

## BLAST Input: Select a search type

### BLAST Assembled Genomes

Contains links to genomic BLAST pages for common organisms, and a link to a complete list of available organism genome BLAST pages

### B.I.P.

### Basic BLAST

Contains links to BLAST forms for the traditional set of databases (e.g., nr, est, etc.). Choose the link for the search type you want. For example, choose "nucleotide blast" to search a nucleotide database using a nucleotide query.

### Specialized BLAST

Contains links to special-purpose BLAST databases and tools such as trace archives and IgBLAST

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

The screenshot shows the NCBI BLAST website interface. The browser address bar displays <http://blast.ncbi.nlm.nih.gov/Blast.cgi>. The page title is "BLAST: Basic Local Alignment Search Tool". The navigation menu includes "Home", "Recent Results", "Saved Strategies", and "Help". The main content area is titled "NCBI BLAST Home" and features a search bar with the text "BLAST finds regions of similarity between biological sequences. more...". Below the search bar, there is a "New" announcement: "Aligning Multiple Protein Sequences? Try the COBALT Multiple Alignment Tool. (Go)". The "BLAST Assembled Genomes" section lists various species genomes to search, including Human, Mouse, Rat, Arabidopsis thaliana, Oryza sativa, Bos taurus, Danio rerio, Drosophila melanogaster, Gallus gallus, Pan troglodytes, Microbes, and Apis mellifera. The "Basic BLAST" section lists search programs: nucleotide blast, protein blast, blastx, tblastn, and tblastx, each with a brief description and available algorithms. The "Specialized BLAST" section lists various specialized search options, such as Primer-BLAST, trace archives, conserved domains, conserved domain architecture (cdart), gene expression profiles (GEO), immunoglobulins (IgBLAST), SNPs (snp), and vector contamination (vecscreen).

# BLAST Input: Input a QUERY sequence and select a SUBJECT database

## Enter a QUERY sequence

Provides a place to input or upload your query sequence, and optionally select a query subrange.

## Select "nr" database

### Choose Search Set

Is where you select a database and optionally limit your search by an organism or Entrez query. The default database "Human genomic + transcript", implicitly limits the search to Human. You can also select a database that is not species-specific (e.g., nr).

### Program selection

Allows you to optimize your search for different scenarios (e.g., intra- vs. inter-species searches). The choices correspond to megablast, discontinuous megablast, and blastn for nucleotide; and blastp, PSI-BLAST and PHI-BLAST for protein.

The screenshot shows the NCBI BLAST search interface in a Mozilla Firefox browser window. The URL is <http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&>. The page title is "Nucleotide BLAST: Search nucleotide...". The interface includes a navigation bar with "Home", "Recent Results", "Saved Strategies", and "Help". The main content area is titled "BLAST" and "Basic Local Alignment Search Tool". It features a "Enter Query Sequence" section with a text area containing a FASTA sequence, a "Query subrange" section with "From" and "To" input fields, and an "Or, upload file" section with a "Browse..." button. Below this is the "Choose Search Set" section, where the "Database" is set to "Nucleotide collection (nr/nt)" and the "Organism" is set to "Human genomic + transcript". The "Program Selection" section shows "Optimize for" set to "Highly similar sequences (megablast)". A "BLAST" button is present, along with a checkbox for "Show results in a new window". A note at the bottom right states "Note: Parameter values that differ from the default are highlighted in yellow".

# BLAST Input: Advanced parameters

## Max Target Sequences

Maximum number of aligned sequences to display in results

## Short Queries

Improve results for short queries

## E-value cutoff

Expected number of chance matches in a random model, i.e. by chance. Default: 10 matches per query

## Word Size

The length of the seed (short DNA/protein sequence) that initiates an alignment.

## Match/Mismatch Scores

Reward and penalty for matching and mismatching bases.

