BLAST Input: Select a search type

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

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

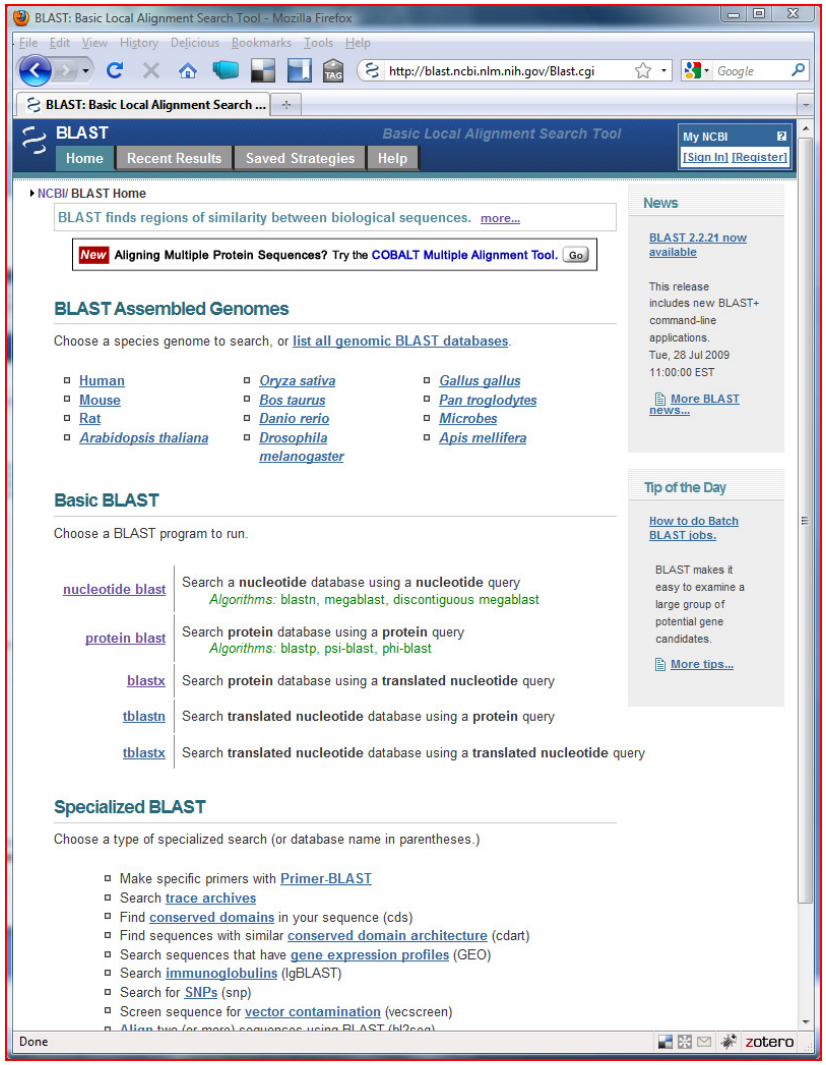
B.I.P.

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



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, discontiguous megablast, and blastn for nucleotide; and blastp, PSI-BLAST and PHI-BLAST for protein.

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

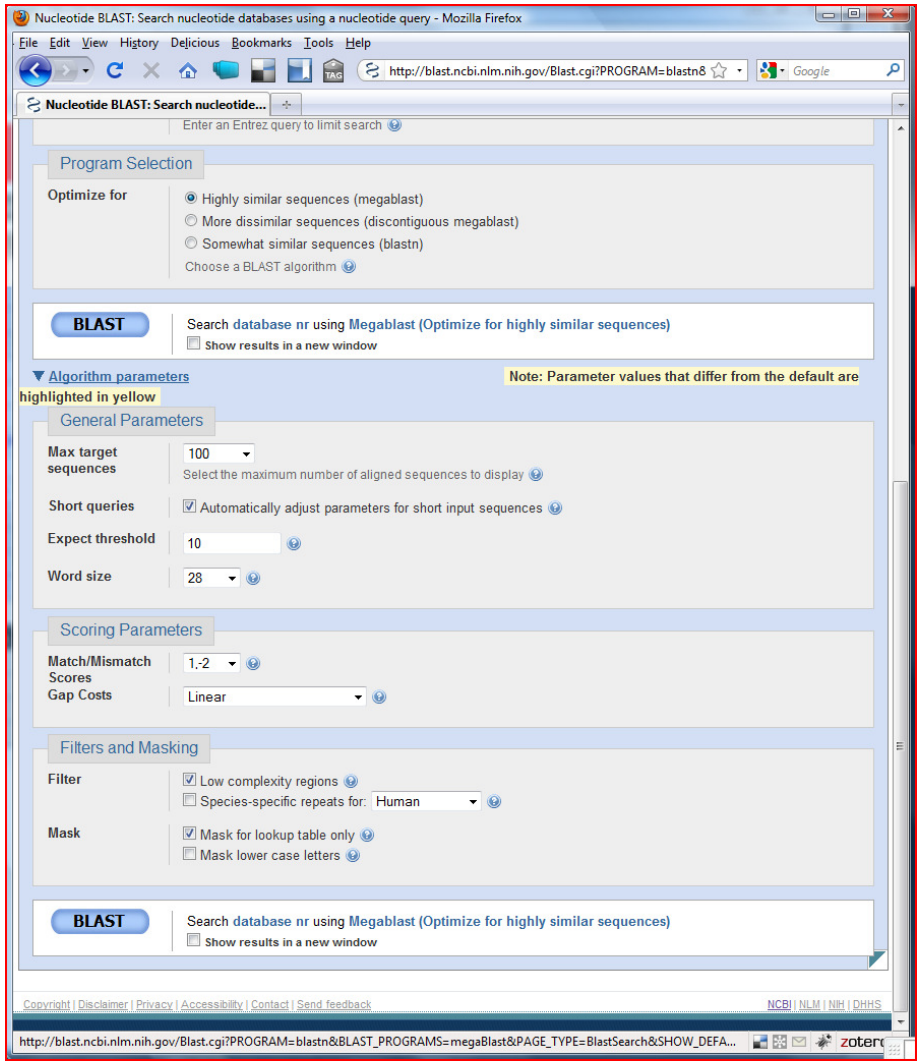
Cost to create and extend a gap in an alignment. Linear costs are available only with megablast and are determined by the match/mismatch scores.

