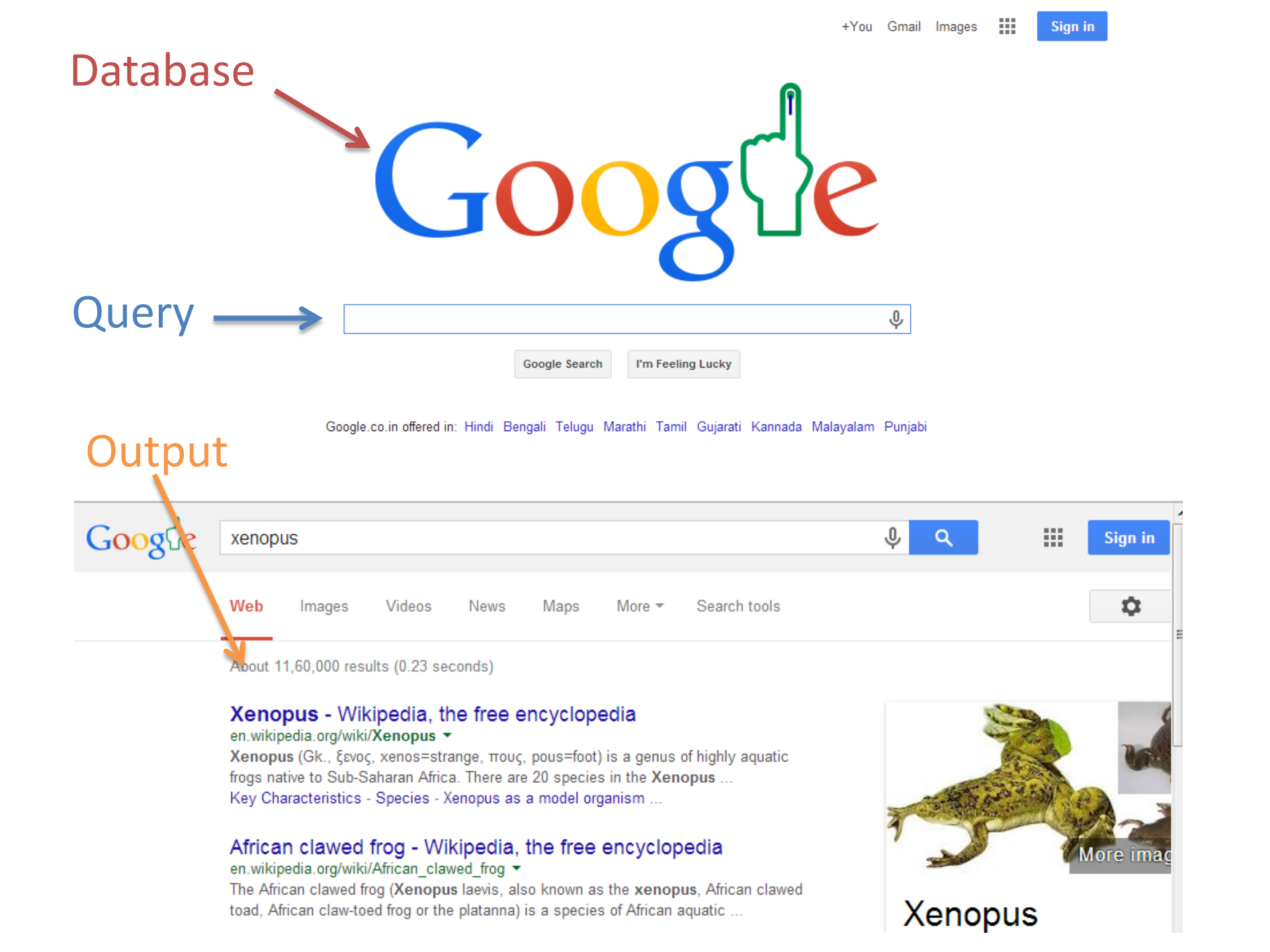
BLAST & other tools for sequence analysis

Taejoon Kwon

University of Texas at Austin

Xenopus Bioinformatics Workshop, May 2014


+You
Gmail
Images
Sign in
Database
Google
Query
Google Search
I'm Feeling Lucky
Google.co.in offered in: Hindi Bengali Telugu Marathi Tamil Gujarati Kannada Malayalam Punjabi
Output
Google
xenopus
Sign in
Web
Images
Videos
News
Maps
More
Search tools
About 11,60,000 results (0.23 seconds)
Xenopus - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Xenopus
Xenopus (Gk., ξενος, xenos=strange, πους, pous=foot) is a genus of highly aquatic frogs native to Sub-Saharan Africa. There are 20 species in the Xenopus ...
Key Characteristics - Species - Xenopus as a model organism ...
African clawed frog - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/African_clawed_frog
The African clawed frog (Xenopus laevis, also known as the xenopus, African clawed toad, African claw-toed frog or the platanna) is a species of African aquatic ...
More imag
Xenopus

NGS search - bowtie2

```
Bowtie 2 version 2.0.0-beta5 by Ben Langmead (blangmea@jhsph.edu)
Usage:
  bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r>} [-S <sam>]

  <bt2-idx>  Index filename prefix (minus trailing .X.bt2).
             NOTE: Bowtie 1 and Bowtie 2 indexes are not compatible.
  <m1>       Files with #1 mates, paired with files in <m2>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <m2>       Files with #2 mates, paired with files in <m1>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <r>        Files with unpaired reads.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <sam>      File for SAM output (default: stdout)

  <m1>, <m2>, <r> can be comma-separated lists (no whitespace) and can be
  specified many times.  E.g. '-U file1.fq,file2.fq -U file3.fq'.
```

bowtie2 [options] –x <DB file(index)> -S <output name> -1/2/U <query file>

MS/MS search - comet

```
Comet version "2013.02 rev. 0"
(c) University of Washington

Comet usage:  /work/taejoon/src.MS/comet/2013020/comet.2013020.linux.exe [options] <input_files>

Supported input formats include mzXML, mzXML, mz5 and ms2 variants (cms2, bms2, ms2)

      options:  -p          to print out a comet.params file (named comet.params.new)
                -P<params>  to specify an alternate parameters file (default comet.params)
                -N<name>    to specify an alternate output base name; valid only with one input file
                -D<dbase>   to specify a sequence database, overriding entry in parameters file
                -F<num>     to specify the first/start scan to search, overriding entry in parameters file
                -L<num>     to specify the last/end scan to search, overriding entry in parameters file
                            (-L option is required if -F option is used)
```

