BLAST Analysis for the MAQC project

Damir Herman
NCBI/NLM/NIH
damir at nih dot gov

June 2, 2006

Abstract

This document goes into details about the parameters used in the probe mapping. We used the human RefSeq database downloaded on March 8, 2006 from the NCBI web site. As we were interested in perfect matches only, we discuss some BLAST command line flags that may be more appropriate for oligo probes, with a possibility of extending the analysis to cross-hybridizing probes.

1 RefSeq

As explained in the **README.txt** file that accompanies data files and mapping tables from the supplementary material, we obtained the human RefSeq database on March 8, 2006 by issuing the following query on the NCBI web site:

```
human[orgn] AND "biomol mrna"[Properties] AND "srcdb refseq"[Properties]
NOT "srcdb refseq model"[Properties]
```

We stored the output in the GenBank format (from the pull-down menu) in the file called `RefSeq_NMz_030806.gb`.

2 BLAST

We downloaded the executables for the latest BLAST version (at the time it was blast-2.2.11), from `ftp://ftp.ncbi.nih.gov/blast/executables/LATEST/`
Please refer to the provided documentation about setting up the lookup path for BLAST.
From the gb file, we generated the fasta file `RefSeq_NMz_030806` and formated it for usage with the standalone BLAST by issuing the following *formatdb* command:

```
formatdb -p F -o T -i RefSeq_NMz_030806
```

For the analysis, we used *blastn*. Here is a short list of switches that modify *blastn*'s default behavior [1, Ch.13]:

-G Initial penalty for opening a gap of length 0. A penalty for extending the gap is controlled by -E. -G 0 invokes default behavior (*blastn 5, others 11*), and setting -G to zero is impossible, unless -g F is set, which turns gaping off. The default gap costs for programs other than blastn depend on the scoring matrix; the value here is for the default BLOSUM62 matrix.

Used -G 2.

-E The penalty for each gap character (*default: blastn 2, others 1*). The -G parameter controls the initial cost of opening a gap. Note that -E 0 is synonymous with the default behavior and it is impossible to set -E to zero unless -g F is set, which turns gaping off.

Used -E 1.

-q Sets the penalty for a nucleotide mismatch (*negative integer, default: -3*). Also see -r. The choices of [Integer] for -q and -r are very important because they determine your target frequencies. The default values -r 1 -q -3 are most effective for aligning sequences that are 99% identical.

Used -q -1.

-r Sets the score of a nucleotide match (*default: 1*).

-W Sets the word size for the initial word search. The minimum word size for blastn is 7 (*default: blastn 11, others 3*). It is estimated that cutting down the word size by one character slows the search down three times.

Used -W 9.

So, instead of the +1/-3 scoring scheme which should give a target frequency of 99% identity, we used +1/-1 that would give a target frequency of 75% identity for ungapped alignments between sequences of infinite length [1, App.B]. In practice, for short sequences and gap alignment these numbers are fairly close.
The exact blastall command line was

```
blastall -p blastn -d RefSeq_NMz_030806 -i <query_file> -o <output_file>
  -F F -G 2 -E 1 -q -1 -m 9 -W 9 -b 50 -v 50
```

The run time scaled with the number of probes and their length. It took between 10-20 minutes for the alternative platforms (TAQ, QGN and GEX) and the cDNA array (EPP), 8-9 hours for oligo arrays and $\simeq$ 2 days of total CPU time for AFX to execute the search on AMD Opteron Dual Processor 248 nodes with 4 gig of memory on the NIH biowulf cluster, `http://biowulf.nih.gov`.

The `<output_file>` was parsed through a script that had pre-hashed all NCBI annotation for every accession and we filtered out only the perfect matches: the alignment length had to be equal to the probe length, without any mismatches or gaps. This output was used for generation of mapping tables.

References

[1] Ian Korf, Mark Yandell & Joseph Bedell, *BLAST*, O'Reilly & Associates, Sebastopol (2003).