Filter

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

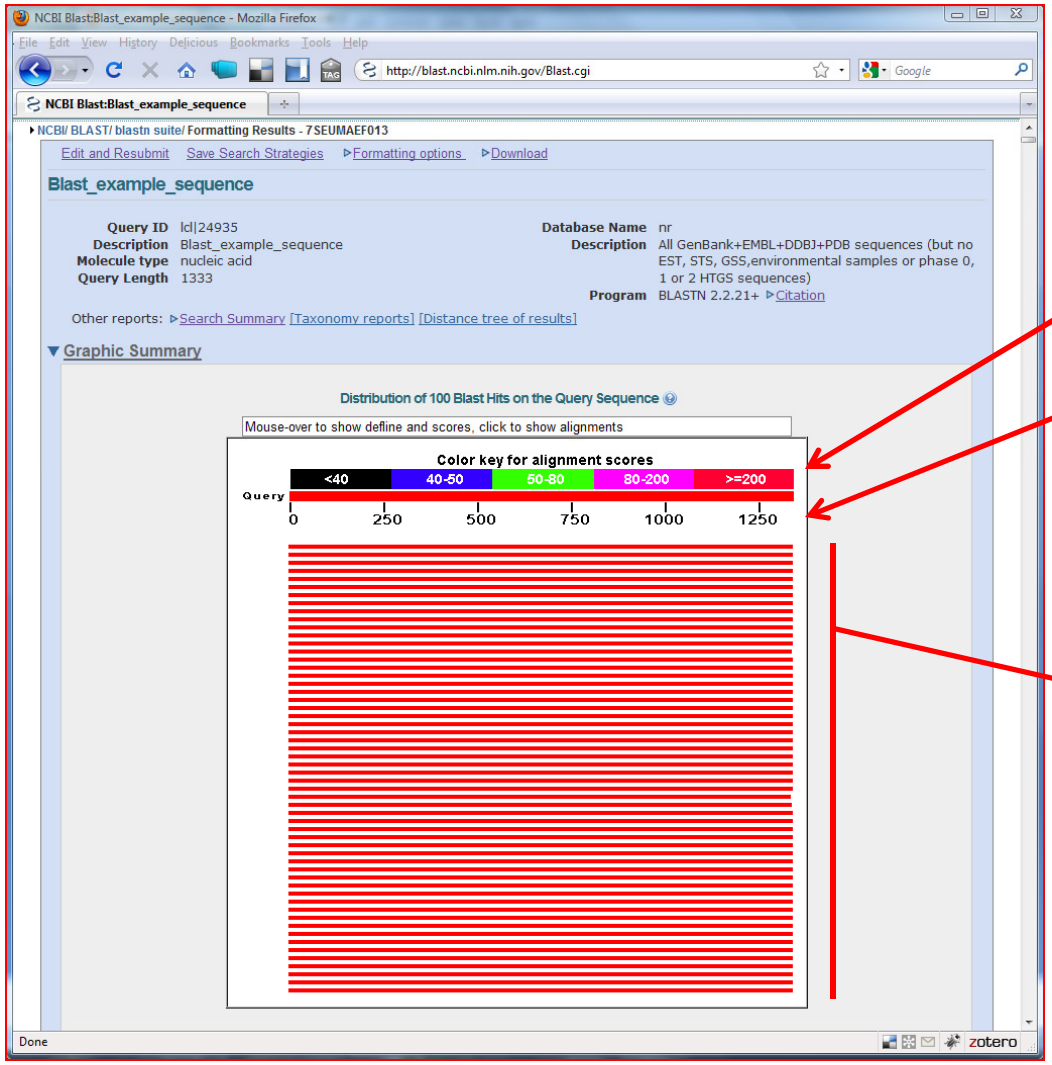
Mask

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

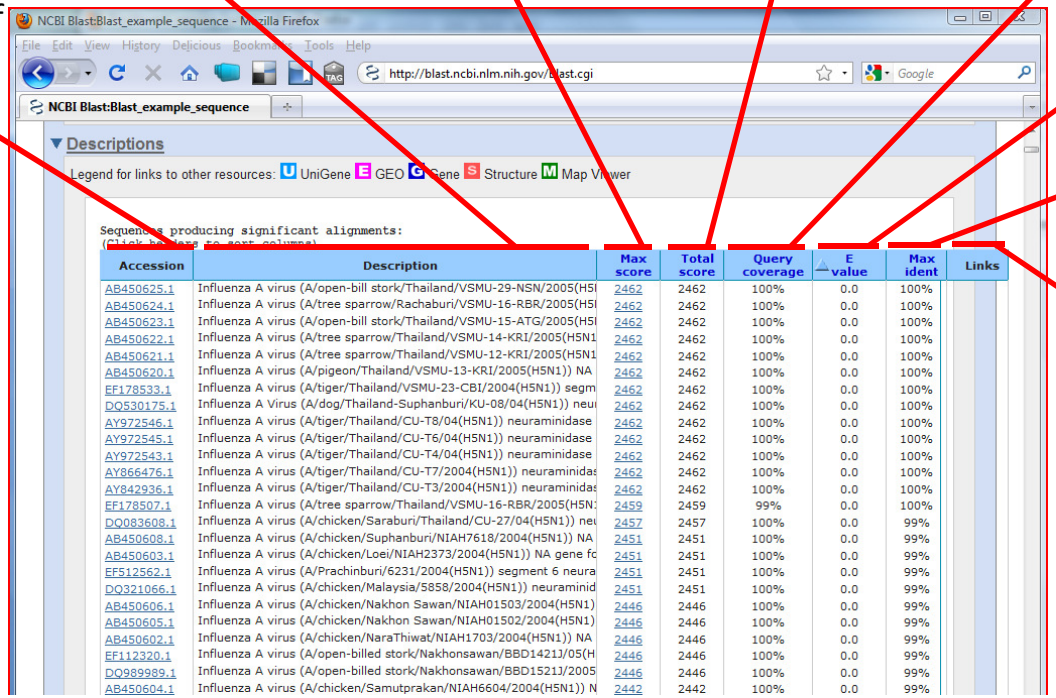
Description
One-line description of SUBJECT