comet.exe –D<DB file> -N<output name> <query file>

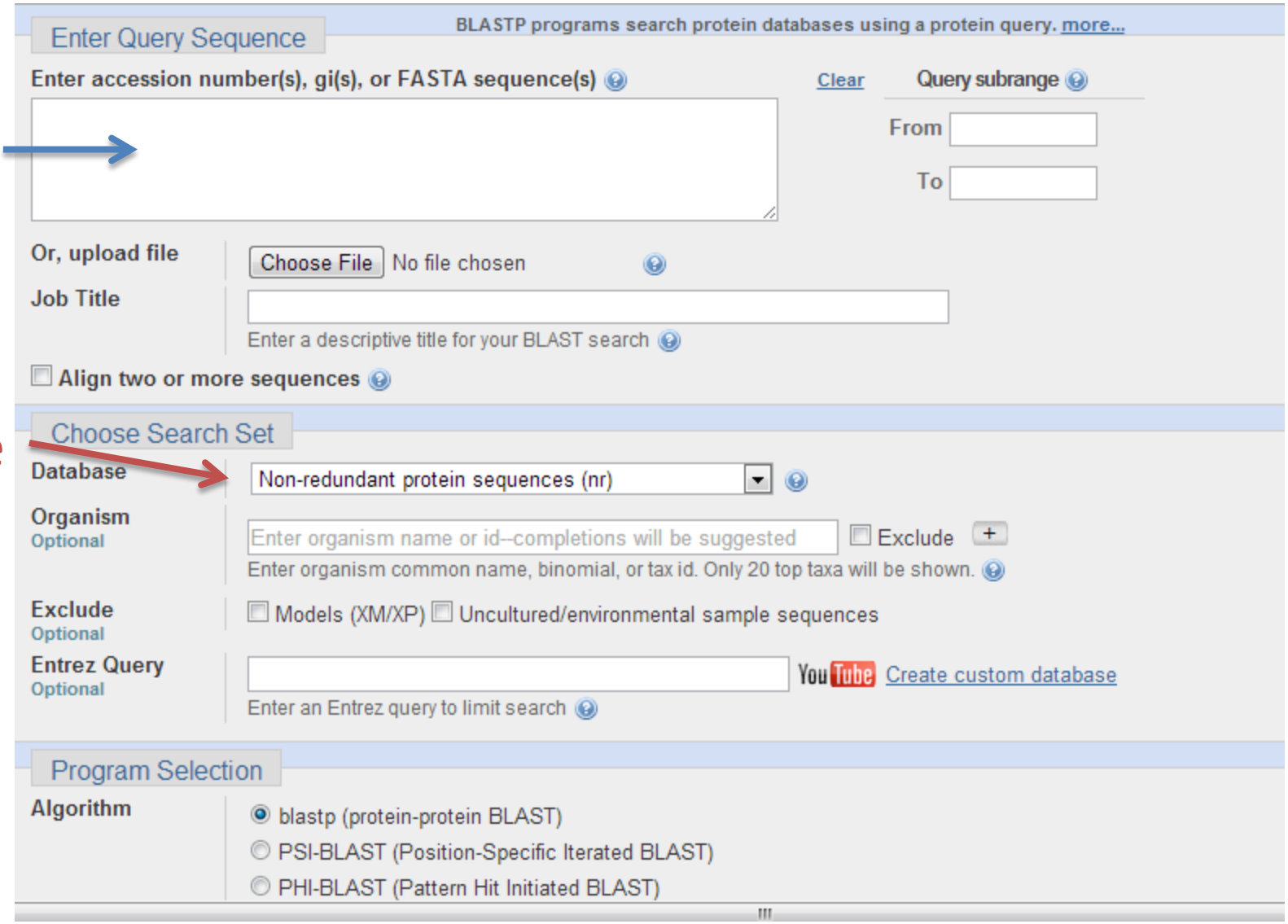
All ‘search’ program should have:

(1) Query

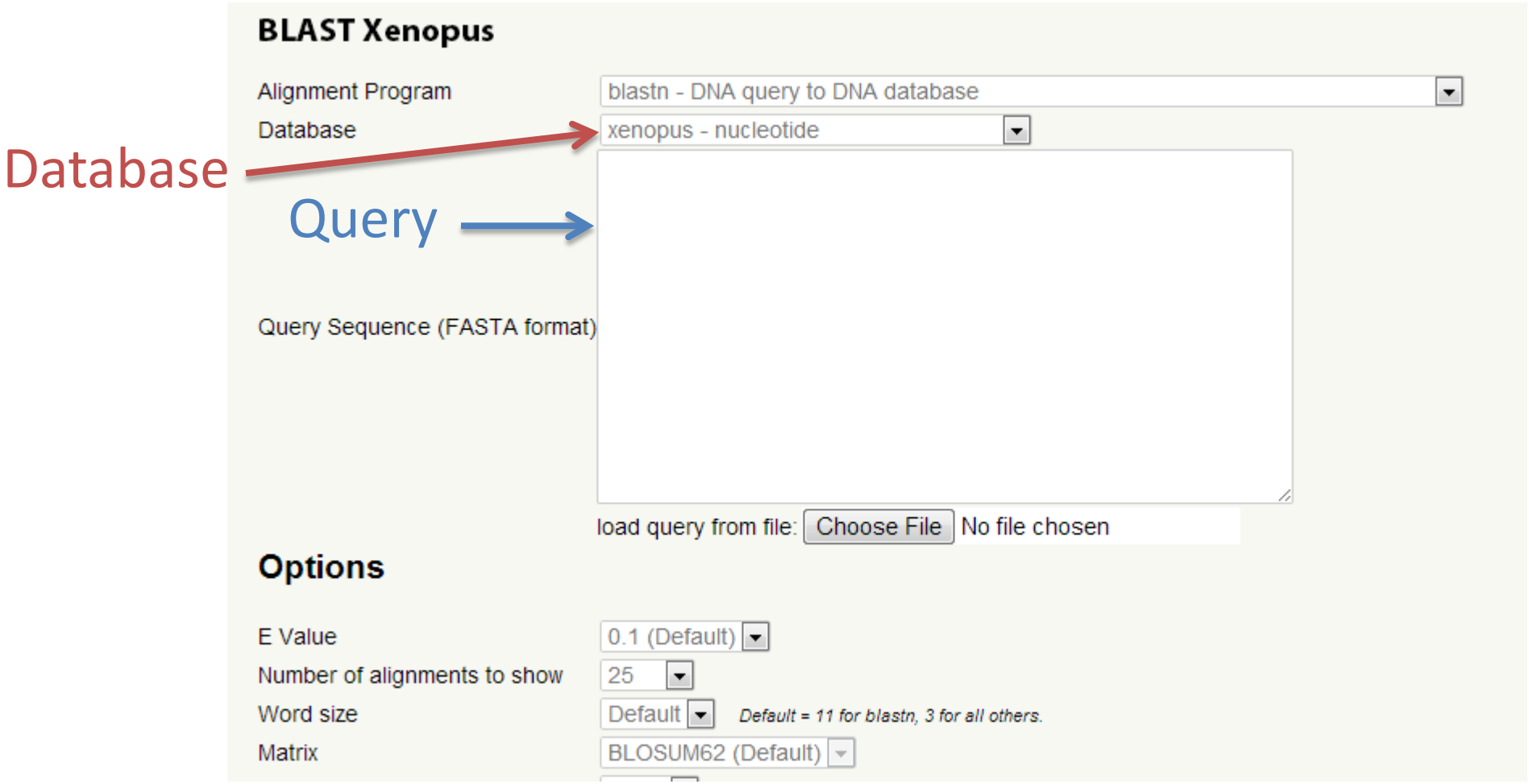
(2) Database (sometimes indexed)

