BLAST

Getting the most from your cycles.

Tom Madden
NCBI/NLM/NIH
madden@ncbi.nlm.nih.gov

The BLAST algorithm

What is BLAST?

- **B**asic **L**ocal **A**lignment **S**earch **T**ool
- Calculates similarity for biological sequences.
- Produces local alignments: only a portion of each sequence must be aligned.
- Uses statistical theory to determine if a match might have occurred by chance.

The BLAST family of programs allows all combinations of DNA or protein query sequences with searches against DNA or protein databases:

Protein-protein (blastp): compares an amino acid sequence against a protein sequence database.

Nucl.-nucl (blastn): compares a nucleotide query sequence against a nucleotide sequence database (in general optimized for speed, not sensitivity).

Translated nucl.-protein (blastx): compares the six-frame conceptual translation products of a nucleotide query against a protein sequence database.

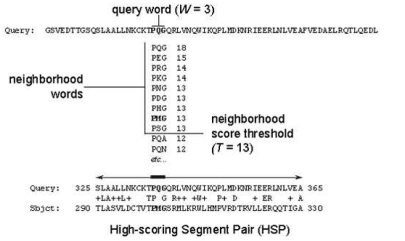
Protein-translated nucl (tblastn): compares a protein query sequence against a sequence database dynamically translated in all six reading frames (useful for searching proteins against EST's).

Translated nucl-translated nucl. (tblastx): compares the six frame translation of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

BLAST is a heuristic.

- A lookup table is made of all the "words" (short subsequences) in the query sequence. In many types of searches "neighboring" words are included.
- The database is scanned for matching words ("hot spots").
- Gapped and un-gapped extensions are initiated from these matches.

The BLAST Search Algorithm

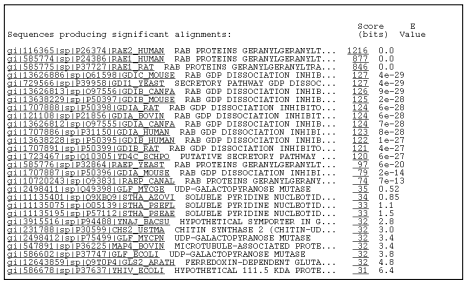


BLAST OUTPUT

There are many different BLAST output formats.

- Pair-wise report
- Query-anchored report
- Hit-table
- Tax BLAST
- Abstract Syntax Notation 1
- XML