Max Score
Score for highest scoring alignment to SUBJECT

Total Score
Sum of scores for all alignmments to SUBJECT

Query Coverage
% of QUERY covered by alignmnets

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

BLAST Output: Alignments

Description
One-line description of SUBJECT

E-value of alignment
Probability of such an alignment by chance

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

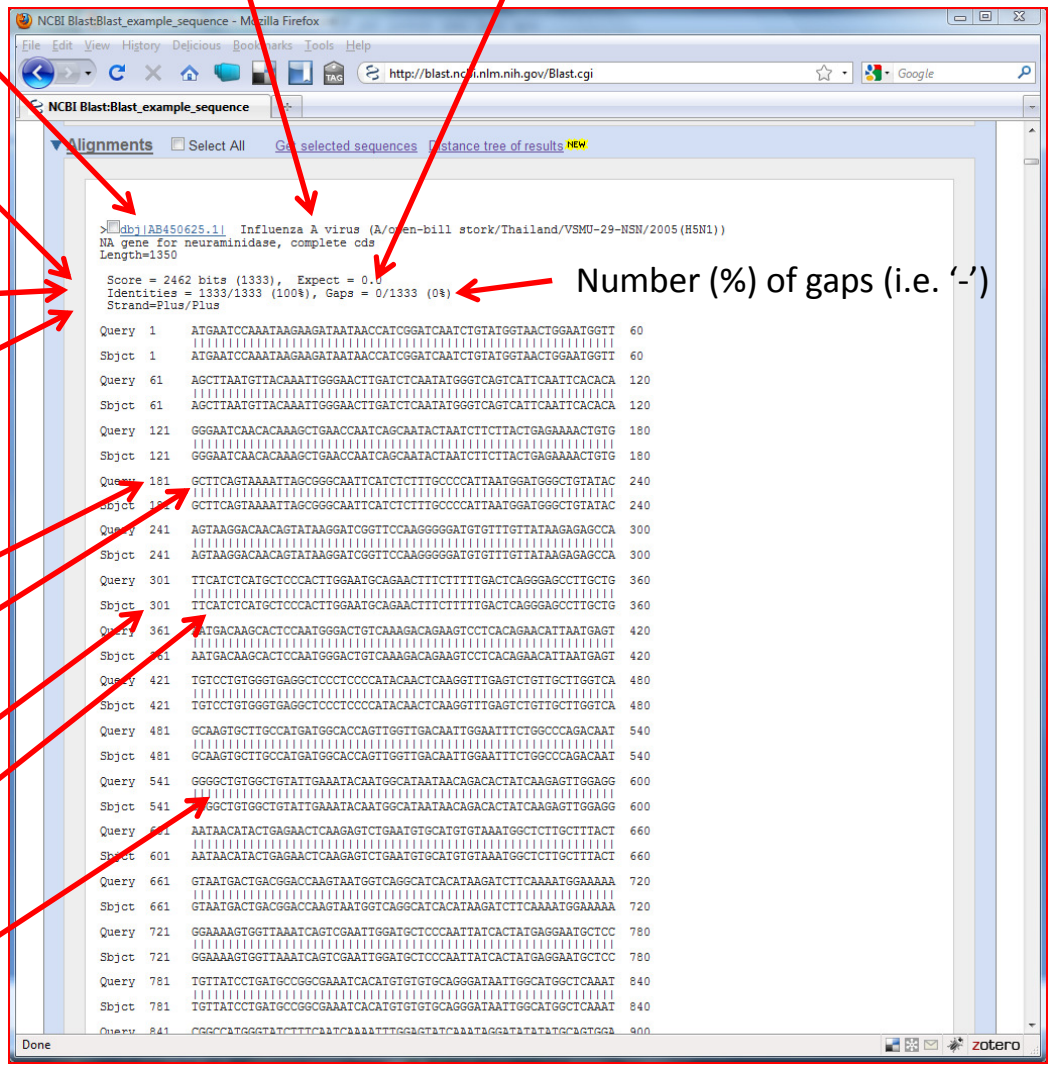
Sequence of QUERY (gap = '-')

Position in SUBJECT

Sequence of SUBJECT (gap = '-')

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

Number (%) of gaps (i.e. '-')