(3) Output

NCBI BLAST server



Or XenBase BLAST server

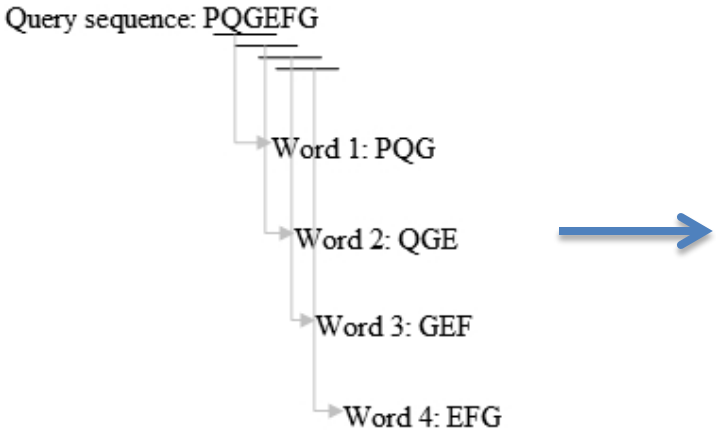


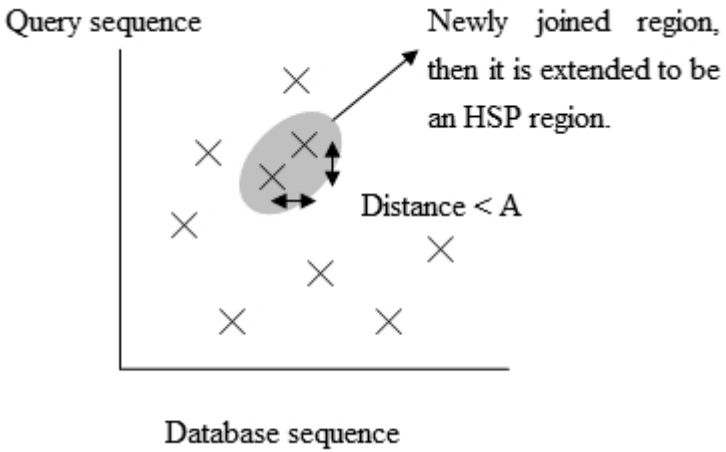
```
BLAST query/options error: Either a BLAST database or subject sequence(s) must be specified
taejoon@cygnus:/work/XenBioinfo2014/seq_align$ ~/src/blast+/current/bin/blastp -help
USAGE
  blastp [-h] [-help] [-import_search_strategy filename]
    [-export_search_strategy filename] [-task task_name] [-db database_name]
    [-dbsize num_letters] [-gilist filename] [-seqidlist filename]
    [-negative_gilist filename] [-entrez_query entrez_query]
    [-db_soft_mask filtering_algorithm] [-db_hard_mask filtering_algorithm]
    [-subject subject_input_file] [-subject_loc range] [-query input_file]
    [-out output_file] [-evalue evalue] [-word_size int_value]
    [-gapopen open_penalty] [-gapextend extend_penalty]
    [-xdrop_ungap float_value] [-xdrop_gap float_value]
    [-xdrop_gap_final float_value] [-searchsp int_value]
    [-max_hsps_per_subject int_value] [-seg SEG_options]
    [-soft_masking soft_masking] [-matrix matrix_name]
    [-threshold float_value] [-culling_limit int_value]
    [-best_hit_overhang float_value] [-best_hit_score_edge float_value]
    [-window_size int_value] [-lcase_masking] [-query_loc range]
    [-parse_deflines] [-outfmt format] [-show_gis]
    [-num_descriptions int_value] [-num_alignments int_value] [-html]
    [-max_target_seqs num_sequences] [-num_threads int_value] [-ungapped]
    [-remote] [-comp_based_stats compo] [-use_sw_tback] [-version]
```

blastp [options] –db <DB file(index)> -out <output name> -in <query file>

blastn [options] –db <DB file(index)> -out <output name> -in <query file>

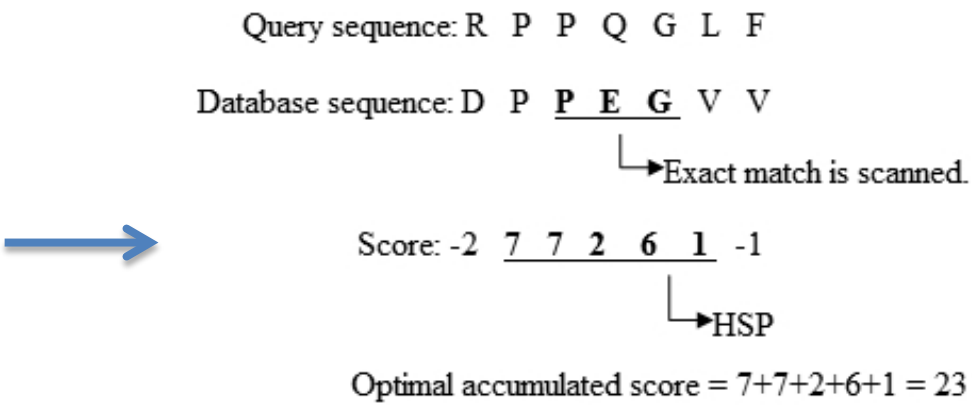
How BLAST works





1. Make k-mer list of “query”

2. Search HSP (High-scoring Segment Pair)



3. Extend exact matches

BLAST is LOCAL alignment

```
Global  FTFTALILLAVAV
        F--TAL-LLA-AV

Local   FTFTALILL-AVAV
        --FTAL-LLAAV--
```