## Gap Costs

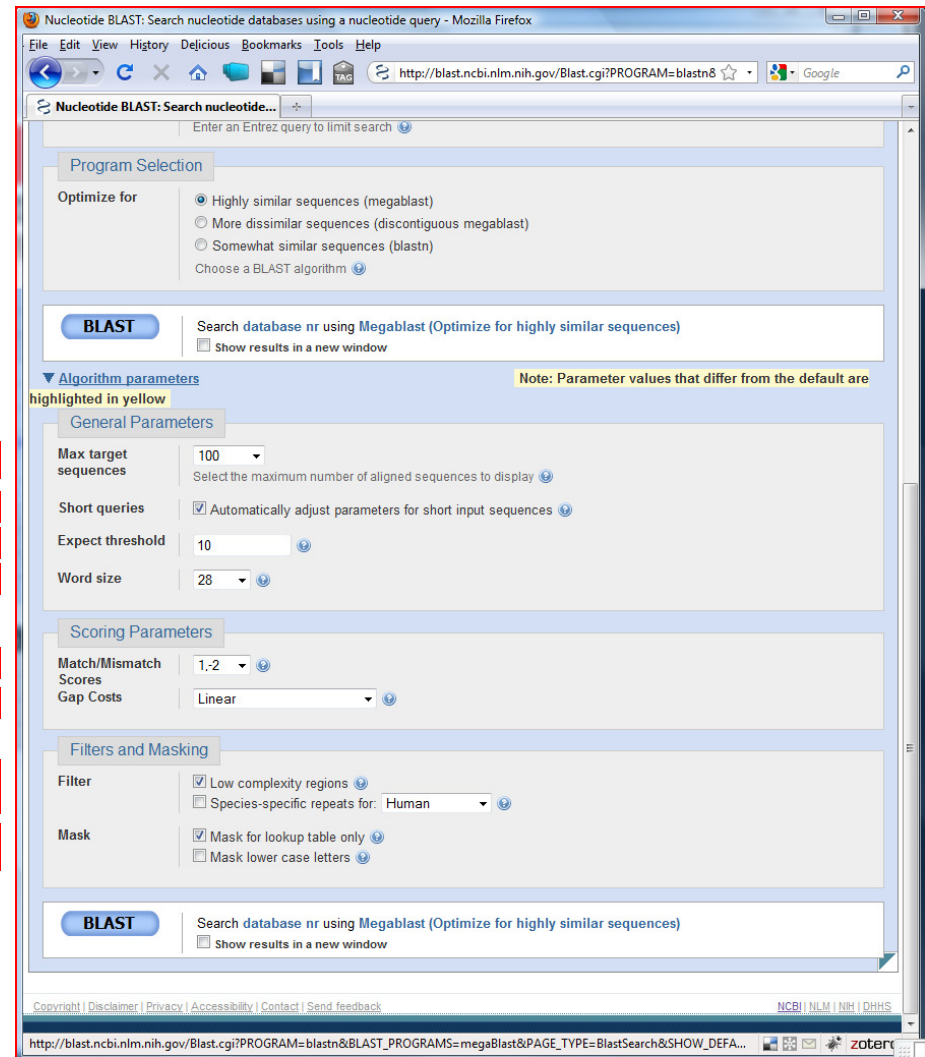
Cost to create and extend a gap in an alignment. Linear costs are available only with megablast and are determined by the match/mismatch scores.

## Filter

Mask regions of low compositional complexity, or species-specific repeats that may cause spurious or misleading results.

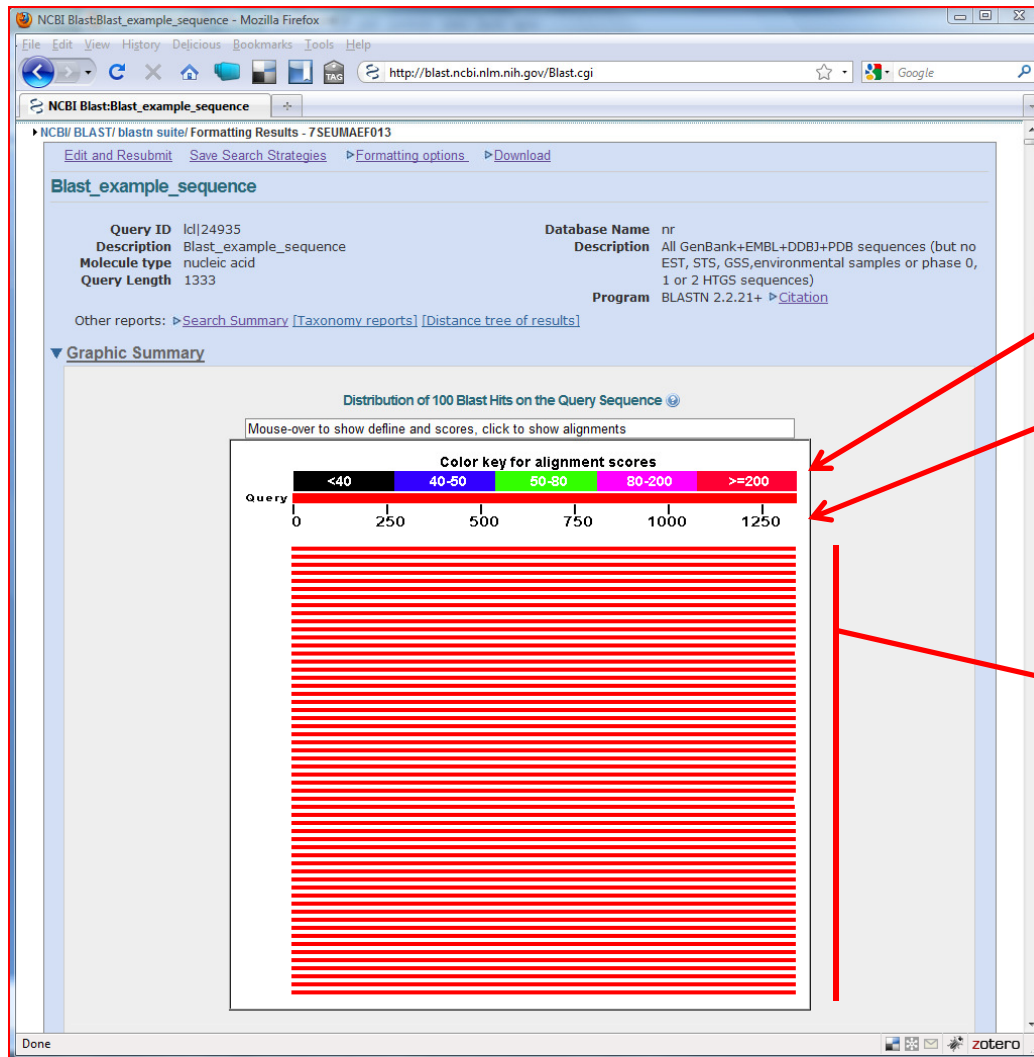
## Mask

Allow extension of alignment through repetitive regions  
Filter lower case bases in QUERY



# BLAST Output: Graphic Summary

This graphic is an overview of database sequences aligned to the query sequence



Alignments are color-coded by score, within one of five score ranges.

