

Heuristic Alignment and Searching

Mark Voorhies

3/28/2012

Types of alignments

Global Alignment Each letter of each sequence is aligned to a letter or a gap (e.g., Needleman-Wunsch).

Local Alignment An optimal pair of subsequences is taken from the two sequences and globally aligned (e.g., Smith-Waterman).

Types of alignments

Global Alignment Each letter of each sequence is aligned to a letter or a gap (e.g., Needleman-Wunsch).

Local Alignment An optimal pair of subsequences is taken from the two sequences and globally aligned (e.g., Smith-Waterman). *This tends to be more biologically relevant.*

The implementation of local alignment is the same as for global alignment, with a few changes to the rules:

- Initialize edges to 0 (*no penalty for starting in the middle of a sequence*)
- The maximum score is never less than 0, and no pointer is recorded unless the score is greater than 0 (*note that this implies negative scores for gaps and bad matches*)
- The trace-back starts from the highest score in the matrix and ends at a score of 0 (*local, rather than global, alignment*)

Because the naive implementation is essentially the same, the time and space requirements are also the same.

Timing CLUSTALW

Timing CLUSTALW from the command line:

```
for i in 50 100 150 200 250 300 350 400 450; do
  head -n $i -q G217B_iron.fasta Pb01_iron.fasta > temp.fasta;
  time clustalw -infile=temp.fasta -type=DNA -align;
done
```

The output looks like this:

```
Sequences (1:2) Aligned. Score: 0
Guide tree file created: [temp.dnd]
```

```
There are 1 groups
Start of Multiple Alignment
```

```
Aligning...
Group 1:                               Delayed
Alignment Score 7238
```

```
CLUSTAL-Alignment file created [temp.aln]
```

```
real 0m3.400s
user 0m3.388s
sys 0m0.012s
```

You can copy the timing results into Excel.
Here is a Python script that does the same thing:

```
#!/usr/bin/env python
# Time-stamp: <ParseTimes.py 2011-03-29 21:10:59 Mark Voorhies>
"""Parse wall times from a log file on stdin and write them as a CSV
formatted column for Excel/OpenOffice/etc on stdout.  If command line
arguments are given, treat them as a second output column."""

from csv import writer
import re
time_re = re.compile("(?P<minutes>[\d]+)m(?P<seconds>[\d]+\.[\d]+)s", re.M)

if (__name__ == "__main__"):
    import sys
    args = sys.argv[1:]
    out = writer(sys.stdout)
    i = 0
    for t in time_re.finditer(sys.stdin.read()):
        try:
            y = args[i]
            i += 1
        except IndexError:
            y = ""
        out.writerow(
            (float(t.group("minutes"))*60+float(t.group("seconds")), y))

    del out
```

You can fit the timing results to a curve in Excel.

$$y = Ax^B \quad (1)$$

$$\log y = \log Ax^B \quad (2)$$

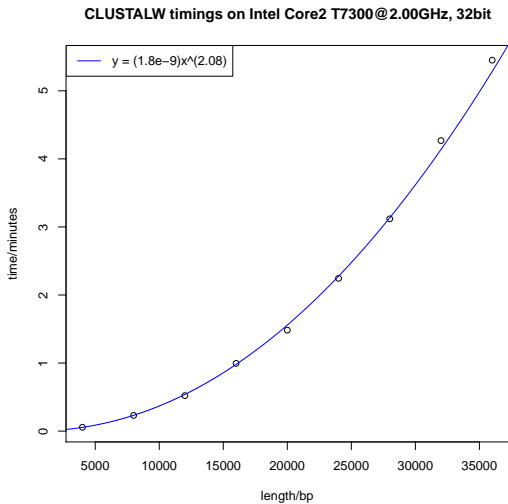
$$= \log A + B \log x \quad (3)$$

$$= A' + B \log x \quad (4)$$

Here is an R script that does the same thing:

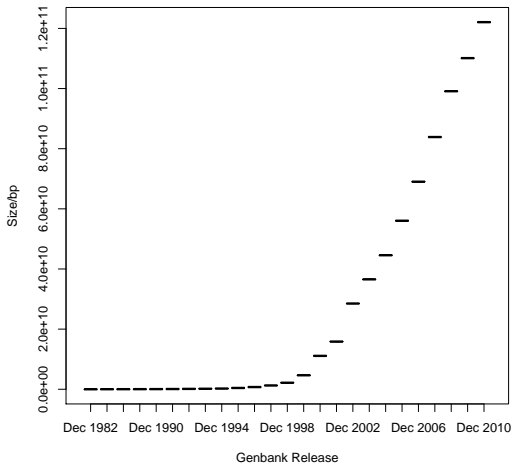
```
data <- read.csv("timings.csv", header = FALSE, col.names = c("t","n"))
x <- log(data$n*80)
y <- log(data$t/60)
f <- lm(y ~ x)
x0 <- 0:40000
a <- exp(f$coeff[1])
b <- f$coeff[2]
pdf("ClustalwTimings.pdf")
plot(data$n*80, data$t/60, xlab = "length/bp", ylab = "time/minutes",
      main = "CLUSTALW_timings_on_Intel_Core2_T7300@2.00GHz,_32bit")
points(x0, a*x0^b, col = "blue", type = "l")
legend("topleft", c("y=-(1.8e-9)x^(2.08)"), col = "blue", lty = 1)
dev.off()
```

CLUSTALW takes $O(MN)$ time



$O(MN)$ time is too slow!

source: <ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt>



Why BLAST?

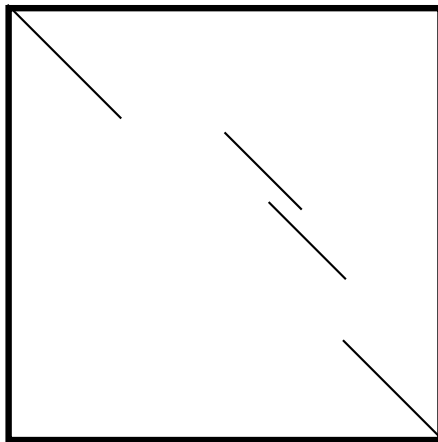
- Fast, heuristic approximation to a full Smith-Waterman local alignment
- Developed with a statistical framework to calculate expected number of false positive hits.
- Heuristics biased towards “biologically relevant” hits.

Seeding searches

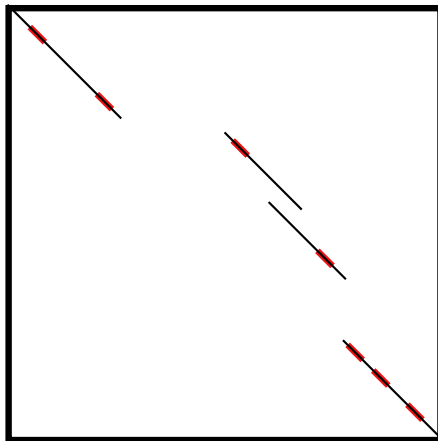
Most of the magic in a sequence-search tool lives in its indexing scheme

Program	Purpose	Indexing
BLAST	Database searching	Target indexing, 3aa or 11nt words
BLAT	mRNA mapping	Query indexing
BOWTIE	RnaSeq	Specialized index for low quality, short reads
e-PCR	Simulated PCR	Annealing-oriented index

