sequences; 1,374 total letters


Query= BD218070.1 Compositions and methods for inhibiting human
immunodeficiency virus infection by down-regulating human cellular
genes

Length=28
                                                                  Score     E
Sequences producing significant alignments:                      (Bits)  Value

   BD218070.1 Compositions and methods for inhibiting human immuno...  52.8    3e-012


 > BD218070.1 Compositions and methods for inhibiting human immunodeficiency
virus infection by down-regulating human cellular
genes
Length=28

 Score = 52.8 bits (28),  Expect = 3e-012
 Identities = 28/28 (100%), Gaps = 0/28 (0%)
 Strand=Plus/Plus

Query  1   CTCGGAATTCAAGCTTATGGATGGATGG  28
           ||||||||||||||||||||||||||||
Sbjct  1   CTCGGAATTCAAGCTTATGGATGGATGG  28


Lambda      K        H
     1.33    0.621     1.12

 Gapped
```

Database used

Single query sequence

E value

Alignment from 1-28 bps

Let us run BLASTP (1)

- **Create the database of sequence_protein.fasta by command as follows:**

makeblastdb -in sequnce_protien.fasta -input_type fasta -dbtype prot –out /path/prot_db

- **Three output files will be created with extension of .nin, nhr, nsq**

Blastp.exe –query sequence_protein_query.fasta –db nucle_db –out blastn-output

BLASTP OUTPUT

Database used

```
Database: sequence_protein.fasta
           11 sequences; 1,458 total letters
```

Single query sequence

```
Query= ABG47031.1 hemoglobin, partial [Homo sapiens]

Length=105
```

E value

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

   ABG47031.1 hemoglobin, partial [Homo sapiens]                  216     5e-079
   CDB75647.1 hemoglobin [Clostridium sp. CAG:265]                18.1    0.23
   EPH11744.1 hemoglobin [Myroides odoratimimus CCUG 12700]       15.0    2.7
   ERF86166.1 hemoglobin [Bradyrhizobium sp. DFCI-1]              14.2    5.6
```

Alignment from 1-60 Amino acids

```
 > ABG47031.1 hemoglobin, partial [Homo sapiens]
Length=105

  Score = 216 bits (551),  Expect = 5e-079, Method: Compositional matrix adjust.
  Identities = 105/105 (100%), Positives = 105/105 (100%), Gaps = 0/105 (0%)

Query  1    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
            MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
Sbjct  1    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60

Query  61   VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFR  105
            VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFR
Sbjct  61   VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFR  105


 > CDB75647.1 hemoglobin [Clostridium sp. CAG:265]
Length=145
```

RESULT INTERPRETATION

1. **How many sequence crossed threshold e value ???**
2. **How many sequences shows > 50 % identity with database ??**
3. **How many sequences shows > 90 % identity with database ??**
4. **Prepare tabular output for BLASTP and BLASTN results.**

Blastx : Let us run

1 . Perform the blastx

2. How many sequences shows 90% identity against the database

3. What is their e-value ??

QUESTIONS

- **Is it possible to localise its position on human genome ?**
- **How to analysis its gene structure ?**
- **For this, Open the UCSC Browser available at https://genome.ucsc.edu/**

UNIVERSITY OF CALIFORNIA SANTA CRUZ UCSC **Genome Browser**

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Our tools

- **Genome Browser**
  interactively visualize genomic data
- **BLAT**
  rapidly align sequences to the genome
- **Table Browser**
  download data from the Genome Browser database
- **Variant Annotation Integrator**
  get functional effect predictions for variant calls
- **Data Integrator**
  combine data sources from the Genome Browser database
- **Gene Sorter**
  find genes that are similar by expression and other metrics
- **Genome Browser in a Box (GBiB)**
  run the Genome Browser on your laptop or server
- **In-Silico PCR**
  rapidly align PCR primer pairs to the genome
- **LiftOver**
  convert genome coordinates between assemblies
- **VisiGene**
  interactively view in situ images of mouse and frog

More tools...

Click "BLAT"

Difference Between BLAST and BLAT

- **Blat is an alignment tool like BLAST, but it is structured differently.**

- **Blat works by keeping an index of an entire genome in memory.**

- **Thus, the target database of BLAT is not a set of GenBank sequences, but instead an index derived from the assembly of the entire genome.**

Advantages of BLAT over BLAST

- Its Speed is very high(no queues, response in seconds).
- The ability to submit a long list of simultaneous queries in fasta format.
- A direct link into the UCSC browser.
- Alignment block details in natural genomic order.
- An option to launch the alignment later as part of a custom track.