Range of length in QUERY

Multiple alignments on the same database sequence are connected by a dashed line (not shown)

Mousing over an alignment shows the alignment definition and score in the box at the top.

Clicking an alignment displays the alignment detail (see below)

# BLAST Output: Descriptions

**Description**

One-line description  
of SUBJECT

**Max Score**

Score for highest  
scoring alignment to  
SUBJECT

**Total Score**

Sum of scores for all  
alignments to SUBJECT

**Query Coverage**

% of QUERY covered by  
alignments

**GenBank Accession**

Database identifier of  
SUBJECT

**E-value**

Probability of such an  
alignment by chance

**Max. Ident.**

Highest percent  
identity

**Links**

Hyperlinks to other  
databases that contain  
the SUBJECT

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
<a href="#">AB450625.1</a>	Influenza A virus (A/open-bill stork/Thailand/VSMU-29-NSN/2005(H5N1))	2462	2462	100%	0.0	100%	
<a href="#">AB450624.1</a>	Influenza A virus (A/tree sparrow/Rachaburi/VSMU-16-RBR/2005(H5N1))	2462	2462	100%	0.0	100%	
<a href="#">AB450623.1</a>	Influenza A virus (A/open-bill stork/Thailand/VSMU-15-ATG/2005(H5N1))	2462	2462	100%	0.0	100%	
<a href="#">AB450622.1</a>	Influenza A virus (A/tree sparrow/Thailand/VSMU-14-KRI/2005(H5N1))	2462	2462	100%	0.0	100%	
<a href="#">AB450621.1</a>	Influenza A virus (A/tree sparrow/Thailand/VSMU-12-KRI/2005(H5N1))	2462	2462	100%	0.0	100%	
<a href="#">AB450620.1</a>	Influenza A virus (A/pigeon/Thailand/VSMU-13-KRI/2005(H5N1)) NA	2462	2462	100%	0.0	100%	
<a href="#">EF178533.1</a>	Influenza A virus (A/tiger/Thailand/VSMU-23-CBI/2004(H5N1)) segm	2462	2462	100%	0.0	100%	
<a href="#">DQ530175.1</a>	Influenza A Virus (A/dog/Thailand-Suphanburi/KU-08/04(H5N1)) neu	2462	2462	100%	0.0	100%	
<a href="#">AY972546.1</a>	Influenza A virus (A/tiger/Thailand/CU-T8/04(H5N1)) neuraminidase	2462	2462	100%	0.0	100%	
<a href="#">AY972545.1</a>	Influenza A virus (A/tiger/Thailand/CU-T6/04(H5N1)) neuraminidase	2462	2462	100%	0.0	100%	
<a href="#">AY972543.1</a>	Influenza A virus (A/tiger/Thailand/CU-T4/04(H5N1)) neuraminidase	2462	2462	100%	0.0	100%	
<a href="#">AY866476.1</a>	Influenza A virus (A/tiger/Thailand/CU-T7/2004(H5N1)) neuraminidase	2462	2462	100%	0.0	100%	
<a href="#">AY842936.1</a>	Influenza A virus (A/tiger/Thailand/CU-T3/2004(H5N1)) neuraminidase	2462	2462	100%	0.0	100%	
<a href="#">EF178507.1</a>	Influenza A virus (A/tree sparrow/Thailand/VSMU-16-RBR/2005(H5N1))	2459	2459	99%	0.0	100%	
<a href="#">DQ083608.1</a>	Influenza A virus (A/chicken/Saraburi/Thailand/CU-27/04(H5N1)) neu	2457	2457	100%	0.0	99%	
<a href="#">AB450608.1</a>	Influenza A virus (A/chicken/Suphanburi/NTAH7618/2004(H5N1)) NA	2451	2451	100%	0.0	99%	
<a href="#">AB450603.1</a>	Influenza A virus (A/chicken/Loei/NTAH2373/2004(H5N1)) NA gene fc	2451	2451	100%	0.0	99%	
<a href="#">EF512562.1</a>	Influenza A virus (A/Prachinburi/6231/2004(H5N1)) segment 6 neura	2451	2451	100%	0.0	99%	
<a href="#">DQ321066.1</a>	Influenza A virus (A/chicken/Malaysia/S858/2004(H5N1)) neuraminid	2451	2451	100%	0.0	99%	
<a href="#">AB450606.1</a>	Influenza A virus (A/chicken/Nakhon Sawan/NTAH01503/2004(H5N1))	2446	2446	100%	0.0	99%	
<a href="#">AB450605.1</a>	Influenza A virus (A/chicken/Nakhon Sawan/NTAH01502/2004(H5N1))	2446	2446	100%	0.0	99%	
<a href="#">AB450602.1</a>	Influenza A virus (A/chicken/NaraThiawat/NTAH1703/2004(H5N1)) NA	2446	2446	100%	0.0	99%	
<a href="#">EF112320.1</a>	Influenza A virus (A/open-billed stork/Nakhonsawan/BBD14211/05(H5N1))	2446	2446	100%	0.0	99%	
<a href="#">DQ989989.1</a>	Influenza A virus (A/open-billed stork/Nakhonsawan/BBD15211/2005(H5N1))	2446	2446	100%	0.0	99%	
<a href="#">AB450604.1</a>	Influenza A virus (A/chicken/Samutprakan/NTAH6604/2004(H5N1)) N	2442	2442	100%	0.0	99%	