One-line descriptions



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

gi|116365|sp|P26374|RAE2_HUMAN  RAB PROTEINS GERANYLGERANYLT...  1216  0.0
gi|1585774|sp|P24386|RAE1_HUMAN  RAB PROTEINS GERANYLGERANYLT...  877  0.0
gi|1585775|sp|P37727|RAE1_RAT  RAB PROTEINS GERANYLGERANYLTRA...  866  0.0
gi|13626886|sp|Q61598|GDIC_MOUSE  RAB GDP DISSOCIATION INHIB...  127  4e-29
gi|729566|sp|P39958|GDI1_YEAST  SECRETORY PATHWAY GDP DISSOC...  127  4e-29
gi|13626813|sp|Q97556|GDIB_CANFA  RAB GDP DISSOCIATION INHIB...  126  9e-29
gi|13638229|sp|P50397|GDIB_MOUSE  RAB GDP DISSOCIATION INHIB...  125  2e-28
gi|1707888|sp|P50398|GDIA_RAT  RAB GDP DISSOCIATION INHIBITO...  124  6e-28
gi|121108|sp|P21856|GDIA_BOVIN  RAB GDP DISSOCIATION INHIBIT...  124  6e-28
gi|13648812|sp|O97555|GDIA_CANFA  RAB GDP DISSOCIATION INHIB...  124  7e-28
gi|1707886|sp|P31150|GDIA_HUMAN  RAB GDP DISSOCIATION INHIBI...  123  8e-28
gi|13638228|sp|P50395|GDIB_HUMAN  RAB GDP DISSOCIATION INHIB...  122  1e-27
gi|1707889|sp|P50399|GDIB_RAT  RAB GDP DISSOCIATION INHIBITO...  121  4e-27
gi|1723467|sp|q10305|YD4C_SCHPO  PUTATIVE SECRETORY PATHWAY ...  120  6e-27
gi|1585776|sp|P32864|RAE1_YEAST  RAB PROTEINS GERANYLGERANYLT...  97  6e-20
gi|1707887|sp|P50396|GDIA_MOUSE  RAB GDP DISSOCIATION INHIBI...  75  2e-14
gi|1072024|sp|Q93831|PAEP_CANAL  RAB PROTEINS GERANYLGERANY...  74  7e-13
gi|2498411|sp|O49398|GLF_MYCSE  UDP-GALACTOPYRANOSE MUTASE  35  0.52
gi|1115401|sp|Q9KBQ9|THA_AZOVI  SOLUBLE PYRIDINE NUCLEOTID...  34  0.85
gi|1115075|sp|Q05139|THA_PSEFL  SOLUBLE PYRIDINE NUCLEOTID...  33  1.1
gi|1115195|sp|P57112|THA_PSEAE  SOLUBLE PYRIDINE NUCLEOTID...  33  1.5
gi|2915516|sp|P54480|YNAJ_BACSU  HYPOTHETICAL SYMPORTER IN G...  33  2.8
gi|1941788|sp|P30599|CHS2_NEUTM  CHITIN SYNTHASE 2 (CHITIN-UD...  32  3.0
gi|2498412|sp|P75499|GLF_MYCPN  UDP-GALACTOPYRANOSE MUTASE  32  3.4
gi|547891|sp|P36225|MAP4_BOVIN  MICROTUBULE-ASSOCIATED PROTE...  32  3.4
gi|586602|sp|P37747|GLF_ECOLI  UDP-GALACTOPYRANOSE MUTASE  32  3.8
gi|12643859|sp|O97DP4|GIG2_ARATH  FERREDOXIN-DEPENDENT GLUTA...  32  4.8
gi|586678|sp|P37637|YHIV_ECOLI  HYPOTHETICAL 111.5 KDA PROTE...  31  6.4
```

Pair-wise alignments

```
>gi|13626886|sp|Q61598|GDIC_MOUSE RAB GDP DISSOCIATION INHIBITOR BETA-2 (RAB GDI BETA-2) (GDI-3)
          Length = 445

 Score =  127 bits (320), Expect = 4e-29
 Identities = 77/264 (29%), Positives = 139/264 (52%), Gaps = 19/264 (7%)

Query: 226 GRRFNIDLVSKLLYSQGLLIDLLIKSDVSRYVEFKNVTRILAFREGKVEQVPCSRADVFN 285
           GR +N+DL+ K L + G L+ +L+ ++V+RY++FK +     ++ GK+ +VP + A+
Sbjct: 69  GRDWNVDLIPKFLMANGQLVKMLLFTEVTRYMDFKVIEGSFVYKGGKIYKVPSTEAEALA 128

Query: 286 SKELTMVEKRMLMKFLTFCLEY-EQHPDEYQAF--RQCSFSEYLKTKKLTPNLQHFVLHS 342
           S  + + EKR   KFL +   + E+ P  ++    ++ S  +  K   L  ++  F  HS
Sbjct: 129 SSLMGLFEKRRFRKFLVYVANFDEKDPRTFEGVDPKKTSMRDVYKKFDLGQDVIDFTGHS 188

Query: 343 IAM------TSESSCTTIDGLNATKNFLQCLGRFGNTPFLFPLYGQGEIPQGFCRMCAVF 396
           +A+        +  C TI   N  K + + L R+G +P+L+PLYG GE+PQGF R+ A++
Sbjct: 189 LALYRTDDYLDQPCCETI---NRIKLYSESLARYGKSPYLYPLYGLGELPQGFARLSAIY 245