Download & Install

ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/

Index of /blast/executables/blast+/LATEST/

Name	Size	Date Modified
[parent directory]		
ChangeLog	85 B	1/6/14 3:35:00 PM
ncbi-blast-2.2.29+-1.i686.rpm	174 MB	1/6/14 3:36:00 PM
ncbi-blast-2.2.29+-1.src.rpm	11.3 MB	1/6/14 3:36:00 PM
ncbi-blast-2.2.29+-1.x86_64.rpm	151 MB	1/6/14 3:36:00 PM
ncbi-blast-2.2.29+-ia32-linux.tar.gz	174 MB	1/6/14 3:37:00 PM
ncbi-blast-2.2.29+-ia32-win32.tar.gz	56.5 MB	1/6/14 3:37:00 PM
ncbi-blast-2.2.29+-src.tar.gz	14.9 MB	1/6/14 3:37:00 PM
ncbi-blast-2.2.29+-src.zip	17.9 MB	1/6/14 3:37:00 PM
ncbi-blast-2.2.29+-universal-macosx.tar.gz	254 MB	1/6/14 3:38:00 PM
ncbi-blast-2.2.29+-win32.exe	56.6 MB	1/6/14 3:38:00 PM
ncbi-blast-2.2.29+-win64.exe	64.7 MB	1/6/14 3:38:00 PM
ncbi-blast-2.2.29+-x64-linux.tar.gz	151 MB	1/6/14 3:39:00 PM
ncbi-blast-2.2.29+-x64-win64.tar.gz	64.5 MB	1/6/14 3:39:00 PM
ncbi-blast-2.2.29+.dmg	255 MB	1/6/14 3:39:00 PM

makeblastdb

```
~/src/blast+/current/bin/makeblastdb
-dbtype prot
-in Wu2013_AMBME.prot_annot.fa
-out Wu2013_AMBME.prot_annot.fa


Building a new DB, current time: 05/16/2014 06:43:12
New DB name:   Wu2013_AMBME.prot_annot.fa
New DB title:  Wu2013_AMBME.prot_annot.fa
Sequence type: Protein
Keep Linkouts: T
Keep MBits: T
Maximum file size: 1073741824B
Adding sequences from FASTA; added 43726 sequences in 1.45152 seconds.
```

makeblastdb

```
$ ~/src/blast+/current/bin/makeblastdb
-dbtype nucl
-in Wu2013_AMBME.cdna_annot.fa
-out Wu2013_AMBME.cdna_annot.fa

Building a new DB, current time: 05/16/2014 06:36:39
New DB name:   Wu2013_AMBME.cdna_annot.fa
New DB title:  Wu2013_AMBME.cdna_annot.fa
Sequence type: Nucleotide
Keep Linkouts: T
Keep MBits: T
Maximum file size: 1073741824B
Adding sequences from FASTA; added 43726 sequences in 2.37648 seconds.
```

blastp

```
USAGE
  blastp [-h] [-help] [-import_search_strategy filename]
    [-export_search_strategy filename] [-task task_name] [-db database_name]
    [-dbsize num_letters] [-gilist filename] [-seqidlist filename]
    [-negative_gilist filename] [-entrez_query entrez_query]
    [-db_soft_mask filtering_algorithm] [-db_hard_mask filtering_algorithm]
    [-subject subject_input_file] [-subject_loc range] [-query input_file]
    [-out output_file] [-evalue evalue] [-word_size int_value]
    [-gapopen open_penalty] [-gapextend extend_penalty]
    [-xdrop_ungap float_value] [-xdrop_gap float_value]
    [-xdrop_gap_final float_value] [-searchsp int_value]
    [-max_hsps_per_subject int_value] [-seg SEG_options]
    [-soft_masking soft_masking] [-matrix matrix_name]
    [-threshold float_value] [-culling_limit int_value]
    [-best_hit_overhang float_value] [-best_hit_score_edge float_value]
    [-window_size int_value] [-lcase_masking] [-query_loc range]
    [-parse_deflines] [-outfmt format] [-show_gis]
    [-num_descriptions int_value] [-num_alignments int_value] [-html]
    [-max_target_seqs num_sequences] [-num_threads int_value] [-ungapped]
    [-remote] [-comp_based_stats compo] [-use_sw_tback] [-version]
```

blastp [options] –db <DB file(index)> -out <output name> -query <query file>

Basic options

```
-task <String, Permissible values: 'blastp' 'blastp-short' 'deltablast' >
  Task to execute
  Default = `blastp'
-db <String>
  BLAST database name
   * Incompatible with:  subject, subject_loc
-out <File_Out>
  Output file name
  Default = `-'
-evalue <Real>
  Expectation value (E) threshold for saving hits
  Default = `10'
-word_size <Integer, >=2>
  Word size for wordfinder algorithm
```

```
-html
  Produce HTML output?

*** Query filtering options
-seg <String>
  Filter query sequence with SEG (Format: 'yes', 'window locut hicut', or
  'no' to disable)
  Default = `no'
```

-outfmt

```
*** Formatting options
-outfmt <String>
  alignment view options:
    0 = pairwise,
    1 = query-anchored showing identities,
    2 = query-anchored no identities,
    3 = flat query-anchored, show identities,
    4 = flat query-anchored, no identities,
    5 = XML Blast output,
    6 = tabular,
    7 = tabular with comment lines,
    8 = Text ASN.1,
    9 = Binary ASN.1,
   10 = Comma-separated values,
   11 = BLAST archive format (ASN.1)
```