# BLAST Output: Alignments

GenBank Accession  
Database identifier of  
SUBJECT

Score of alignment  
Reward for Matches, minus  
penalty for mismatches and gaps

Number (%) of matches (i.e. '|')

Strand orientation of QUERY to  
SUBJECT

Position in QUERY

Sequence of QUERY (gap = '-')

Position in SUBJECT

Sequence of SUBJECT (gap = '-')

Alignment (match = '|', mismatch = ':')

Description

One-line description  
of SUBJECT

E-value of alignment

Probability of such an alignment by chance

The screenshot shows the NCBI Blast results page for a query sequence. The top part of the output displays the subject description: `>|dbj|AB450625.1 Influenza A virus (A/can-bill stork/Thailand/VSMU-29-NSN/2005(H5N1))`. Below this, the alignment statistics are shown: `Score = 2462 bits (1333), Expect = 0.0`, `Identities = 1333/1333 (100%), Gaps = 0/1333 (0%)`, and `Strand=Plus/Plus`. The alignment itself is presented in a table-like format with columns for Query position, Query sequence, Subject position, and Subject sequence. Red arrows from the labels on the left point to the following elements in the screenshot: the subject description, the score and identity statistics, the alignment statistics (Identities and Gaps), the first few lines of the alignment (Query 1, Subject 1), the strand orientation (Plus/Plus), the position in the query (361), the sequence of the query (AATGACAAGCACTCCAAATGGGACTGTCAAAGCAGAACTCTCACAGAACATTAATGAGT), the position in the subject (481), the sequence of the subject (GCAAGTGCCTGCCATGATGCCACAGTTGGTTGACAATTGGAAATTTCTGGCCAGACAAAT), and the alignment symbols (vertical bars for matches and colons for mismatches).

# NCBI GenBank: GenBank format

GenBank Accession  
Database identifier of  
SUBJECT

Description  
One-line description  
of SUBJECT

Source of the sequence

Organism name and taxonomy

Publications using this  
sequence

Source details i.e. strain type,  
country of isolation, etc.

Positional features in the  
sequence i.e. genes, non-  
coding elements, repeats, etc.

Sequence of SUBJECT

## Database record of SUBJECT

The screenshot shows the NCBI Nucleotide database record for Influenza A virus (A/open-bill stork/Thailand/VSMU-29-NSN/2005(H5N1)) NA gene for neuraminidase, complete cds. The record includes the following information:

- LOCUS:** AB450625 1350 bp RNA linear VRL 06-DEC-2008
- DEFINITION:** Influenza A virus (A/open-bill stork/Thailand/VSMU-29-NSN/2005(H5N1)) NA gene for neuraminidase, complete cds.
- ACCESSION:** AB450625
- VERSION:** AB450625.1
- GI:** 210144900
- KEYWORDS:** .
- SOURCE:** Influenza A virus (A/open-bill stork/Thailand/VSMU-29-NSN/2005(H5N1))
- ORGANISM:** Influenza A virus (A/open-bill stork/Thailand/VSMU-29-NSN/2005(H5N1)); Viruses; ssRNA negative-strand viruses; Orthomyxoviridae; Influenzavirus A.
- REFERENCE:** 1. Uchida, Y., Chaichoune, K., Wiriyarat, W., Watanabe, C., Hayashi, T., Patchimasiri, T., Nuansrichay, B., Parchariyanon, S., Okamatsu, M., Tsukamoto, K., Takemae, H., Ratanakorn, P., Yamaguchi, S. and Saito, T. Molecular epidemiological analysis of highly pathogenic avian influenza H5N1 subtype isolated from poultry and wild bird in Thailand. *Virus Res.* 138 (1-2), 70-80 (2008). [PubMed 18801394](#)
- REMARK:** Publication\_Status: Available-Online
- REFERENCE:** 2 (bases 1 to 1350)
- AUTHORS:** Uchida, Y. and Saito, T.
- TITLE:** Direct Submission
- JOURNAL:** Submitted (30-JUL-2008) Contact:Yuko Uchida National Institute of Animal Health; Kannondai 3-1-5, Tsukuba 305-0856, Japan

**FEATURES:**

- source**  
1..1350  
/organism="Influenza A virus (A/open-bill stork/Thailand/VSMU-29-NSN/2005(H5N1))"  
/mol\_type="genomic RNA"  
/strain="A/open-bill stork/Thailand/VSMU-29-NSN/2005"  
/serotype="H5N1"  
/db\_xref="taxon:551274"  
/country="Thailand"
- gene**  
1..1350  
/gene="NA"
- CDS**  
1..1350  
/gene="NA"  
/codon\_start=1  
/product="neuraminidase"  
/protein\_id="BA080861.1"  
/db\_xref="GI:210144901"  
/translation="MNPNKLIITIGSICMTGMSLMLQIGNLISIWWSHSIHTGNQH  
KAEPISINLLTEKIVASVKLAGNSLCPINGNAVYKSDNSIRIGSKDVFVIREPFI  
SCSHLECRIFFLTQALLNDKHSNGTVKDRSPHRTLMSCPVGEAPSPYNSRFESVANS  
ASACHDGI SWLTIIGISGPDNGAVLVLYNGIITDIKSWRNILRLTQESACACVNSC  
FTVMTDGFSGNQASHKIFPKMEKGVVKSVELDAPNYHYEECSCTPDAGEITCVCRDNW  
HGSNRFVWSFNQMLEYQIGYICGVFGDNPFRNDGAGSCGFPVSNNGAYGVKGFSEKYG  
NGVWIGRTASTNSRSGFEMWDENGTETDSSFVSKQDIVAITDWSGYSGSFVQHPEL  
TGLDCIRPCFWVELINGRKESTIWTSGSSISFCVNSDVTGVNSWDFDGAELFTIDK"