Query: 397 GGIYCLRHKVQCFVVDKESGRCKAIIDHFGQRINAKYFIVEDSYLSEETCSNVQYKQISR 456
           GG Y L   ++  +V  ++G+   +   G+   K  I + SY+ +       +  Q+ R
Sbjct: 246 GGTYMLNKPIEEIIV--QNGKVVGVKSE-GEIARCKQLICDPSYVKDRV---EKVGQVIR 299

Query: 457 AVLITDQSILKTDLDQQTSILIVP 480
            + I    I  T+ D  +  +I+P
Sbjct: 300 VICILSHPIKNTN-DANSCQIIIP 322


 Score = 52.0 bits (123), Expect = 3e-06
 Identities = 22/44 (50%), Positives = 32/44 (72%)

Query: 8  EFDVVIIGTGLPESILAAACSRSGQRVLHIDSRSYYGGNWASFS 51
          E+DV+++GTGL E IL+  S +G++VLH+D  YYGG  AS +
Sbjct: 4  EYDVIVLGTGLTECILSGIMSVNGKKVLHMDQNPYYGGESASIT 47
```

BLAST report designed for human readability.

- One-line descriptions provide overview designed for human “browsing”.
- Redundant information is presented in the report (e.g., one-line descriptions and alignments both contain expect values, scores, descriptions) so a user does not need to move back and forth between sections.
- HTML version has lots of links for a user to explore.
- It can change as new features/information becomes available.

Hit-table

```
# BLASTN 2.2.2 [Dec-14-2001]
# Database: ecoli
# Query: gi|4730899 Homo sapiens genomic DNA of 21q22.1, GART and AML, f43D11-119B8 region, segment 5/10, complete seq
# Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s.
gi|4730899      gi|2367099|gb|AE000133.1|AE000133       100.00  1198    0       0       52913   54110   10943   9746
gi|4730899      gi|1789919|gb|AE000427.1|AE000427       100.00  1195    0       0       52913   54107   2347    1153
gi|4730899      gi|1789607|gb|AE000401.1|AE000401       100.00  1195    0       0       52913   54107   6037    4843
gi|4730899      gi|1788338|gb|AE000294.1|AE000294       100.00  1195    0       0       52913   54107   5700    4506
gi|4730899      gi|1787588|gb|AE000231.1|AE000231       100.00  1195    0       0       52913   54107   4146    5340
gi|4730899      gi|1786875|gb|AE000170.1|AE000170       100.00  1195    0       0       52913   54107   2321    1127
gi|4730899      gi|1786751|gb|AE000160.1|AE000160       100.00  1195    0       0       52913   54107   9133    7939
gi|4730899      gi|1788508|gb|AE000308.1|AE000308       99.92   1195    1       0       52913   54107   11740   10546
gi|4730899      gi|2367181|gb|AE000381.1|AE000381       99.83   1195    2       0       52913   54107   2030    3224
gi|4730899      gi|1788298|gb|AE000291.1|AE000291       99.58   1198    5       0       52910   54107   5290    4093
gi|4730899      gi|1787633|gb|AE000234.1|AE000234       93.21   1105    72      3       52912   54014   1146    43
```

- Contains no sequence or definition lines, but does contain sequence identifiers, starts/stops (one-offset), percent identity of match as well as expect value etc.
- Simple format is ideal for automated tasks such as screening of sequence for contamination or sequence assembly.

There are drawbacks to parsing the BLAST report and Hit-table.

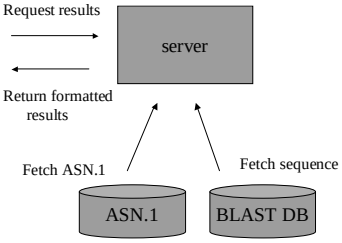
- No way to automatically check for truncated output.
- No way to rigorously check for syntax changes in the output.

Structured output allows automatic and rigorous checks for syntax errors and changes.

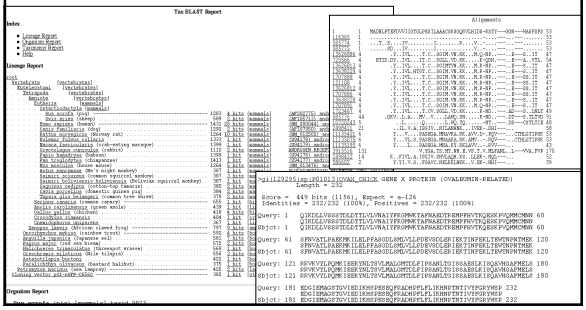
Abstract Syntax Notation 1 (ASN.1)

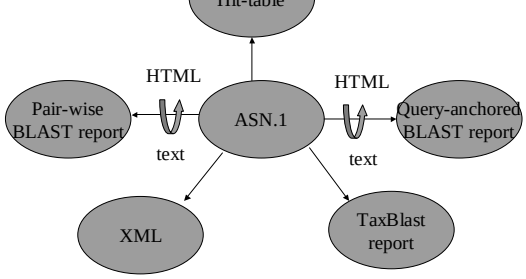
- Is an International Standards Organization (ISO) standard for describing structured data and reliably encoding it.
- Used extensively in the telecommunications industry.
- Both a binary and a text format.
- NCBI data model is written in ASN.1.
- Asntool can produce C object loaders from an ASN.1 specification.
- Datatool can produce C++ classes from an ASN.1 specification.

ASN.1 is used for the NCBI BLAST Web page.



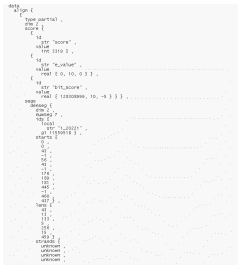
Different reports can be produced from the ASN.1 of one search.





The BLAST ASN.1 (“SeqAlign”) contains:

- Start, stop, and gap information (zero-offset).
- Score, bit-score, expect-value.
- Sequence identifiers.
- Strand information.



Three flavors of Seq-Align, Score-block(s) plus one of:

- Dense-diag: series of unconnected diagonals. No coordinate “stretching” (e.g., cannot be used for protein-nucl. alignments). Used for ungapped BLASTN/BLASTP.
- Dense-seg: describes an alignment containing many segments. No coordinate “stretching”. Used for gapped BLASTN/BLASTP.
- Std-seg: a collection of locations. No restriction on stretching of coordinates. Used for gapped/ungapped translating searches. Generic.

Score Block

SEQUENCE is an ordered list of elements, each of which is an ASN.1 type. Required unless DEFAULT or OPTIONAL