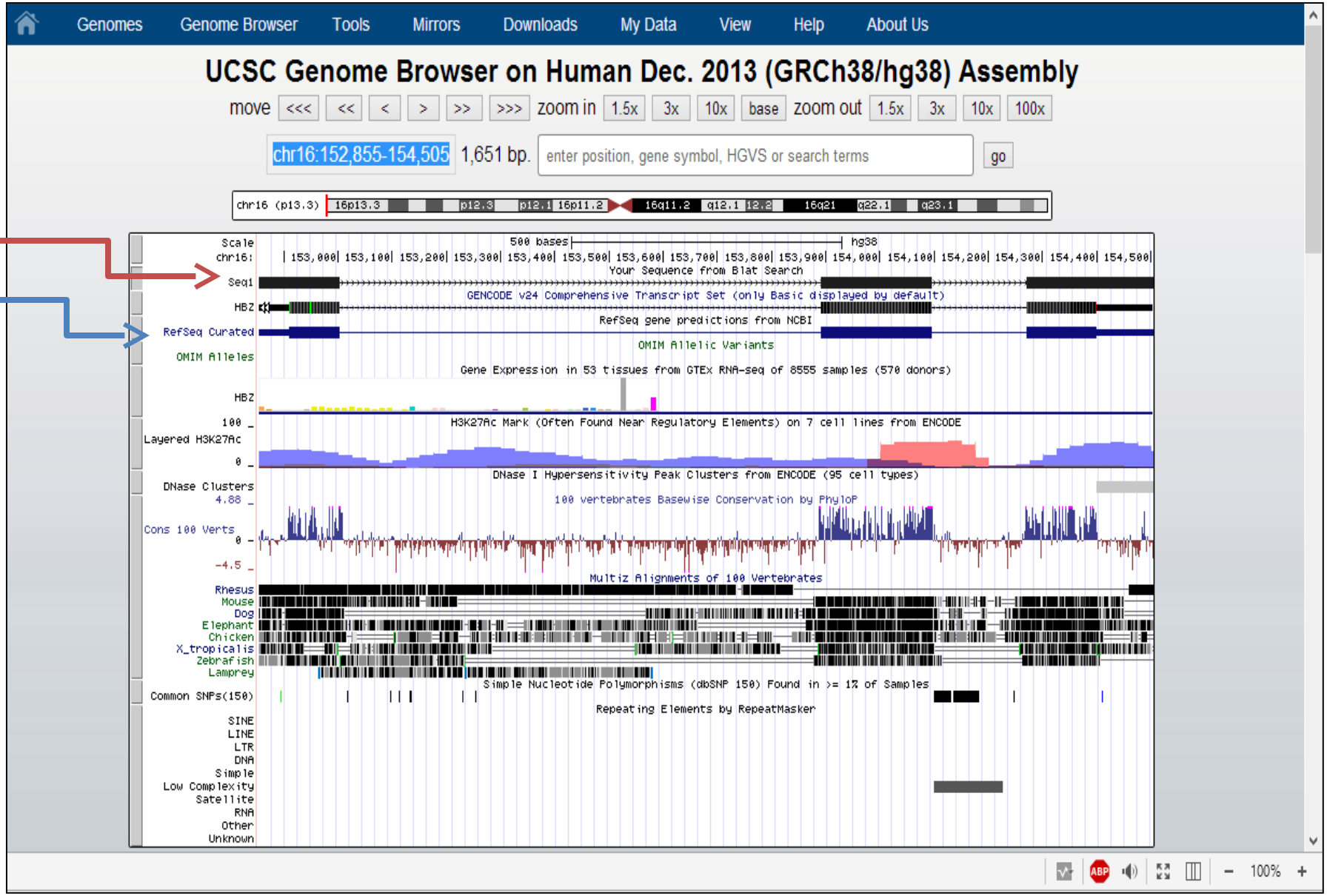
- **Paste following sequence into Query search Box and click Submit**

```
>Seq1
ACCAAGGCCAGTCCTGAGCAGGCCCAACTCCAGTGCAGCTGCCCACCCTGCCGCCATGTC
TCTGACCAAGACTGAGAGGACCATCATTGTGTCCATGTGGGCCAAGATCTCCACGCAGGC
CGACACCATCGGCACCGAGACTCTGGAGAGGCTCTTCCTCAGCCACCCGCAGACCAAGAC
CTACTTCCCGCACTTCGACCTGCACCCGGGGTCCGCGCAGTTGCGCGCGCACGGCTCCAA
GGTGGTGGCCGCCGTGGGCGACGCGGTGAAGAGCATCGACGACATCGGCGGCGCCCTGTC
CAAGCTGAGCGAGCTGCACGCCTACATCCTGCGCGTGGACCCGGTCAACTTCAAGCTCCT
GTCCCACTGCCTGCTGGTCACCCTGGCCGCGCGCTTCCCCGCCGACTTCACGGCCGAGGC
CCACGCCGCCTGGGACAAGTTCCTATCGGTCGTATCCTCTGTCCTGACCGAGAAGTACCG
CTGAGCGCCGCCTCCGGGACCCCCAGGACAGGCTGCGGCCCCTCCCCCGTCCTGGAGGTT
CCCCAGCCCCACTTACCGCGTAATGCGCCAATAAACCAATGAACGAAGC
```

What did you get ??