-outfmt

```
Options 6, 7, and 10 can be additionally configured to produce
a custom format specified by space delimited format specifiers.
The supported format specifiers are:
         qseqid means Query Seq-id
            qgi means Query GI
           qacc means Query accesion
        qaccver means Query accesion.version
           qlen means Query sequence length
         sseqid means Subject Seq-id
      sallseqid means All subject Seq-id(s), separated by a ';'
        gapopen means Number of gap openings
           gaps means Total number of gaps
           ppos means Percentage of positive-scoring matches
         frames means Query and subject frames separated by a '/'
         qframe means Query frame
         sframe means Subject frame
           btop means Blast traceback operations (BTOP)
  When not provided, the default value is:
  'qseqid sseqid pident length mismatch gapopen qstart qend sstart send
  evalue bitscore', which is equivalent to the keyword 'std'
  Default = `0'
```

blastn

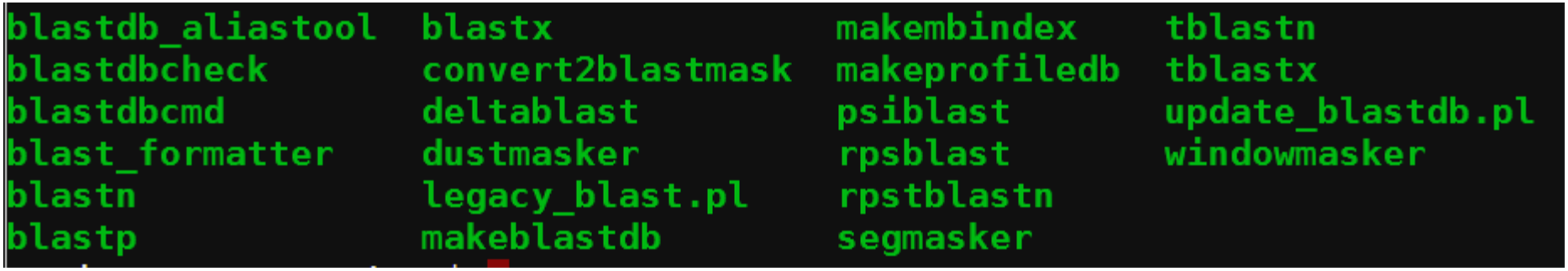
```
blastn [-h] [-help] [-import_search_strategy filename]
  [-export_search_strategy filename] [-task task_name] [-db database_name]
  [-dbsize num_letters] [-gilist filename] [-seqidlist filename]
  [-negative_gilist filename] [-entrez_query entrez_query]
  [-db_soft_mask filtering_algorithm] [-db_hard_mask filtering_algorithm]
  [-subject subject_input_file] [-subject_loc range] [-query input_file]
  [-out output_file] [-evalue evalue] [-word_size int_value]
  [-gapopen open_penalty] [-gapextend extend_penalty]
  [-perc_identity float_value] [-xdrop_ungap float_value]
  [-xdrop_gap float_value] [-xdrop_gap_final float_value]
  [-searchsp int_value] [-max_hsps_per_subject int_value] [-penalty penalty]
  [-reward reward] [-no_greedy] [-min_raw_gapped_score int_value]
  [-template_type type] [-template_length int_value] [-dust DUST_options]
  [-filtering_db filtering_database]
  [-window_masker_taxid window_masker_taxid]
  [-window_masker_db window_masker_db] [-soft_masking soft_masking]
  [-ungapped] [-culling_limit int_value] [-best_hit_overhang float_value]
  [-best_hit_score_edge float_value] [-window_size int_value]
  [-off_diagonal_range int_value] [-use_index boolean] [-index_name string]
  [-lcase_masking] [-query_loc range] [-strand strand] [-parse_deflines]
  [-outfmt format] [-show_gis] [-num_descriptions int_value]
  [-num_alignments int_value] [-html] [-max_target_seqs num_sequences]
  [-num_threads int_value] [-remote] [-version]
```

blastn [options] –db <DB file(index)> -out <output name> -query <query file>

Basic options

```
-task <String, Permissible values: 'blastn' 'blastn-short' 'dc-megablast'
                'megablast' 'rmblastn' >
  Task to execute
  Default = `megablast'
-db <String>
  BLAST database name
    * Incompatible with:  subject, subject_loc
-out <File_Out>
  Output file name
  Default = `-'
-evalue <Real>
  Expectation value (E) threshold for saving hits
  Default = `10'
-word_size <Integer, >=4>
  Word size for wordfinder algorithm (length of best perfect match)
```

What about all others?



- blastp: Query=protein → DB=protein
- blastn: Query=dna → DB=dna
- blastx : Query=dna + translation → DB=protein
- tblastn: Query=protein → DB=dna + translation
- tblastx: Query=dna + translation → DB=dna + translation
- psiblast: Position-Specific Initiated BLAST

BLAT

```
blat - Standalone BLAT v. 34 fast sequence search command line tool
usage:
   blat database query [-ooc=11.ooc] output.psl
where:
   database and query are each either a .fa , .nib or .2bit file,
   or a list these files one file name per line.
   -ooc=11.ooc tells the program to load over-occurring 11-mers from
               and external file.  This will increase the speed
               by a factor of 40 in many cases, but is not required
   output.psl is where to put the output.
   Subranges of nib and .2bit files may specified using the syntax:
      /path/file.nib:seqid:start-end
   or
      /path/file.2bit:seqid:start-end
   or
      /path/file.nib:start-end
```

blat <DB file> <query file> <output name>

-type

```
-t=type     Database type.  Type is one of:
              dna - DNA sequence
              prot - protein sequence
              dnax - DNA sequence translated in six frames to protein
            The default is dna
-q=type     Query type.  Type is one of:
              dna - DNA sequence
              rna - RNA sequence
              prot - protein sequence
              dnax - DNA sequence translated in six frames to protein
              rnax - DNA sequence translated in three frames to protein
            The default is dna
-prot       Synonymous with -t=prot -q=prot
-ooc=N.ooc  Use overused tile file N.ooc.  N should correspond to
            the tileSize
-tileSize=N sets the size of match that triggers an alignment.
            Usually between 8 and 12
            Default is 11 for DNA and 5 for protein.
-stepSize=N spacing between tiles. Default is tileSize.
-oneOff=N   If set to 1 this allows one mismatch in tile and still
            triggers an alignments.  Default is 0.
```

-out

```
-out=type   Controls output file format.  Type is one of:
                psl - Default.  Tab separated format, no sequence
                pslx - Tab separated format with sequence
                axt - blastz-associated axt format
                maf - multiz-associated maf format
                sim4 - similar to sim4 format
                wublast - similar to wublast format
                blast - similar to NCBI blast format
                blast8- NCBI blast tabular format
                blast9 - NCBI blast tabular format with comments
-fine       For high quality mRNAs look harder for small initial and
            terminal exons.  Not recommended for ESTs
-maxIntron=N  Sets maximum intron size. Default is 750000
-extendThroughN - Allows extension of alignment through large blocks of N's
```