NCBI GenBank: GenBank format

GenBank Accession
Database identifier of SUBJECT

Description
One-line description of SUBJECT

Source of the sequence

Organism name and taxonomy

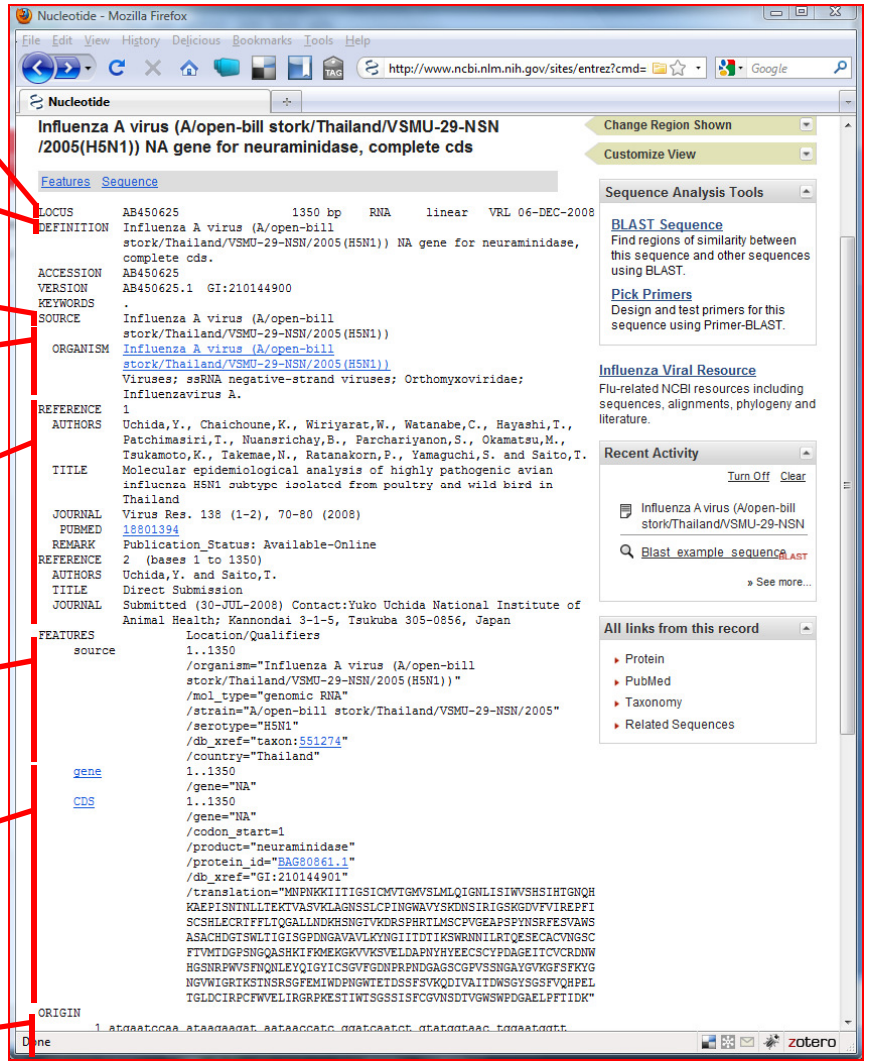
Publications using this sequence

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



NCBI GenBank: FASTA format

GenBank id Database GenBank Accession

Database identifier of SUBJECT

Description

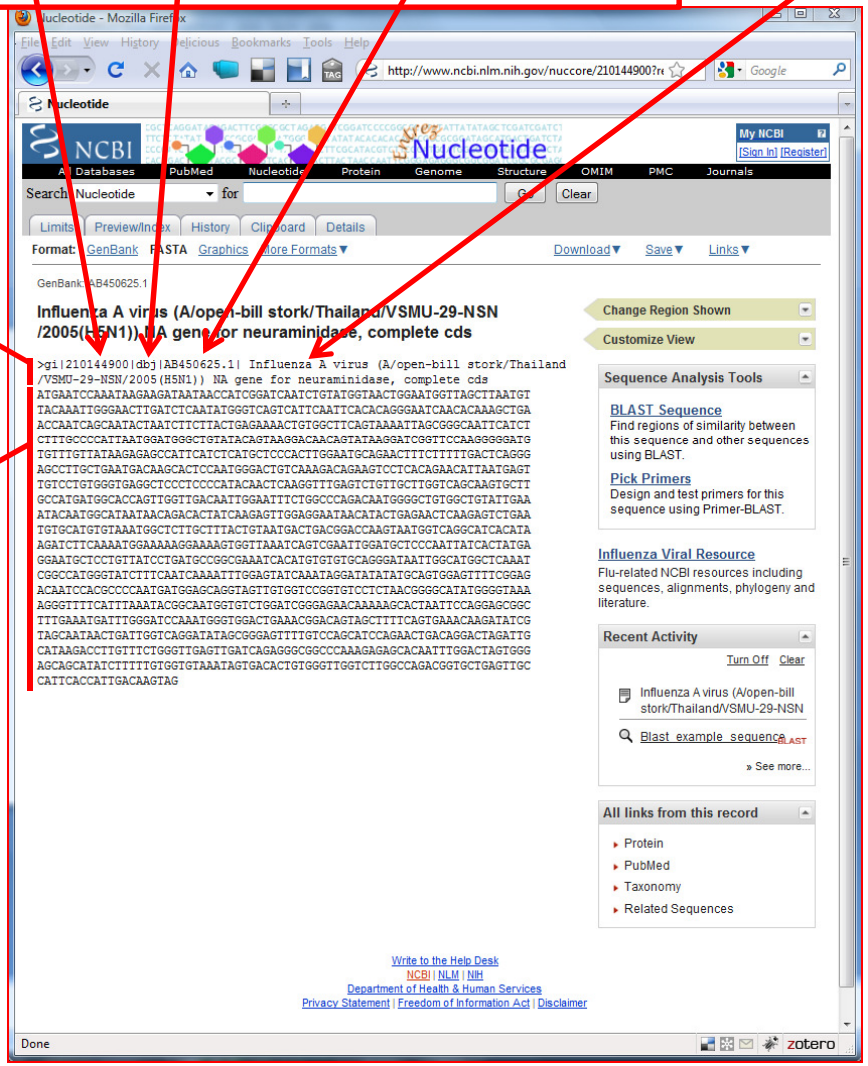
One-line description of SUBJECT

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?

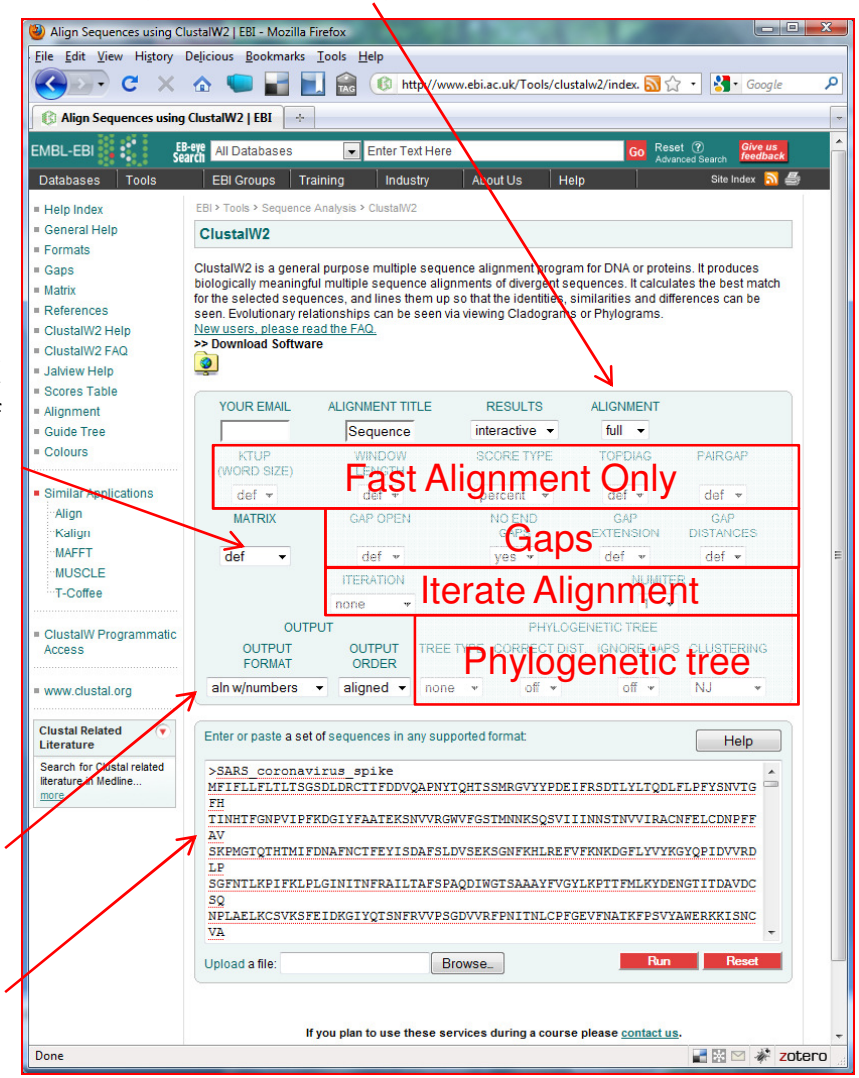
'Fast' for quick peek at alignment, but choose 'full' for best result.