```
Score ::= SEQUENCE {
   id Object-id OPTIONAL ,      -- identifies Score type
   value CHOICE {               -- actual value
      real REAL ,               -- floating point value
      int INTEGER  } }          -- integer
```

Dense-seg definition

```
Dense-seg ::= SEQUENCE {          -- for (multiway) global or partial alignments

  dim INTEGER DEFAULT 2 ,         -- dimensionality

  numseg INTEGER ,                -- number of segments here
  ids SEQUENCE OF Seq-id ,        -- sequences in order
  starts SEQUENCE OF INTEGER ,    -- start OFFSETS in ids order within segs
  lens SEQUENCE OF INTEGER ,      -- lengths in ids order within segs
  strands SEQUENCE OF Na-strand OPTIONAL ,
  scores SEQUENCE OF Score OPTIONAL }  -- score for each seg
```

SEQUENCE OF is an ordered list of the same type of element.

Demo program ("blreplay") to reproduce BLAST results from ASN.1

- Start/stops and identifiers read in from ASN.1 (SeqAlign).
- Sequences and definition lines fetched from BLAST databases.

Asntool can produce XML from ASN.1

- Really a transliteration, not a new specification
- A Document Type Definition (DTD) can also be produced.

ASN.1 and XML validation differences.

- XML can be “well-formed” (does not break any XML syntax rules) or “validated” (checked against a DTD).
- ASN.1 must always be valid (checked against a specification).

Special purpose XML

- NCBI specification does not fit the needs of some users (the sequence is not provided in the SeqAlign, when fetched the sequence is packed 2/4 bp’s per byte).
- Possible to produce XML with more/less information or in a different format.
- First done as an ASN.1 specification, which is then dumped as XML.

BLAST XML designed to be self-contained.

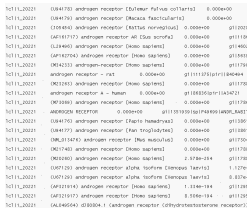
- Query sequence, database sequence, etc.
- Sequence definition lines.
- Start, stop, etc. (one-offset).
- Scores, expect values, % identity etc.
- Produced by BLAST binaries and on NCBI Web page.

Overview of the BLAST XML