**ORIGIN**  
1 ATGATCCCAATATAGAGAGATATAGACAGCAGGACAGTCTGATAGGATGACAGGATAGGAT

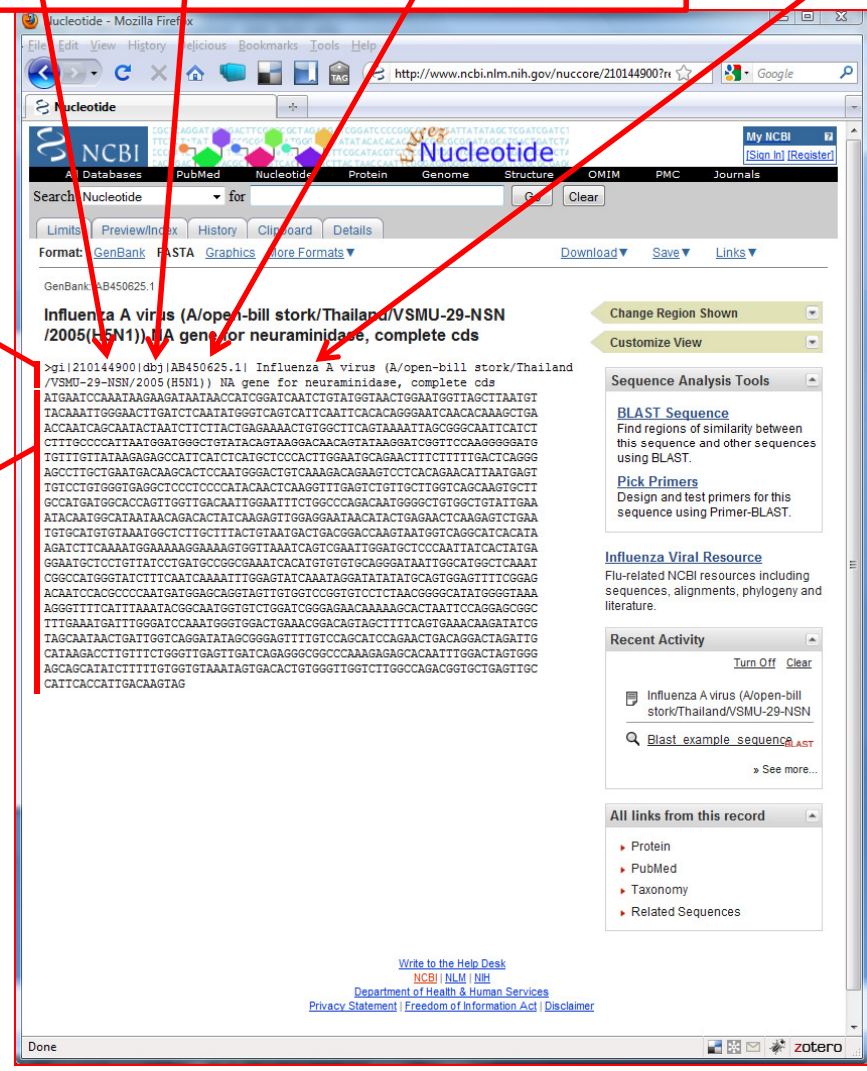
# NCBI GenBank: FASTA format

GenBank id Database    GenBank Accession  
Database identifier of  
SUBJECT

Description  
One-line description  
of SUBJECT

FASTA Header  
One line AND starts with a '>' followed by identifier/name and OPTINALLY a short description

FASTA Sequence  
Maximum 80 characters per line





# CLUSTAL: Input

## Alignment: full or fast?

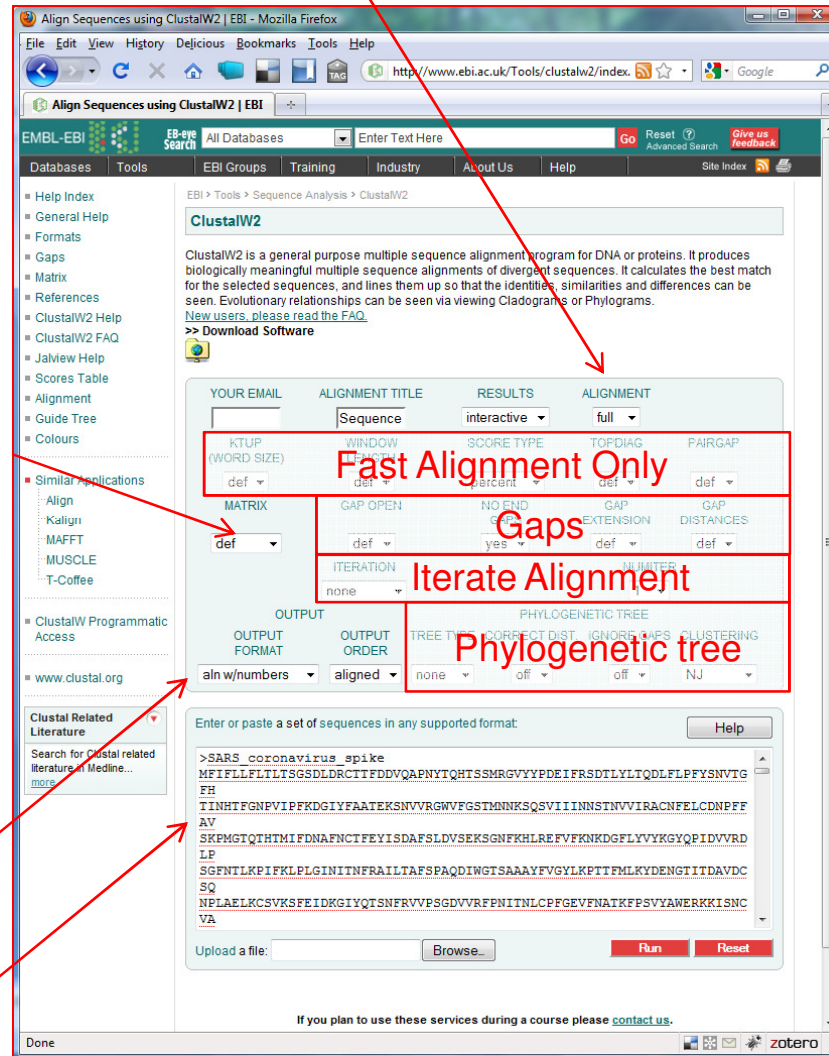
'Fast' for quick peek at alignment, but choose 'full' for best result.