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).

Amino Acid Substitution Matrix

Used to determine the **similarity of non-identical amino acids**

Based c**hemical similarity:**
Hydro**phobic** >> Hydro**phobic** is more **conservative** (similar) than Hydro**phobic** >> Hydro**philic**

There are **3 types** of matrices (**PAM, BLOSUM, GONNET**):
- Date created, 1970s onward
- Different datasets (protein sequences) and/or,
- Evolutionary models.

Each matrix has multiple versions that **model different evolutionary distances** i.e. low to high sequence divergence.

BLOSUM (Henikoff). Best available for database similarity (homology searches) i.e. BLAST.
Versions: Blosum80 (high), 62, 40 and 30 (low).

PAM (Dayhoff). Legacy matrices from the 1970s. Versions: PAM 120 (low), 160, 250 and 350 (high).

Gonnet. Same as PAM, but are much more up to date and are based on a far larger data set.
Versions: GONNET 40 (low), 80, 120, 160, 250 and 350 (high).

Closely related sequences:
- BLOSUM w/ higher numbers or,
- PAM matrices w/ lower numbers

PAM20 – Highly similar sequences

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	6	-8	-5	-4	-8	-5	-3	-3	-8	-6	-7	-8	-6	-9	-2	-1	-1	-16	-9	-3	-5	-4	-4	-19
R	-8	9	-7	-12	-9	-2	-11	-11	-3	-6	-10	-1	-5	-10	-5	-4	-8	-3	-11	-9	-9	-5	-7	-19
N	-5	-7	8	1	-13	-5	-3	-4	-1	-6	-8	-2	-11	-10	-7	-1	-3	-9	-5	-9	6	-4	-4	-19
D	-4	-12	1	8	-16	-4	2	-4	-5	-9	-15	-6	-13	-17	-9	-5	-6	-17	-13	-9	6	0	-7	-19
C	-8	-9	-13	-16	10	-16	-16	-11	-8	-7	-17	-16	-16	-15	-9	-4	-9	-18	-5	-7	-14	-16	-11	-19
Q	-5	-2	-5	-4	-16	9	0	-8	0	-9	-6	-4	-5	-15	-4	-6	-7	-15	-14	-8	-4	7	-6	-19
E	-3	-11	-3	2	-16	0	8	-5	-6	-6	-10	-5	-8	-16	-7	-5	-7	-19	-9	-8	0	6	-6	-19
G	-3	-11	-4	-4	-11	-8	-5	7	-10	-13	-12	-8	-10	-10	-7	-3	-7	-17	-16	-7	-4	-6	-6	-19
H	-8	-3	-1	-5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-5	-7	-8	-8	-4	-7	-2	-2	-6	-19
I	-6	-6	-6	-9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-10	-8	-3	-16	-7	1	-7	-7	-6	-19
L	-7	-10	-8	-15	-17	-6	-10	-12	-7	-2	7	-9	0	-4	-8	-9	-8	-7	-8	-3	-10	-8	-7	-19
K	-8	-1	-2	-6	-16	-4	-5	-8	-8	-7	-9	7	-3	-16	-8	-5	-4	-14	-10	-10	-3	-5	-6	-19
M	-6	-5	-11	-13	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-2	-12	-6	-6	-19
F	-9	-10	-10	-17	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-9	-12	-16	-9	-19
P	-2	-5	-7	-9	-9	-4	-7	-7	-5	-10	-8	-8	-9	-11	8	-3	-5	-16	-16	-7	-8	-5	-6	-19
S	-1	-4	-1	-5	-4	-6	-5	-3	-7	-8	-9	-5	-6	-7	-3	7	0	-6	-8	-8	-2	-6	-4	-19
T	-1	-8	-3	-6	-9	-7	-7	-7	-8	-3	-8	-4	-5	-10	-5	0	7	-15	-7	-4	-4	-7	-5	-19
W	-16	-3	-9	-17	-18	-15	-19	-17	-8	-16	-7	-14	-15	-6	-16	-6	-15	13	-6	-18	-11	-17	-13	-19
Y	-9	-11	-5	-13	-5	-14	-9	-16	-4	-7	-8	-10	-13	1	-16	-8	-7	-6	10	-8	-7	-11	-9	-19
V	-3	-9	-9	-9	-7	-8	-8	-7	-7	1	-3	-10	-2	-9	-7	-8	-4	-18	-8	7	-9	-8	-6	-19
B	-5	-9	6	6	-14	-4	0	-4	-2	-7	-10	-3	-12	-12	-8	-2	-4	-11	-7	-9	6	-1	-6	-19
Z	-4	-5	-4	0	-16	7	6	-6	-2	-7	-8	-5	-6	-16	-5	-6	-7	-17	-11	-8	-1	6	-6	-19
X	-4	-7	-4	-7	-11	-6	-6	-6	-6	-6	-7	-6	-6	-9	-6	-4	-5	-13	-9	-6	-6	-6	-6	-19
*	-19	-19	-19	-19	-19	-19	-19	-19	-19	-19	-19	-19	-19	-19	-19	-19	-19	-19	-19	-19	-19	-19	-19	1