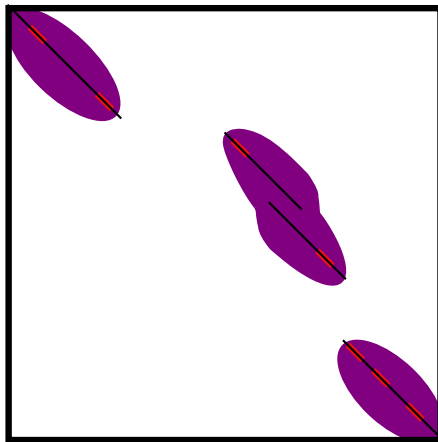
BLAST: A quick overview



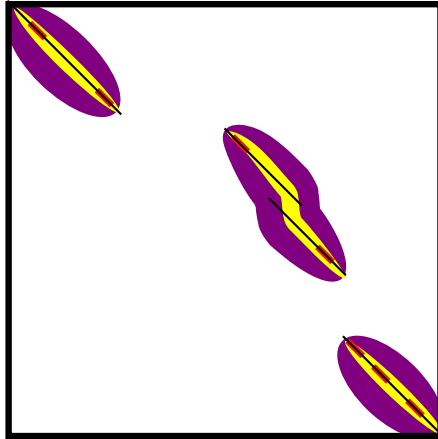
BLAST: Seed from exact word hits



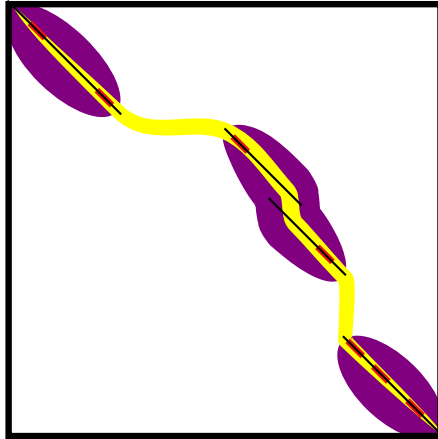
BLAST: Myers and Miller local alignment around seed pairs



BLAST: High Scoring Pairs (HSPs)



Gapped BLAST: Merge neighboring HSPs



How fast is BLAST?

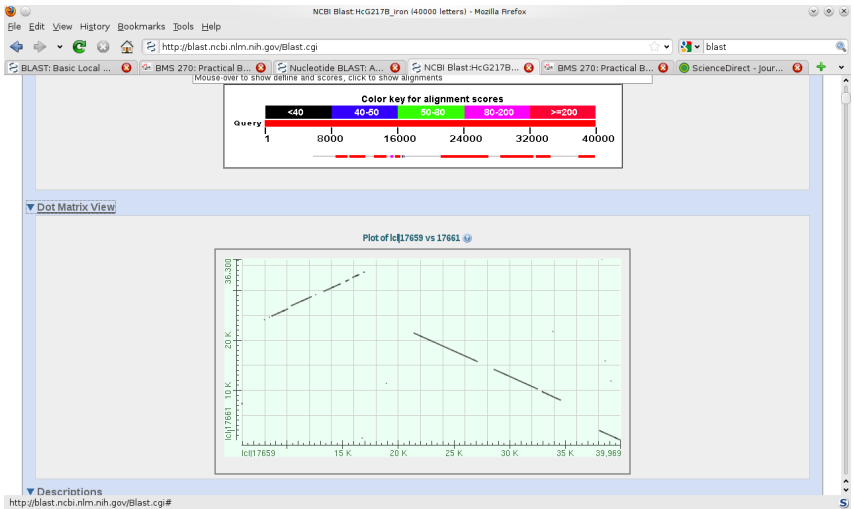
The screenshot shows a web browser window titled "Nucleotide BLAST: Align two or more sequences using BLAST - Mozilla Firefox". The address bar shows the URL: `http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=MegaBlast&PROGRAM=blastn&BLAST_PROGRAMS=megaBlast&PAGE=...`. The browser tabs include "BLAST: Basic Local...", "BMS 270: Practical B...", "Nucleotide BLAST: A...", "NCBI Blast:HcG217B...", "BMS 270: Practical B...", and "ScienceDirect - jour...".