PSL

psLayout version 3

match	mis-match	rep. match	N's	Q gap count	Q gap bases	T gap count	T gap bases	strand	Q name	Q size	Q start	Q end	T name	T size	T start	T end	block count	blockSizes	qStarts	tStarts
563	2	0	0	0	0	0	0	+	xenTro3_ct_ncRNAsall_7135_TCONS_00000075	572	7	572	XENTR_JGI7a.scaffold_4	140444050	10363498	10363498	6	103635551	1	565,	7,	103634986,
997	1	0	0	0	0	0	0	+	xenTro3_ct_ncRNAsall_7135_TCONS_00000755	998	0	998	XENTR_JGI7a.scaffold_4	140444050	10103073	10103073	2	101031730	1	998,	0,	101030732,
30	0	0	0	0	0	0	0	+	xenTro3_ct_ncRNAsall_7135_TCONS_00000755	998	14	44	XENTR_JGI7a.scaffold_3	117219152	11634050	11634050	4	116340534	1	30,	14,	116340504,
41	2	0	0	0	0	0	0	-	xenTro3_ct_ncRNAsall_7135_TCONS_00000755	998	1	44	XENTR_JGI7a.scaffold_8	108038044	10836483	10836483	2	10836526	1	43,	954,	10836483,
197	10	0	0	2	12	2	2	-	xenTro3_ct_ncRNAsall_7135_TCONS_00000755	998	734	953	XENTR_JGI7a.scaffold_4	140444050	10102235	10102235	7	101022566	4	58,69,9,71,	45,104,173,193,	101022357,101022415,101022485,101022495,
42	0	0	0	1	7	1	11	-	xenTro3_ct_ncRNAsall_7135_TCONS_00000755	998	17	66	XENTR_JGI7a.scaffold_3	117219152	88136809	88136809	2	88136862	2	12,30,	932,951,	88136809,88136832,
40	2	0	0	1	2	0	0	-	xenTro3_ct_ncRNAsall_7135_TCONS_00000755	998	3	47	XENTR_JGI7a.scaffold_25	948372	726341	726383	2	23,19,	951,976,	726341,726364,
38	1	0	0	0	0	0	0	-	xenTro3_ct_ncRNAsall_7135_TCONS_00000755	998	5	44	XENTR_JGI7a.scaffold_2	150913983	89650904	89650904	2	89650943	1	39,	954,	89650904,
41	3	0	0	0	0	0	0	-	xenTro3_ct_ncRNAsall_7135_TCONS_00000755	998	0	44	XENTR_JGI7a.scaffold_1	215906545	12067967	12067967	9	120679723	1	44,	954,	120679679,
37	3	0	0	0	0	0	0	-	xenTro3_ct_ncRNAsall_7135_TCONS_00000755	998	4	44	XENTR_JGI7a.scaffold_1	215906545	20703034	20703034	9	207030389	1	40,	954,	207030349,
77	2	0	0	0	0	0	0	+	xenTro3_ct_ncRNAsall_7135_TCONS_00000756	1074	995	1074	XENTR_JGI7a.scaffold_99	313365	153439	153518	1	79,	995,	153439,
76	3	0	0	0	0	0	0	+	xenTro3_ct_ncRNAsall_7135_TCONS_00000756	1074	995	1074	XENTR_JGI7a.scaffold_98	308099	252867	252946	1	79,	995,	252867,
73	6	0	0	0	0	0	0	+	xenTro3_ct_ncRNAsall_7135_TCONS_00000756	1074	995	1074	XENTR_JGI7a.scaffold_98	308099	194847	194926	1	79,	995,	194847,
76	2	0	0	1	1	0	0	+	xenTro3_ct_ncRNAsall_7135_TCONS_00000756	1074	995	1074	XENTR_JGI7a.scaffold_94	329321	273812	273890	2	42,36,	995,1038,	273812,273854,
77	2	0	0	0	0	0	0	+	xenTro3_ct_ncRNAsall_7135_TCONS_00000756	1074	995	1074	XENTR_JGI7a.scaffold_93	317421	266269	266348	1	79,	995,	266269,
91	1	0	0	1	57	1	285	+	xenTro3_ct_ncRNAsall_7135_TCONS_00000756	1074	925	1074	XENTR_JGI7a.scaffold_9	100021500	6677333	6677710	2	13,79,	925,995,	6677333,6677631,
78	1	0	0	0	0	0	0	+	xenTro3_ct_ncRNAsall_7135_TCONS_00000756	1074	995	1074	XENTR_JGI7a.scaffold_9	100021500	24044258	24044337	1	79,	995,	24044258,
78	1	0	0	0	0	0	0	+	xenTro3_ct_ncRNAsall_7135_TCONS_00000756	1074	995	1074	XENTR_JGI7a.scaffold_9	100021500	26092421	26092500	1	79,	995,	26092421,

GMAP (& gsnap)

```
gmap_build: Builds a gmap database for a genome to be used by GMAP or GSNAP.
Part of GMAP package, version 2013-07-20.

A simplified alternative to using the program gmap_setup, which creates a Makefile.

Usage: gmap_build [options...] -d <genomename> <fasta_files>

Options:
    -D, --dir=STRING        Destination directory for installation (defaults to gmapdb directory
 specified at configure time)
    -d, --db=STRING         Genome name
    -T STRING               Temporary build directory
```

```
Usage: gmap [OPTIONS...] <FASTA files...>, or
       cat <FASTA files...> | gmap [OPTIONS...]

Input options (must include -d or -g)
  -D, --dir=directory           Genome directory
  -d, --db=STRING               Genome database.  If argument is '?' (with
                                  the quotes), this command lists available databases.

  -k, --kmer=INT                kmer size to use in genome database (allowed values: 16 or less
).
                                  If not specified, the program will find the highest available
                                  kmer size in the genome database
```

gmap_build [options] **–d <DB name> -D <DB directory> <DB file>**

gmap [options] **–d <DB name> -D <DB directory> -f <output name> <query file>**

MUSCLE: multiple sequence alignment with high accuracy and high throughput

Robert C. Edgar*

195 Roque Moraes Drive, Mill Valley, CA 94941, USA

An unemployed gentleman scholar

Posted on May 4, 2010 | 7 Comments

A comment to a previous post asks me for some personal information: *"I've noticed that you never list a university, firm, or non-profit affiliation on your papers or website. Would you mind writing a post about how you got to be where you are, who supports your work, the reactions of reviewers to papers from outside the university/well-known industrial research lab circle, and the like? I for one would be terribly inte*

So to answer the question: after selling my business I had some financial independence, and I've supported myself and my research from my savings for the past decade. You can think of me as unemployed, independent and/or a gentleman scholar of modest means (like, say, my fellow-countryman Charles Darwin). It's hard to know how people have perceived my unconventional status. Everyone feels misunderstood and dismissed by peers sometimes (reviewers, editors, conference organizers...), so it's impossible to say whether I would have been better accepted if I had a conventional affiliation. MUSCLE has been helpful because many people have heard of it (almost 3,000 citations so far per Google Scholar). That gave me some street cred early on, and surely helped open other doors.