```
<!ELEMENT BlastOutput (
        BlastOutput_program ,          BLAST program, e.g., blastp, etc
        BlastOutput_version ,          version of BLAST engine (e.g., 2.1.2)
        BlastOutput_reference ,        Reference about algorithm
        BlastOutput_db ,               Database(s) searched
        BlastOutput_query-ID ,         query identifier
        BlastOutput_query-def ,        query definition
        BlastOutput_query-len ,        query length
        BlastOutput_query-seq? ,       query sequence
        BlastOutput_param ,            BLAST search parameters
        BlastOutput_iterations         BLAST results for each iteration/run
)>
```

Parsing BLAST XML with Expat.

- Expat is a popular free-ware used for parsing XML.
- Non-validating.
- Simple C (demo) program to parse BLAST output.



Output sizes for a BLASTP search of gi|178628 vs. nr.

- Hit-table: 16 kb
- Binary ASN.1 (SeqAlign): 35 kb
- Text ASN.1 (SeqAlign): 144 kb
- XML (SeqAlign): 392 kb
- XML: 288 kb
- BLAST report (text): 232 kb
- BLAST report (html): 272 kb

Specification (i.e., “data model”) issues should not be confused with the question about whether to use ASN.1 or XML.

Structured output is not a panacea.

- Design issues must still be addressed.
- Semantic issues still exist, e.g. is a start/stop value zero-offset or one-offset.
- Data issues still exist, e.g., is the correct sequence shown, are the offsets correct, was the DNA translated with the correct genetic code?

IMPROVING BLAST THROUGHPUT

Use megablast to align very similar sequences.

- Best if alignments between query and target will be 97-99% identical.
- Word-size: 28; an exact match of 28-31 bases required to initiate extensions.
- A greedy gapped alignment routine with non-affine gapping (constant cost per insertion/deletion) used to perform extensions
- Use for aligning sequences from the same organism.

Example: search u93237 (Human MEN1 gene) vs human EST's with filtering for low-complexity and human repeats and expect value 1.0e-6

MEGABLAST:

- word size: 28
- run time: 19 seconds
- 469 alignments found
- 345 alignments more than 98% identical

BLASTN

- word size: 11
- run time: 148 seconds
- 491 alignments found
- 359 alignments more than 98% identical

BLASTN alignment:

```
 Score =  224 bits (113), Expect = 6e-55
 Identities = 113/113 (100%)
 Strand = Plus / Plus


Query: 6647 ggagatctacaaggagttctttgaagtagccaatgatgtcatccccaacctgctgaaggga 6706
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 79   ggagatctacaaggagttctttgaagtagccaatgatgtcatccccaacctgctgaaggga 138


Query: 6707 ggcagccagcttgctggaggcgggcgaggagcggcgggggagcaaagccagg 6759
            |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 139  ggcagccagcttgctggaggcgggcgaggagcggcgggggagcaaagccagg 191
```

MEGABLAST alignment:

```
 Score =  228 bits (115), Expect = 4e-56
 Identities = 136/142 (95%), Gaps = 6/142 (4%)
 Strand = Plus / Plus


Query: 6623 ctacaa-ctac-tgccgggaa-gac-ga-ggagatctacaaggagttctttgaagtagcc 6677
            |||||| ||| || ||||||| ||| || ||||||||||||||||||||||||||||||
Sbjct: 51   ctacaagctacatg-cgggaatgacggatggagatctacaaggagttctttgaagtagcc 109