## Substitution matrix

Determines the similarity of two amino acids (next slide)

Format Output

Input Sequences (FASTA)



## Fast Alignment Only

- for large number of sequences
- makes CLUSTAL behave more like BLAST ... runs quicker.

## Gaps

- Penalty to start a gap?
- Penalize edge gaps?
- Penalty to start a gap to make it longer?
- Penalty to separate gaps?

## Iterate alignment

Align sequences multiple times under different conditions to get 'optimal' alignment

## Phylogenetic tree

- Input aligned sequences
- Output type (nj, phylip, dist)
- Correct for multiple substitutions (divergent sequences only).
- Ignore gaps (i.e. possibly ambiguous parts of alignment)
- Algorithm:
  - NJ (branch tips proportional to change)
  - UPGMA (branch tips equal).



# CLUSTAL: Alignment Output

Rows: Input Sequences

Columns: Aligned amino acids

```

Human_coronavirus_NL63_spike      -----MKLFLILLVL 10
Human_coronavirus_229E_spike     -----
Porcine_epidemic_diarrhea_viru   -----MRSLIYFWLLLPVLP TLSLPQDVTRC 26
Transmissible_gastroenteritis_  MKKLFVVLVVMPLIYGDNFPCSKLTNRTIGNQWNLIETFLNYSRLPPN 50
Human_coronavirus_OC43_spike    -----MFLILLISLPTAFAVIGDL 19
Bovine_coronavirus_spike       -----MFLILLISLPTAFAVIGDL 19
Murine_hepatitis_virus_spike    -----MLFVFILFLP SCLGYIGDF 19
Human_coronavirus_HKU1_spike    -----MLLIIFILPTTLAVIG 16
SARS_coronavirus_spike         -----MFIFLLFLTITSG 13
Avian_infectious_bronchitis_vi  -----
    
```

End Gap

Sequence Positions

Internal Gap

Conserved?

Human\_coronavirus\_NL63\_spike  
Human\_coronavirus\_229E\_spike

AVFPMILW	RED	Small (small+ hydrophobic (incl.aromatic -Y))
DE	BLUE	Acidic
RK	MAGENTA	Basic
STYHCNGQ	GREEN	Hydroxyl + Amine + Basic - Q
Others	Gray	

```

SGIREFSNLVLN NCTKYNIYDYVGTGIIR-----SSNQSLAGGIT--YVS 656
EGVSSFMNVILDKCTKYNIYDVSGVGVIR-----VSNDTFLNGIT--YTS 475
EGITDVSFMTLDVCTKYTIYGFKGEGII T-----LTNSSILAGVY--YTS 675
SGVHDL SVLHLD SCDYNIYGRITGVGIIR-----QTNRTLLSGLY--YTS 713
DLQKANTDIILGVCVNYDLYGILGQGFIV-----EVNATYYNSWQNLLYD 660
DLQKSNTDIILGVCVNYDLYGITGQGFIV-----EVNATYYNSWQNLLYD 662
DLQLPNTDEVVTGICV KYDLYGITGQGVFK-----EVKADYYNSWQTLLYD 611
DLLQPNTDEVFTDVCVDYDLYGITGQGFIV-----EVSAVYYNSWQNLLYD 652
CGPKLSTDLIKNQCVNFENGLTGTGVLT-----PSSK-RFQPFQQFGRD 554
ITQNNYNNITLNTCVDYNIYGRITGQGFITNVTDSAVSYNYLADAGLAILD 467
: . * . : . * * : . .
    
```

Identical

Conserved

Semi-Conserved

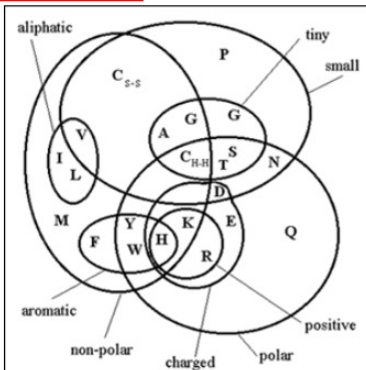


Figure 1. A Venn diagram showing the relationship of the 20 naturally occurring amino acids to a selection of physico-chemical properties thought to be important in the determination of protein structure.