The main content area is the NCBI BLAST interface, titled "BLASTN programs search nucleotide subjects using a nucleotide query, more...". It features several sections:

- Enter Query Sequence:** A text input field for the query sequence, a "Clear" button, and a "Query subrange" section with "From" and "To" input fields.
- Or, upload file:** A file selection button showing the path `/home/mvoorhie/Projects/Cc` and a "Browse..." button.
- Job Title:** A text input field with the placeholder text "Enter a descriptive title for your BLAST search".
- Align two or more sequences:** A checked checkbox.
- Enter Subject Sequence:** A text input field for the subject sequence, a "Clear" button, and a "Subject subrange" section with "From" and "To" input fields.
- Or, upload file:** A file selection button showing the path `/home/mvoorhie/Projects/Cc` and a "Browse..." button.
- Program Selection:** A section titled "Optimize for" with three radio button options:
 - Highly similar sequences (megablast)
 - More dissimilar sequences (discontiguous megablast)
 - Somewhat similar sequences (blastn)A "Choose a BLAST algorithm" link is also present.

The status bar at the bottom left of the browser window shows "Done".

How fast is BLAST?



How fast is BLAST?

```
time bl2seq -p blastn -i G217B_iron.fasta -j Pb01_iron.fasta -e 1e-6 > temp.blastn
```

```
real    0m0.342s  
user    0m0.080s  
sys     0m0.032s
```

The basic flavors of BLAST

Target Query	Protein	DNA
Protein	BLASTP	TBLASTN
DNA	BLASTX	BLASTN TBLASTX

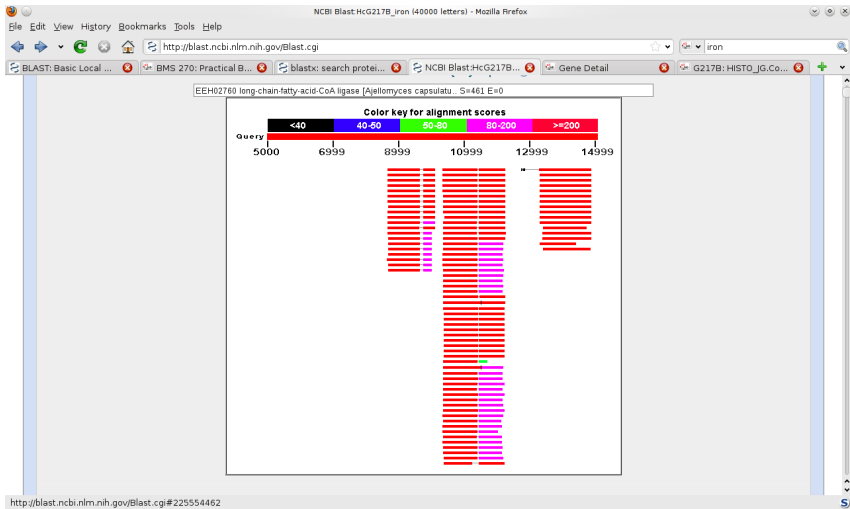
BLASTX: Nucleotide query vs. Protein Database

The screenshot shows the NCBI BLASTX web interface in a Mozilla Firefox browser window. The browser title is "blastx: search protein databases using a translated nucleotide query - Mozilla Firefox". The address bar shows the URL: "http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastx&BLAST_PROGRAMS=blastx&PAGE_TYPE=BlastSearch&SI...". The search results bar shows "size of nr ncbi".

The interface includes the following sections:

- Exclude** (Optional):
 - Models (XM/XP)
 - Uncultured/environmental sample sequences
- Entrez Query** (Optional):
 - Enter an Entrez query to limit search
- BLAST** button
- Search database** Non-redundant protein sequences (nr) using Blastx (search protein databases using a translated nucleotide query)
- Show results in a new window
- Algorithm parameters**
 - General Parameters**
 - Max target sequences**: 100 (Select the maximum number of aligned sequences to display)
 - Expect threshold**: 10
 - Word size**: 3
 - Max matches in a query range**: 0
 - Scoring Parameters**
 - Matrix**: BLOSUM62
 - Gap Costs**: Existence: 11 Extension: 1
 - Filters and Masking**
 - Filter**: Low complexity regions
 - Mask**:
 - Mask for lookup table only
 - Mask lower case letters

BLASTX: Nucleotide query vs. Protein Database



Sometimes it's still worth running locally...