Query: 6678 aatgatgtcatccccaacctgctgaaggaggcagccagcttgctggaggcgggcgaggag 6737
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 110  aatgatgtcatccccaacctgctgaaggaggcagccagcttgctggaggcgggcgaggag 169


Query: 6738 cggccgggggagcaaagccagg 6759
            ||||||||||||||||||||
Sbjct: 170  cggccgggggagcaaagccagg 191
```

Increasing the threshold for blast[px]/tblast[nx] speeds up search.

- these programs use exact and "neighboring" (three-letter) words as initial hits.
- increasing threshold decreases the number of neighboring words.
- fewer neighboring words mean fewer extensions.
- if the threshold is a high value (e.g., 100) only exact matches are used.
- more subtle alignments will be missed.

Example: blastx search of NT_078011 (41887 bases human contig) against nr.

Threshold	Time (seconds)	Alignments with expect <= 1.0e-6
12	4050	962
13	2452	955
14	1581	954
15	1139	947
17	770	916
100	658	879

```
>ref|NP_524599.1| Serotonin receptor 7 CG12073-PA [Drosophila melanogaster] ...
          Length = 564

 Score = 63.2 bits (152), Expect = 4e-07
 Identities = 28/70 (40%), Positives = 37/70 (52%)
 Frame = -3

Query: 41255 ARGKSFTFVLAVVMGVFVLCWFPFFFSYSLYGICREACQVPGPLFKFFFWIGYCNSSLNP 41076
           A+ K  +  L ++M  F +CW PFF    +     E   VP  L   F W+GY NS LNP
Sbjct: 447   AKEKKASTTLGIIMSAFTVCWLPFFILALIRPF--ETMHVPASLSSLFLWLGYANSLLNP 504

Query: 41075 VIYTVFNQDF 41046
           +IY   N+DF
Sbjct: 505   IIYATLNRDF 514
```

This alignment is found with thresholds 12,13,14,15, and 17. It is not found if only exact matches are used (threshold = 100).

The default mode of blast[px] and tblast[nx] requires two hits on the same diagonal to initate an extension.

Do as little work (with the database) as possible.

- Scanning the database and/or reading it from disk can be a significant portion of some searches.
- Use the “concatenation” feature of megablast to concatenate multiple queries and scan the database only once for all queries.
- The OS will cache recently used files. Make use of this by grouping together queries for one database before searching another one.
- Buy more memory if you cannot fit a database into memory.

MEGABLAST concatenates queries

- searching 50 human EST’s (total of 15,700 bases) against the human EST database took 23.5 seconds (668 bases/second).
- searching one human EST (gi|272208, 211 bases) took 13 seconds (16 bases/second).
- Concatenation minimizes time spent scanning the database.
- The more stringent the search, the more the savings.

Three different strategies to search three EST’s against the nt and est databases.

- 26 minutes (wall clock time) if searches are grouped by query.
- 10 minutes if searches are grouped by database.
- 7.5 minutes if megablast concatenates the queries.

BLAST DATABASES

BLAST databases:

- can be produced with stand-alone formatdb and a FASTA file.
- are always (?) produced with the formatdb API (e.g., stand-alone formatdb).
- are almost always read with the readdb API (recommended).
- pack nucleotide sequences 4-to-1.
- are architecture independent.

The (physical) BLAST databases comprise files in binary format.

pin or nin	Index into sequence and header files
psq or nsq	Sequence data
phr or nhr	Sequence identifier, definition, taxonomic information, etc. stored as binary ASN.1
pni or nni *	ISAM index file for GI identifiers
pnd or nnd *	ISAM data file for GI identifiers
psi or nsi *	ISAM index file for other identifiers
psd or nsd *	ISAM data file for other identifiers

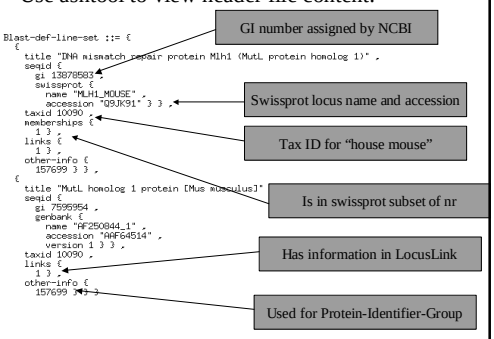
* only created if “-o” option used with formatdb.

ASN.1 spec. used for header files