More negative,
penalize changes more

PAM350 – Highly Divergent sequences

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	2	-1	0	1	-2	0	1	2	-1	0	-2	-1	-1	-4	1	1	1	-7	-4	0	1	0	0	-10
R	-1	7	1	-1	-4	2	0	-2	2	-2	-3	4	0	-5	0	0	-1	4	-5	-3	0	1	-1	-10
N	0	1	2	2	-4	1	2	1	2	-2	-3	1	-2	-4	0	1	1	-5	-3	-2	2	2	0	-10
D	1	-1	2	4	-6	2	4	1	1	-2	-4	1	-3	-6	0	1	0	-8	-5	-2	3	3	-1	-10
C	-2	-4	-4	-6	18	-6	-6	-4	-4	-3	-7	-6	-6	-5	-3	0	-2	-10	1	-2	-5	-6	-3	-10
Q	0	2	1	2	-6	4	3	-1	3	-2	-2	1	-1	-5	1	0	0	-5	-5	-2	2	3	0	-10
E	1	0	2	4	-6	3	4	1	1	-2	-4	0	-2	-6	0	0	0	-8	-5	-2	3	3	0	-10
G	2	-2	1	1	-4	-1	1	5	-2	-2	-4	-1	-3	-6	0	1	1	-8	-6	-1	1	0	-1	-10
H	-1	2	2	1	-4	3	1	-2	7	-2	-2	[illegible]	-2	-2	0	-1	-1	-3	0	-2	1	2	0	-10
I	0	-2	-2	-2	-3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-2	-1	0	-6	0	4	-2	-2	0	-10
L	-2	-3	-3	-4	-7	-2	-4	-4	-2	[illegible]	[illegible]	-3	5	[illegible]	-3	-3	-2	-2	0	3	-4	-3	-1	-10
K	-1	4	1	1	-6	1	0	-1	1	-2	-3	5	0	-6	-1	0	0	-4	-5	-2	1	1	-1	-10
M	-1	0	-2	-3	-6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-5	[illegible]	2	-2	-2	0	-10
F	-4	-5	-4	-6	-5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-1	-5	-6	-2	-10
P	1	0	0	0	-3	1	0	0	0	-2	-3	-1	-2	-5	6	1	1	-7	-6	-1	0	0	0	-10
S	1	0	1	1	0	0	0	1	-1	-1	-3	0	-2	-4	1	1	1	-3	-3	-1	1	0	0	-10
T	1	-1	1	0	-2	0	0	1	-1	0	-2	0	-1	-3	1	1	2	-6	-3	0	0	0	0	-10
W	-7	4	-5	-8	-10	-5	-8	-8	-3	-6	-2	-4	-5	1	-7	-3	-6	27	1	-7	-6	-7	-5	-10
Y	-4	-5	-3	-5	1	-5	-5	-6	0	0	0	-5	-2	11	-6	-3	-3	1	14	-2	-4	-5	-2	-10
V	0	-3	-2	-2	-2	-2	-2	-1	-2	4	3	-2	2	-1	-1	-1	0	-7	-2	5	-2	-2	0	-10
B	1	0	2	3	-5	2	3	1	1	-2	-4	1	-2	-5	0	1	0	-6	-4	-2	3	2	0	-10
Z	0	1	2	3	-6	3	3	0	2	-2	-3	1	-2	-6	0	0	0	-7	-5	-2	2	3	0	-10
X	0	-1	0	-1	-3	0	0	-1	0	0	-1	-1	0	-2	0	0	0	-5	-2	0	0	0	-1	-10
*	-10	-10	-10	-10	-10	-10	-10	-10	-10	-10	-10	-10	-10	-10	-10	-10	-10	-10	-10	-10	-10	-10	-10	1

Less negative,
penalize changes less

CLUSTAL: Alignment Output

Rows: Input Sequences
Columns: Aligned amino acids

End Gap

Sequence Positions

```
Human_coronavirus_NL63_spike          ----------------------------------------MKLFLILLVL 10
Human_coronavirus_229E_spike          --------------------------------------------------
Porcine_epidemic_diarrhea_viru        ------------------------MRSLIYFWLLLPVLPTLSLPQDVTRC 26
Transmissible_gastroenteritis_        MKKLFVVLVVMPLIYGDNFPCSKLTNRTIGNQWNLIETFLLNYSSRLPPN 50
Human_coronavirus_OC43_spike          --------------------------------MFLILLISLPTAFAVIGDL 19
Bovine_coronavirus_spike              --------------------------------MFLILLISLPTAFAVIGDL 19
Murine_hepatitis_virus_spike          --------------------------------MLFVFILFLPSCLGYIGDF 19
Human_coronavirus_HKU1_spike          -----------------------------------MLLIIFILPTTLAVIG 16
SARS_coronavirus_spike                -------------------------------------MFIFLLFLTLTSG 13
Avian_infectious_bronchitis_vi        --------------------------------------------------
```

Internal Gap

```
SGIREFSNLVLNNCTKYNIYDYVGTGIIR-----SSNQSLAGGIT--YVS 656
EGVSSFMNVTLDKCTKYNIYDVSGVGVIR-----VSNDTFLNGIT--YTS 475
EGITDVSFMTLDVCTKYTIYGFKGEGIIT-----LTNSSILAGVY--YTS 675
SGVHDLSVLHLDSCTDYNIYGRTGVGIIR-----QTNRTLLSGLY--YTS 713
DLQKANTDIILGVCVNYDLYGILGQGIFV-----EVNATYYNSWQNLLYD 660
DLQKSNTDIILGVCVNYDLYGITGQGIFV-----EVNATYYNSWQNLLYD 662
DLQLPNTEVVTGICVKYDLYGITGQGVFK-----EVKADYYNSWQTLLYD 611
DLLQPNTEVFTDVCVDYDLYGITGQGIFK-----EVSAVYYNSWQNLLYD 652
CGPKLSTDLIKNQCVNFNFNGLTGTGVLT-----PSSK-RFQPFQQFGRD 554
ITQNNYNNITLNTCVDYNIYGRTGQGFITNVTDSAVSYNYLADAGLAILD 467
       :  . *..: : .  * *.:        .           .
```

Identical

Conserved

Semi-Conserved

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	

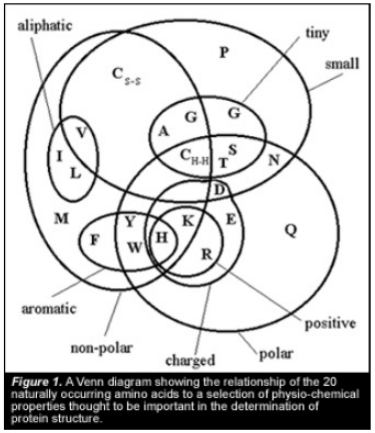


Figure 1. A Venn diagram showing the relationship of the 20 naturally occurring amino acids to a selection of physio-chemical properties thought to be important in the determination of protein structure.

Relationship between sequences … groups?