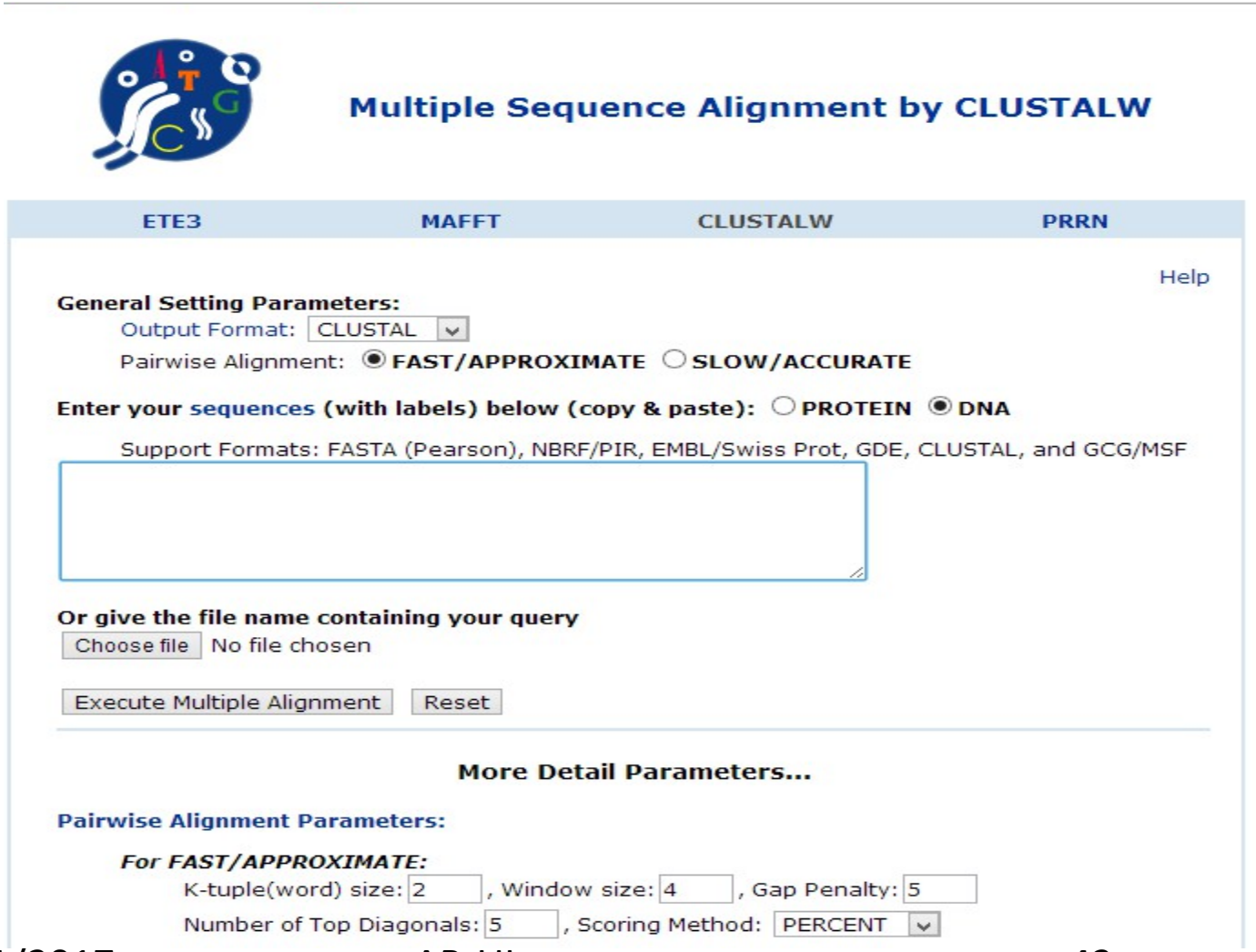
Can you see your sequence ? How ?

How many exons are present in your sequence ?