Download & Install

http://www.drive5.com/muscle/downloads.htm

MUSCLE home

Documentation

Support

Source code
v3.8.31

Operating system	Processor	Bits	Executable file
Linux	Intel i86	32	muscle3.8.31_i86linux32.tar.gz
Linux	Intel i86	64	muscle3.8.31_i86linux64.tar.gz
Mac OSX	Intel i86	32	muscle3.8.31_i86darwin32.tar.gz
Mac OSX	Intel i86	64	muscle3.8.31_i86darwin64.tar.gz
Mac	PPC	64	muscle3.8.31_macppc.tar.gz
Windows	Intel i86	32	muscle3.8.31_i86win32.exe
Windows/Cygwin	Intel i86	32	muscle3.8.31_i86cygwin32.exe

Files are in tarball format. Use tar -zxvf *filename* to extract.

See also

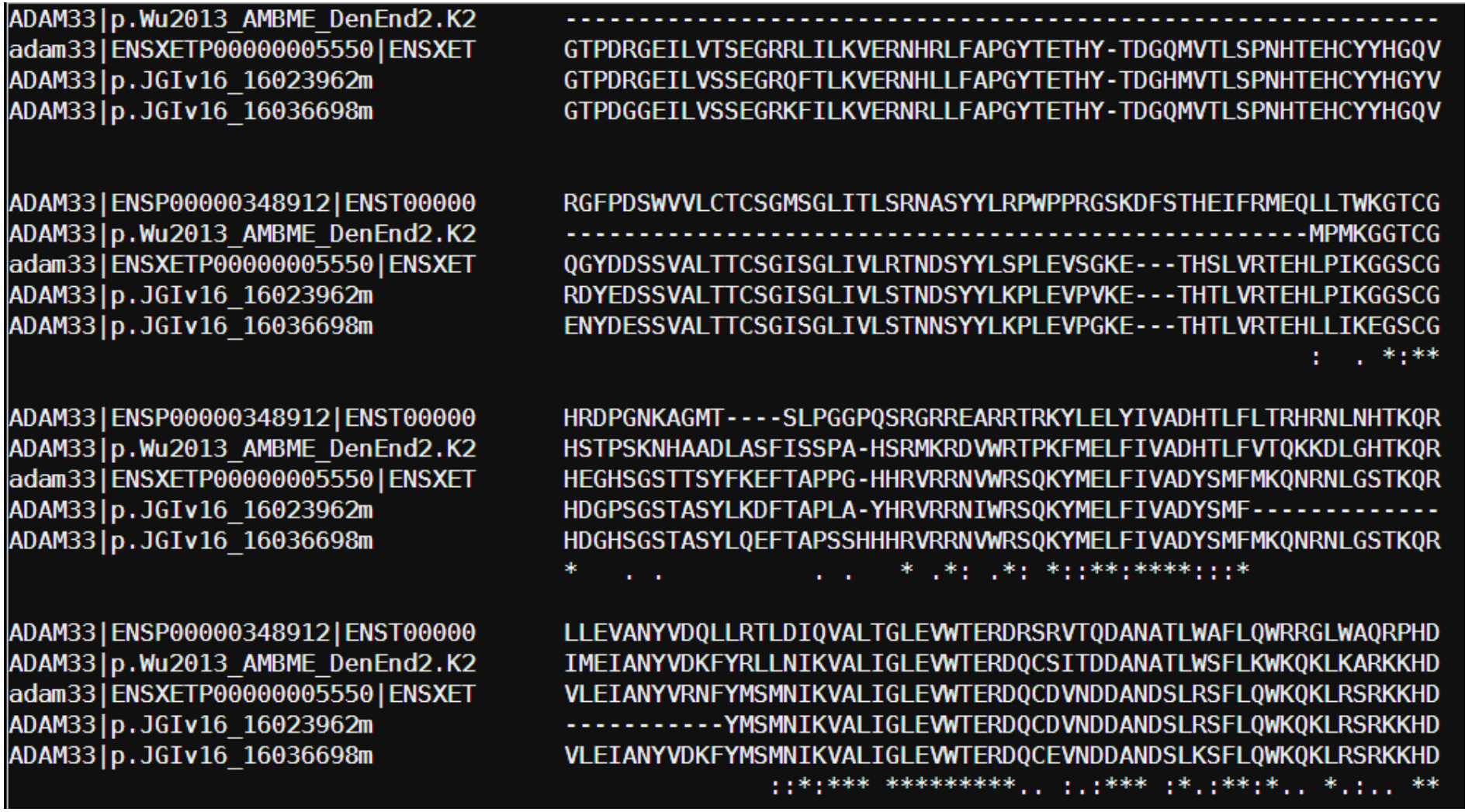
Older versions

muscle3.8.425_binaries.tar.gz (with bug fix for long sequences).