NCBI Blast:HcG217B_iron (40000 letters) - Mozilla Firefox

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

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI [Sign In] [Logout]

NCBI/BLAST/blastx/Formatting Results - T73YCW0E01S

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options...](#) [Download](#)

An error has occurred on the server. Please, contact blast-help@ncbi.nlm.nih.gov

Informational Message: [blastsrv4.REAL]: Error: CPU usage limit was exceeded, resulting in SIGXCPU (24).

HcG217B_iron (40000 letters)

Query ID	lcl 2207	Database Name	nr
Description	HcG217B_iron	Description	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule type	nucleic acid	Program	BLASTX 2.2.25+ Citation
Query Length	40000		

No significant similarity found. For reasons why, [click here](#)

Other reports: [Search Summary](#)

Sequence Viewer

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback

NCBI | BLAST | Help | Databases

Done

S

$$E = kmne^{-\lambda S} \quad (5)$$

- S : HSP score
- E : Expected number of “random” hits in a database of this size scoring *at least* S .
- m : Query length
- n : Database size
- k : Correction for similar, overlapping hits
- λ : normalization factor for scoring matrix

$$E = kmne^{-\lambda S} \quad (5)$$

- S : HSP score
- E : Expected number of “random” hits in a database of this size scoring *at least* S .
- m : Query length
- n : Database size
- k : Correction for similar, overlapping hits
- λ : normalization factor for scoring matrix

A variant of this formula is used to generate sum probabilities for combined HSPs.

$$E = kmne^{-\lambda S} \quad (5)$$

- S : HSP score
- E : Expected number of “random” hits in a database of this size scoring *at least* S .
- m : Query length
- n : Database size
- k : Correction for similar, overlapping hits
- λ : normalization factor for scoring matrix

A variant of this formula is used to generate sum probabilities for combined HSPs.

$$p = 1 - e^{-E} \quad (6)$$

$$E = kmne^{-\lambda S} \quad (5)$$

- S : HSP score
- E : Expected number of “random” hits in a database of this size scoring *at least* S .
- m : Query length
- n : Database size
- k : Correction for similar, overlapping hits
- λ : normalization factor for scoring matrix

A variant of this formula is used to generate sum probabilities for combined HSPs.

$$p = 1 - e^{-E} \quad (6)$$

(If you care about the difference between E and p , you're already in trouble)

Important points:

- Extreme value distribution
- Assumption of infinite sequence length
- No rigorous framework for gap statistics (hmmer3 tries to fill this gap)

- BLAST is very fast, at the expense of not guaranteeing globally optimal results

Summary

- BLAST is very fast, at the expense of not guaranteeing globally optimal results
- But the trade-offs that it makes are biased towards “biologically relevant” results

Summary

- BLAST is very fast, at the expense of not guaranteeing globally optimal results
- But the trade-offs that it makes are biased towards “biologically relevant” results
- And it provides a statistical framework for evaluating its results.

Summary

- BLAST is very fast, at the expense of not guaranteeing globally optimal results
- But the trade-offs that it makes are biased towards “biologically relevant” results
- And it provides a statistical framework for evaluating its results.
- We can, and should, treat our computer work as we would an experiment:
 - Document protocols and observations
 - Run positive and negative controls
 - Keep results organized and dated

- Search your favorite proteins and collate interesting hits in one FASTA file per query – play with adding informative names and annotations (we will use these FASTA files tomorrow).
- Play with the BLAST book protocols (chapter 9) on the NCBI website
- Play with positive and negative controls (including permuted sequences)