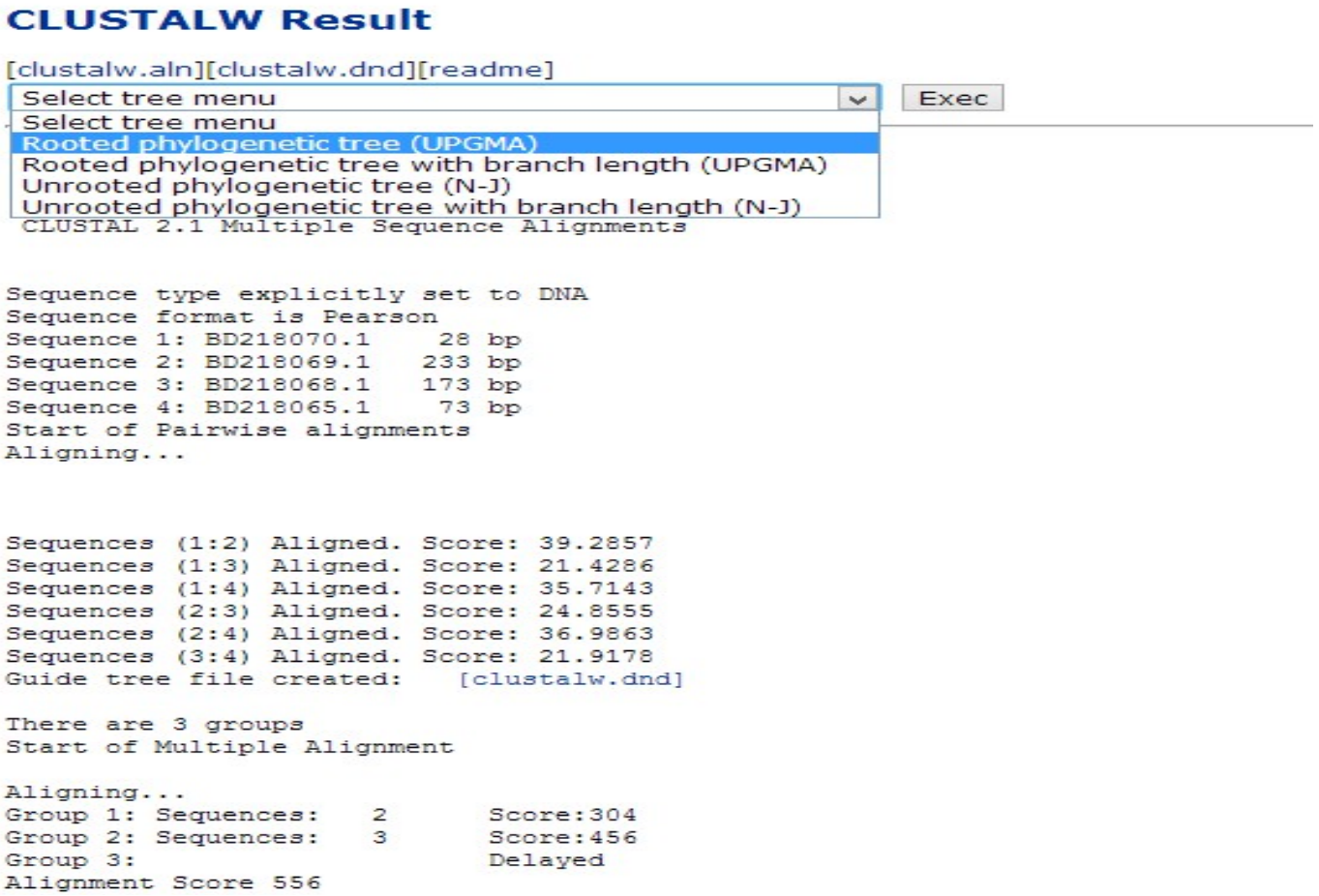

Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us
UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr16:152,855-154,505 1,651 bp.
enter position, gene symbol, HGVS or search terms
go
chr16 (p13.3)
Scale
chr16:
500 bases
hg38
Your Sequence from Blat Search
Seq1
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
HBZ
RefSeq gene predictions from NCBI
RefSeq Curated
OMIM Allelic Variants
OMIM Alleles
Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples (570 donors)
HBZ
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
DNase Clusters
100 vertebrates Basewise Conservation by PhyloP
Cons 100 Verts
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Simple Nucleotide Polymorphisms (dbSNP 150) Found in >= 1% of Samples
Common SNPs(150)
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
100%

Multiple Sequence Alignment

- **Open the link http://www.genome.jp/tools-bin/clustalw**



- **Copy DNA sequence from sample file.**
- **Check RESULT and Select UPGMA from dropdown option**



Interpretation of result

1. Two Highly closely related sequence

2 . Both will share same functionality

Multiple Sequence Alignment : Protein Sequence

- Copy Protein sequence from sample file.
- Check RESULT and Select UPGMA from dropdown option
- How tree look like ?
- What are highly similar sequence? Define their Gene Ids.

QUESTIONS