```
-- one Blast-def-line-set for each entry
Blast-def-line-set ::= SEQUENCE OF Blast-def-line

Blast-def-line ::= SEQUENCE {
   title VisibleString OPTIONAL,            -- simple title
   seqid SEQUENCE OF Seq-id,                -- Regular NCBI Seq-Id
   taxid  INTEGER OPTIONAL,                 -- taxonomy id
   memberships SEQUENCE OF INTEGER OPTIONAL, -- bit arrays
   links SEQUENCE OF INTEGER OPTIONAL,      -- bit arrays
   other-info SEQUENCE OF INTEGER OPTIONAL  -- future use
}
```

Use asntool to view header file content.



asntool -m fastadl.asn -M asn.all -d nr.phr -t Blast-def-line-set -p stdout

Asntool can also produce header files in XML.

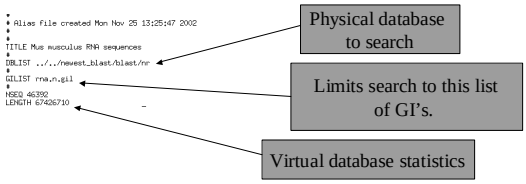
```
<Blast-def-line-set>
  <Blast-def-line>
    <Blast-def-line_title>DNA mismatch repair protein Mlh1 (MutL protein homolog 1)</Blast-def-line_title>
    <Blast-def-line_seqid>
      <Blast-def-line_seqid_E>
        <Seq-id_gi>13878503</Seq-id_gi>
      </Blast-def-line_seqid_E>
      <Blast-def-line_seqid_E>
        <Seq-id_swissprot>
          <Textseq-id>
            <Textseq-id_name>MLH1_MOUSE</Textseq-id_name>
            <Textseq-id_accession>Q9JK91</Textseq-id_accession>
          </Textseq-id>
        </Seq-id_swissprot>
      </Blast-def-line_seqid_E>
    </Blast-def-line_seqid>
    <Blast-def-line_taxid>10090</Blast-def-line_taxid>
    <Blast-def-line_memberships>
      <Blast-def-line_memberships_E>1</Blast-def-line_memberships_E>
    </Blast-def-line_memberships>
    <Blast-def-line_links>
      <Blast-def-line_links_E>1</Blast-def-line_links_E>
    </Blast-def-line_links>
    <Blast-def-line_other-info>
      <Blast-def-line_other-info_E>157699</Blast-def-line_other-info_E>
    </Blast-def-line_other-info>
  </Blast-def-line>
  <Blast-def-line>
    <Blast-def-line_title>MutL homolog 1 protein [Mus musculus]</Blast-def-line_title>
    <Blast-def-line_seqid>
      <Blast-def-line_seqid_E>
        <Seq-id_gi>7595954</Seq-id_gi>
      </Blast-def-line_seqid_E>
      <Blast-def-line_seqid_E>
        <Seq-id_genbank>
          <Textseq-id>
            <Textseq-id_name>AF250844_1</Textseq-id_name>
            <Textseq-id_accession>AAF64514</Textseq-id_accession>
            <Textseq-id_version>1</Textseq-id_version>
          </Textseq-id>
        </Seq-id_genbank>
      </Blast-def-line_seqid_E>
    </Blast-def-line_seqid>
    <Blast-def-line_taxid>10090</Blast-def-line_taxid>
    <Blast-def-line_links>
      <Blast-def-line_links_E>1</Blast-def-line_links_E>
    </Blast-def-line_links>
    <Blast-def-line_other-info>
      <Blast-def-line_other-info_E>157699</Blast-def-line_other-info_E>
    </Blast-def-line_other-info>
  </Blast-def-line>
</Blast-def-line-set>
```

asntool -m fastadl.asn -M asn.all -d nr.phr -t Blast-def-line-set -x stdout

Alias files