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

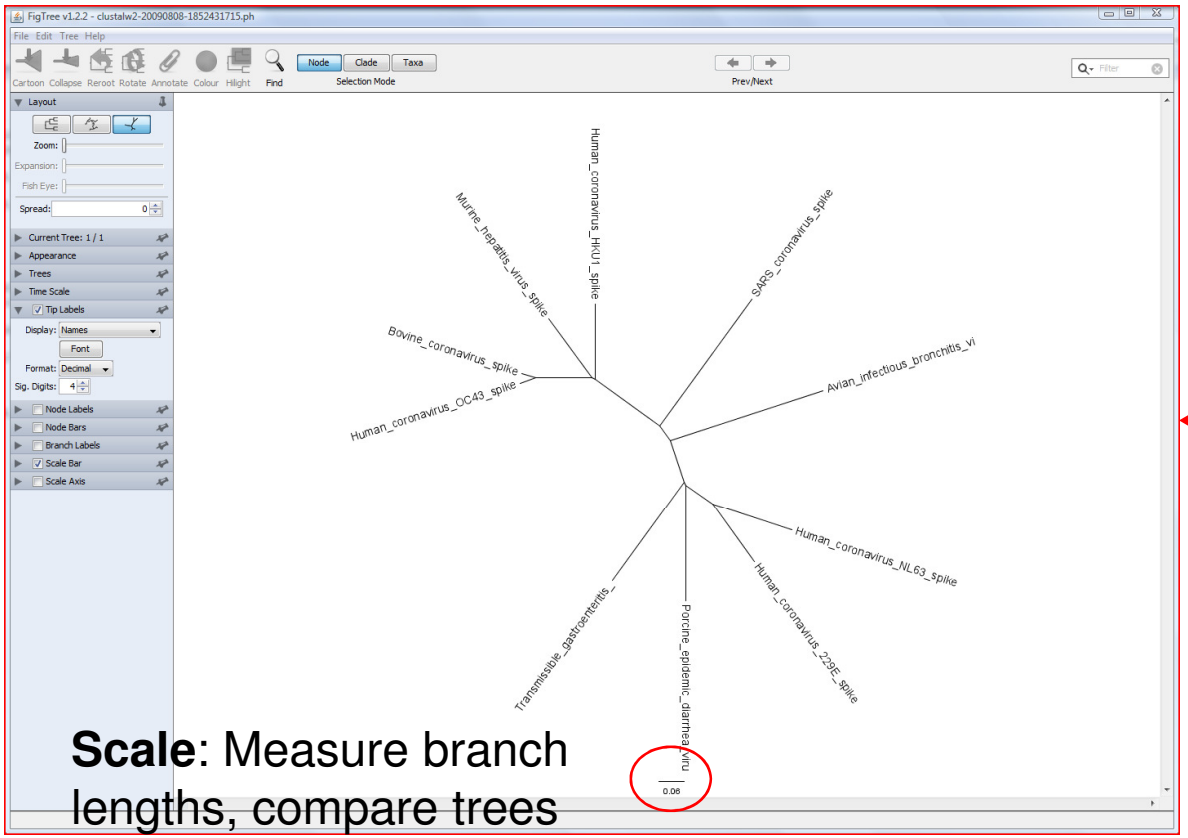
Phylip format:

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

```
(
(
(
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);
```

Figtree

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



Scale: Measure branch lengths, compare trees

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

What do …
Groupings tell us?
Branch Lengths tell us?

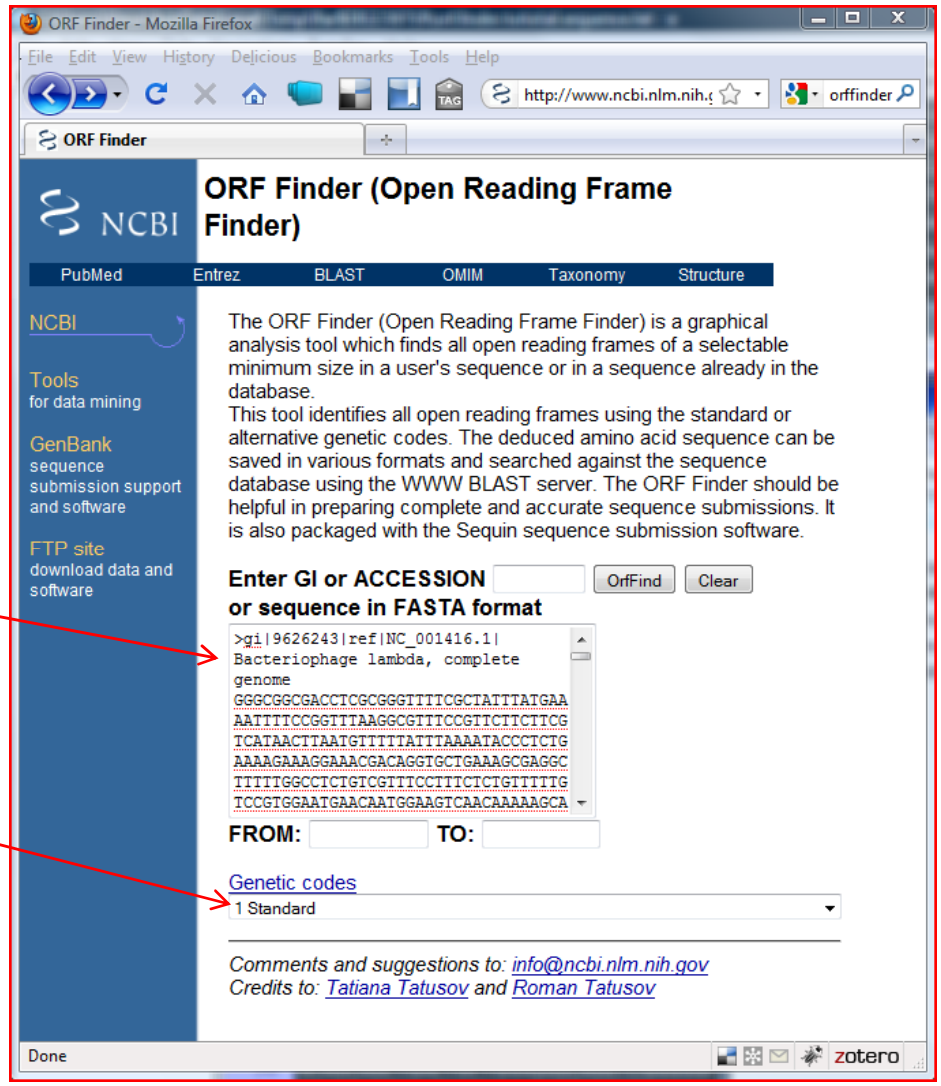
ORFFinder: Input

Goal:

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

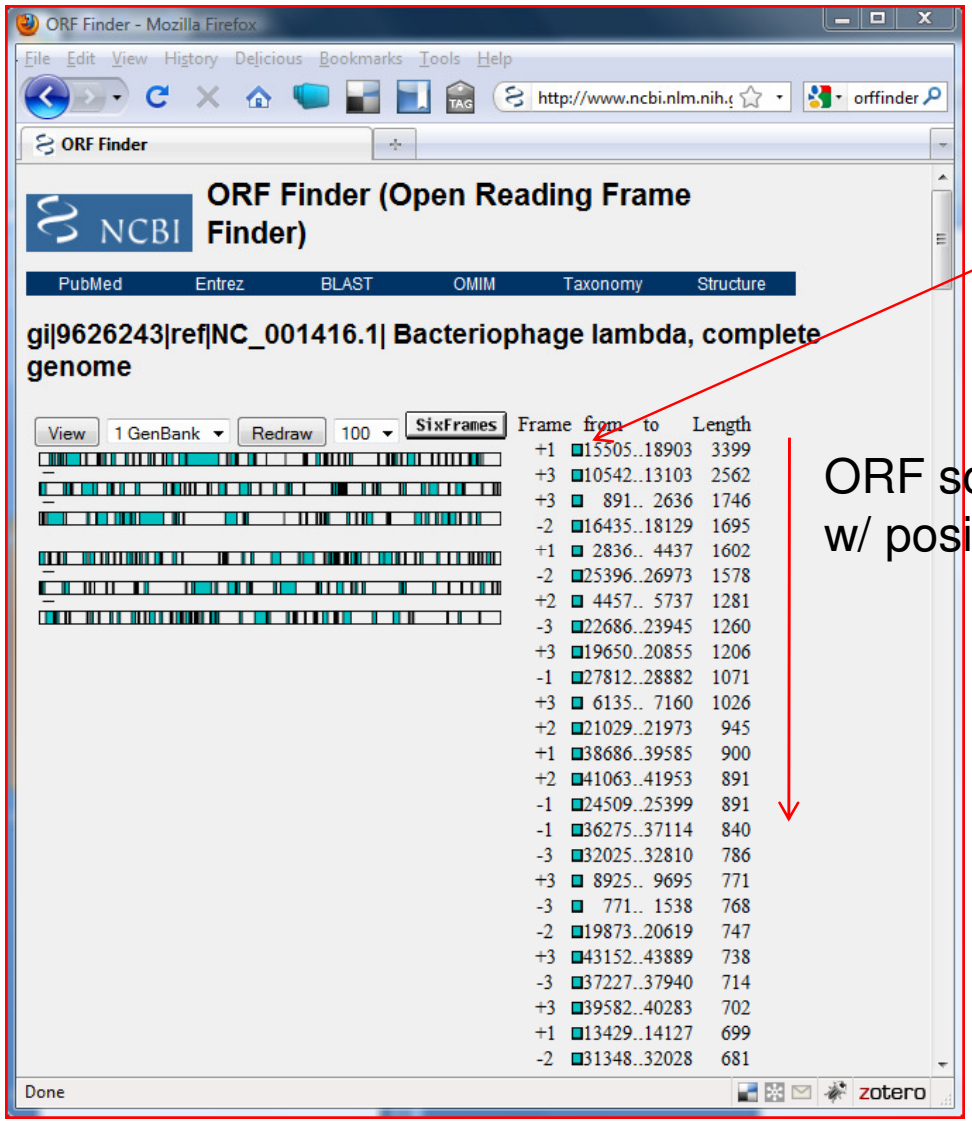
Genomic sequence in FASTA format

Genetic code



ORFFinder: Output

Graphic of ORF locations (light blue)



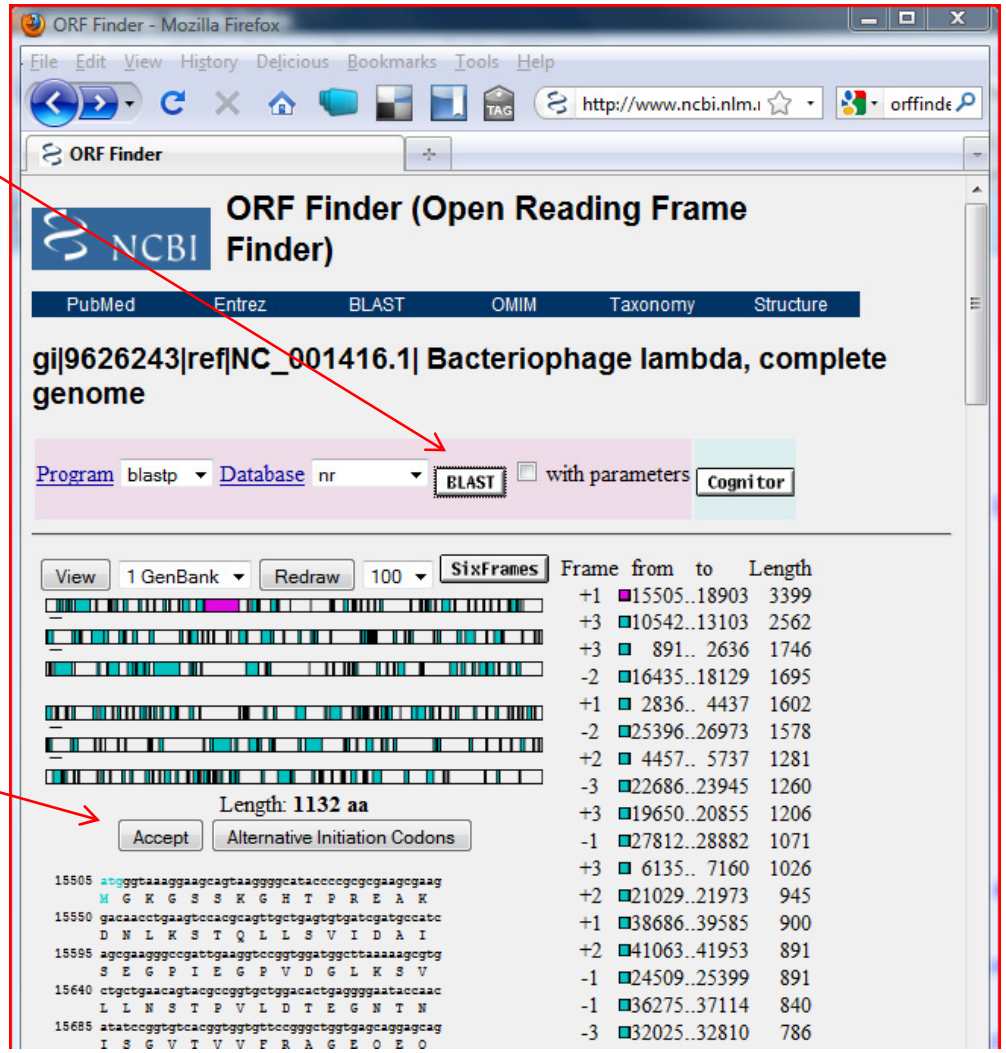
Click box to get ORF detail

ORF sorted by length w/ positions

ORFFinder: Output

BLAST protein sequence: Inspect GenBank record and report pertinent information

Access FASTA formatted sequence



Genomic and amino acid sequence of ORF