Relationship between sequences ... groups?

# CLUSTAL: Tree Output (phylip format) and FigTree: Phylogenetic tree Viewer

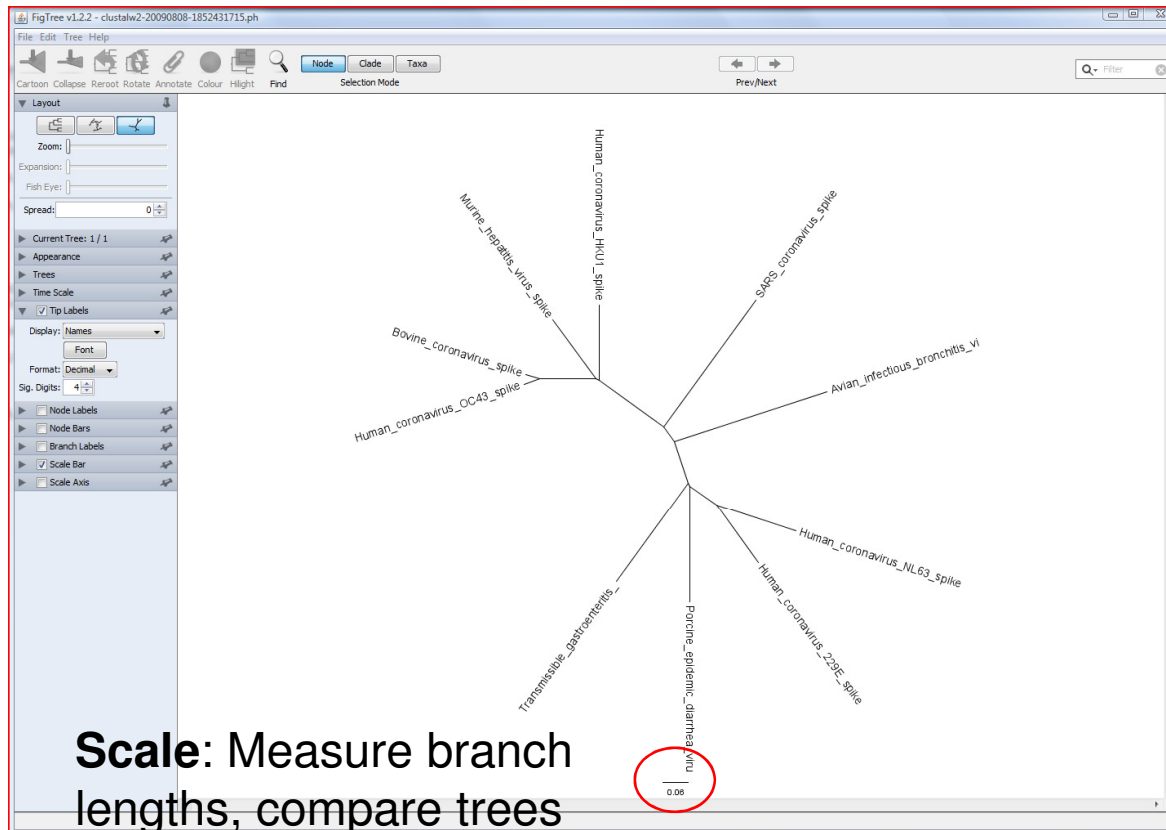
## Phylip format:

- Specifies pairwise relationship between sequences (grouping)
- Branch length indicates amount of change
- Not human readable

## Figtree

- Visually represent phylip data as tree.
- Different types of trees: rooted, unrooted

```
(  
(  
(  
Human_coronavirus_NL63_spike:0.19282,  
Human_coronavirus_229E_spike:0.16396)  
:0.07664,  
Porcine_epidemic_diarrhea_viru:0.26760)  
:0.00763,  
Transmissible_gastroenteritis_:0.28022,  
(  
(  
(  
(  
Human_coronavirus_OC43_spike:0.03789,  
Bovine_coronavirus_spike:0.03391)  
:0.12957,  
Murine_hepatitis_virus_spike:0.16837)  
:0.00956,  
Human_coronavirus_HKU1_spike:0.17385)  
:0.18285,  
SARS_coronavirus_spike:0.36212)  
:0.04149,  
Avian_infectious_bronchitis_vi:0.37160)  
:0.10239);
```



1. Save to 'ph' file,
2. Open w/ FigTree
3. Select right tree

What do ...

Groupings tell us?

Branch Lengths tell us?

# ORFFinder: Input

## Goal:

- 1) Find possible coding sequences
- 2) BLAST coding sequences
- 3) Infer function of coding sequences

Genomic sequence  
in FASTA format

Genetic code

ORF Finder (Open Reading Frame Finder)

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database. This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software.