- Virtual databases uses alias files to instruct BLAST which physical database(s) to search.
- Alias files have extensions ".nal" or ".pal.
- Alias files can specify multiple databases.
- The search can be limited by a list of GI's or ordinal id's (sequences in BLAST database) specified in the alias file.
- Alias files hide physical databases with the same (root) name.

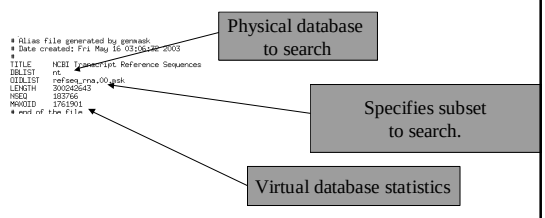
Alias file using GI list



The GI list can be either text or binary; formatdb can produce a binary GI list from a text one.

Formatdb can be used to produce this alias file with statistics.

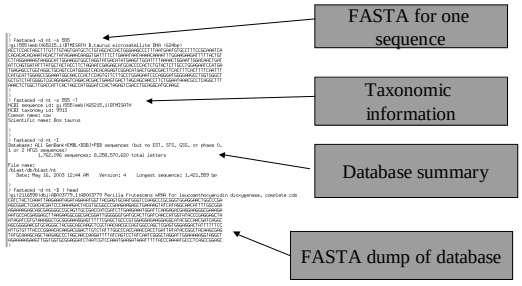
Alias file using ordinal ID list.



Very large data sets.

- An array of four byte integers is used in the pin/nin file to record offsets in the sequence and header files.
- The header and sequence files are limited to about 4 gig (4 billion residues or 16 billion bases) for each physical database.
- Using four byte integers for the offsets means the pin/nin file stays small for all databases.
- Multiple volumes are used for larger data sets.
- Multiple volumes are produced by formatdb as needed along with the corresponding alias file.

Fastacmd: a command-line program to query BLAST databases.



Resources

- BLAST Home page: http://www.ncbi.nlm.nih.gov/BLAST/
- NCBI Information Engineering Branch home page: http://www.ncbi.nlm.nih.gov/IEB/
- Demonstration programs (parsing XML with EXPAT, blreplay.c, doblast.c, db2fasta.c): ftp://ftp.ncbi.nih.gov/blast/demo
- NCBI Handbook chapter: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowSection&rid=handbook.chapter.610

ASN.1 RESOURCES

- The Open Book : A Practical Perspective on OSI by Marshall T. Rose (Prentice Hall).
- OSS Nokalva Web site: http://www.oss.com/asn1/overview.html
- NCBI toolkit documentation on ASN.1:

http://www.ncbi.nlm.nih.gov/IEB/ToolBox/SDKDOCS/ASNLIB.HTML

Email addresses

- General questions about running BLAST: blast-help@ncbi.nlm.nih.gov
- Questions about compiling the toolkit and requests for hard-copy of documentation: toolbox@ncbi.nlm.nih.gov

Selected BLAST References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389-402.
- Altschul SF. Amino acid substitution matrices from an information theoretic perspective. J Mol Biol. 1991 Jun 5;219(3):555-65.
- Altschul SF, Boguski MS, Gish W, Wootton JC. Issues in searching molecular sequence databases. Nat Genet. 1994 Feb;6(2):119-29.

(Some of the) People (currently) working on BLAST

- Kevin Bealer
- Christiam Camacho
- George Coulouris
- Ilya Dondoshansky
- Tom Madden
- Yuri Merezhuk
- Yan Raytselis
- Jian Ye
- Richa Agarwala
- Stephen Altschul
- Peter Cooper
- Susan Dombrowski
- David Lipman
- Wayne Matten
- Scott McGinnis
- Alexander Morgulis
- Alejandro Schaffer
- Tao Tao
- David Wheeler