muscle3.8.425_src.tar.gz (Source for v3.8.425).

```
Basic usage

    muscle -in <inputfile> -out <outputfile>

Common options (for a complete list please see the User Guide):

    -in <inputfile>    Input file in FASTA format (default stdin)
    -out <outputfile>  Output alignment in FASTA format (default stdout)
    -diags             Find diagonals (faster for similar sequences)
    -maxiters <n>      Maximum number of iterations (integer, default 16)
    -maxhours <h>      Maximum time to iterate in hours (default no limit)
    -html              Write output in HTML format (default FASTA)
    -msf               Write output in GCG MSF format (default FASTA)
    -clw               Write output in CLUSTALW format (default FASTA)
    -clwstrict         As -clw, with 'CLUSTAL W (1.81)' header
    -log[a] <logfile>  Log to file (append if -loga, overwrite if -log)
    -quiet             Do not write progress messages to stderr
    -version           Display version information and exit
```

```
taejoon@cygnus:/work/XenBioinfo2014/seq_align$ ~/src/muscle/muscle3.8.31_i86linux64 -in adam33_prot.fa -out
 adam33_prot.clw -clw

MUSCLE v3.8.31 by Robert C. Edgar

http://www.drive5.com/muscle
This software is donated to the public domain.
Please cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.

adam33_prot 5 seqs, max length 914, avg  length 851
00:00:00    10 MB(-1%)  Iter   1  100.00%  K-mer dist pass 1
00:00:00    10 MB(-1%)  Iter   1  100.00%  K-mer dist pass 2
00:00:01    16 MB(-2%)  Iter   1  100.00%  Align node
00:00:01    16 MB(-2%)  Iter   1  100.00%  Root alignment
00:00:01    16 MB(-2%)  Iter   2  100.00%  Refine tree
00:00:01    16 MB(-2%)  Iter   2  100.00%  Root alignment
00:00:01    16 MB(-2%)  Iter   2  100.00%  Root alignment
00:00:01    16 MB(-2%)  Iter   3  100.00%  Refine biparts
00:00:01    16 MB(-2%)  Iter   4  100.00%  Refine biparts
00:00:01    16 MB(-2%)  Iter   5  100.00%  Refine biparts
00:00:01    16 MB(-2%)  Iter   6  100.00%  Refine biparts
00:00:01    16 MB(-2%)  Iter   7  100.00%  Refine biparts
00:00:01    16 MB(-2%)  Iter   8  100.00%  Refine biparts
00:00:01    16 MB(-2%)  Iter   9  100.00%  Refine biparts
00:00:01    16 MB(-2%)  Iter  10  100.00%  Refine biparts
```


ADAM33|p.Wu2013_AMBME_DenEnd2.K2 ----------
adam33|ENSXETP00000005550|ENSXET GTPDRGEILVTSEGRRLILKVERNHRLFAPGYTETHY-TDGQMVTLSPNHTEHCYYHGQV
ADAM33|p.JGIv16_16023962m GTPDRGEILVSSEGRQFTLKVERNHLLFAPGYTETHY-TDGHMVTLSPNHTEHCYYHGYV
ADAM33|p.JGIv16_16036698m GTPDGGEILVSSEGRKFILKVERNRLLFAPGYTETHY-TDGQMVTLSPNHTEHCYYHGQV
ADAM33|ENSP00000348912|ENST00000 RGFPDSWVVLCTCSGMSGLITLSRNASYYLRPWPPRGSKDFSTHEIFRMEQLLTWKGTCG
ADAM33|p.Wu2013_AMBME_DenEnd2.K2 ----------MPMKGGTCG
adam33|ENSXETP00000005550|ENSXET QGYDDSSVALTTCSGISGLIVLRTNDSYYLSPLEVSGKE---THSLVRTEHLPIKGGSCG
ADAM33|p.JGIv16_16023962m RDYEDSSVALTTCSGISGLIVLSTNDSYYLKPLEVPVKE---THTLVRTEHLPIKGGSCG
ADAM33|p.JGIv16_16036698m ENYDESSVALTTCSGISGLIVLSTNNSYYLKPLEVPGKE---THTLVRTEHLLIKEGSCG
ADAM33|ENSP00000348912|ENST00000 HRDPGNKAGMT----SLPGGPQSRGRREARRTRKYLELYIVADHTLFLTRHRNLNHTKQR
ADAM33|p.Wu2013_AMBME_DenEnd2.K2 HSTPSKNHAADLASFISSPA-HSRMKRDVWRTPKFMELFIVADHTLFVTQKKDLGHTKQR
adam33|ENSXETP00000005550|ENSXET HEGHSGSTTSYFKEFTAPPG-HHRVRRNVWRSQKYMELFIVADYSMFMKQNRNLGSTKQR
ADAM33|p.JGIv16_16023962m HDGPSGSTASYLKDFTAPLA-YHRVRRNIWRSQKYMELFIVADYSMF----------
ADAM33|p.JGIv16_16036698m HDGHSGSTASYLQEFTAPSSHHHRVRRNVWRSQKYMELFIVADYSMFMKQNRNLGSTKQR
ADAM33|ENSP00000348912|ENST00000 LLEVANYVDQLLRTLDIQVALTGLEVWTERDRSRVTQDANATLWAFLQWRRGLWAQRPHD
ADAM33|p.Wu2013_AMBME_DenEnd2.K2 IMEIANYVDKFYRLLNIKVALIGLEVWTERDQCSITDDANATLWSFLKWKQKLKARKKHD
adam33|ENSXETP00000005550|ENSXET VLEIANYVRNFYMSMNIKVALIGLEVWTERDQCDVNDDANDSLRSFLQWKQKLRSRKKHD
ADAM33|p.JGIv16_16023962m ----------YMSMNIKVALIGLEVWTERDQCDVNDDANDSLRSFLQWKQKLRSRKKHD
ADAM33|p.JGIv16_16036698m VLEIANYVDKFYMSMNIKVALIGLEVWTERDQCEVNDDANDSLKSFLQWKQKLRSRKKHD

EXONERATE

a generic tool for sequence alignment

```
Guy St.C. Slater. guy@ebi.ac.uk. 2000-2008.


Examples of use:

1. Ungapped alignment of any DNA or protein sequences:
    exonerate queries.fa targets.fa
2. Gapped alignment of Mouse proteins to Fugu proteins:
    exonerate --model affine:local mouse.fa fugu.fa
3. Find top 10 matches of each EST to a genome:
    exonerate --model est2genome --bestn 10 est.fa genome.fa
4. Find proteins with at least a 50% match to a genome:
    exonerate --model protein2genome --percent 50 p.fa g.fa
5. Perform a full Smith-Waterman-Gotoh alignment:
    exonerate --model affine:local --exhaustive yes a.fa b.fa
6. Many more combinations are possible.  To find out more:
    exonerate --help
    man exonerate
```

Summary – What should I use?

- Protein query vs Protein database
  - “BLASTP” or “BLAT –prot”
- cDNA query against cDNA database
  - BLASTN
  - BLAT or GMAP
- cDNA/ncDNA query against genome
  - BLAT or GMAP
  - Exonerate
  - Or BLASTN maybe…
- Protein query against genome
  - Exonerate (very slow!)
- Multiple sequence alignment
  - Muscle