Enter GI or ACCESSION

or sequence in FASTA format

```
>gi|9626243|ref|NC_001416.1|
Bacteriophage lambda, complete
genome
GGGCGGCGACCTCGCGGGTTTTTCGCTATTTATGAA
AAATTTCCGGTTAAGGCGTTTCGGTTCITCTCG
TCATAACTTAATGTTTTTATTTAAATACCCCTCG
AAAAGAAAGGAACGACAGGTGCTGAAAGCGAGGC
TTTTGGCCCTCIGTCGTTTCCTTTCIGTTTTIG
TCCGTGGAATGAACAATGGAAGTCACAAAAGCA
```

FROM:  TO:

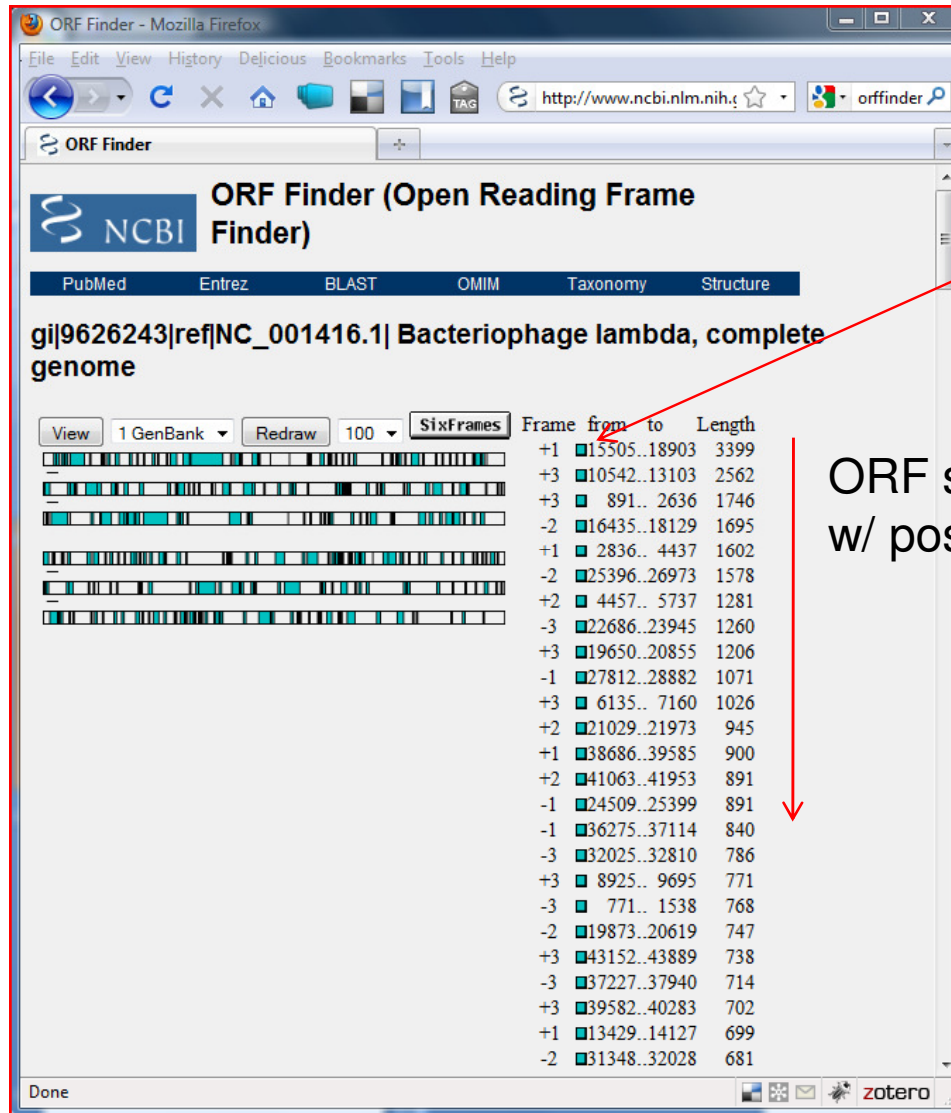
Genetic codes

1 Standard

Comments and suggestions to: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov)  
Credits to: [Tatiana Tatusov](#) and [Roman Tatusov](#)

# ORFFinder: Output

Graphic of ORF locations (light blue)



Click box to get ORF detail

ORF sorted by length w/ positions

## ORFFinder: Output

BLAST protein sequence:  
Inspect GenBank record  
and report pertinent  
information

Access FASTA formatted  
sequence

ORF Finder (Open Reading Frame Finder)

PubMed Entrez BLAST OMIM Taxonomy Structure

gi|9626243|ref|NC\_001416.1| Bacteriophage lambda, complete genome

Program: blastp Database: nr BLAST with parameters Cognitor

View	1 GenBank	Redraw	100	SixFrames	Frame from to	Length
					+1 15505..18903	3399
					+3 10542..13103	2562
					+3 891.. 2636	1746
					-2 16435..18129	1695
					+1 2836.. 4437	1602
					-2 25396..26973	1578
					+2 4457.. 5737	1281
					-3 22686..23945	1260
					+3 19650..20855	1206
					-1 27812..28882	1071
					+3 6135.. 7160	1026
					+2 21029..21973	945
					+1 38686..39585	900
					+2 41063..41953	891
					-1 24509..25399	891
					-1 36275..37114	840
					-3 32025..32810	786

Length: 1132 aa

Accept Alternative Initiation Codons

```
15505 atgggtaaaggaagcagtaaggggcatabccccgcgcgaagcgaag
M G K G S S K G H T P R E A K
15550 gacaacctgaagtccaagcagttgctgagtgatcgatgccatc
D N L K S T Q L L S V I D A I
15595 agcgaaggccgattgaaaggtccggtagtgcttaaaaagcgtg
S E G P I E G F V D G L K S V
15640 ctgctgaacagtaagccgggtgctggacaactgaggggaaataccaac
L L N S T P V L D T E G N T N
15685 atatccgggtgtaacgggtggttccgggctggtagcaggaagcag
I S G V T V V F R A G E O E O
```

Genomic and amino acid  
sequence of ORF