ESTmapper: A Tool for Fast Genome Mapping of Large cDNA Sequence Sets

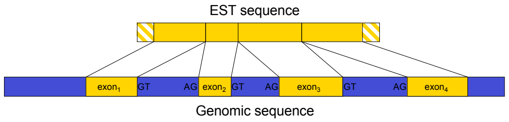
Brian Walenz and Liliana Florea, Informatics Research, Applied Biosystems, Rockville MD, 20850

ABSTRACT

Mapping large sets of cDNA and other sequence features to a genome is an essential precursor to a number of critical bioinformatics workflows, including genome annotation, SNP discovery and detection of alternative splicing. ESTmapper is a software package developed for high-throughput mapping of large cDNA data sets to a eukaryotic genome. It uses an efficient 20-mer index to rapidly locate genomic regions potentially containing the query, then generates a nucleotide-level spliced alignment between the query and each of the selected regions using an optimized version of the sim4 [1] algorithm. We evaluated ESTmapper against BLAT [2] using the dbEST sequence set (http://www.ncbi.nih.gov/dbEST/), and found the results comparable both in terms of sensitivity and computational performance. With the amount of genomic data ever increasing, ESTmapper can become an essential tool for large genome sequencing and analysis projects.

INTRODUCTION

Mapping an expressed DNA sequence on a target genome entails determining its exon model and location on the genome, including clear delimitation of the exon and intron boundaries, alignment quality indicators, and the orientation of the gene from which the sequence was sampled. Sim4 [1], EST_GENOME [3] and Spidey [4] were all designed to align a cDNA and a genomic sequence containing that gene, however, they are not capable of handling the massive mapping tasks required to produce an up-to-date annotation.



ESTmapper maps the 5.7 million human dbEST sequences to the human genome in 250 CPU hours, roughly two orders of magnitude faster than any of the methods above, and achieves the same accuracy as sim4, which it uses to produce the nucleotide-level alignment.

Salient features:

- Position index of all 20-mers in the genome for efficient match detection
- Filtering of genomic regions to identify likely matches
- Detection of multiple occurrences of the query in the genomic sequence
- Optional identification of best-matches
- Large-scale computing capability applicable to whole chromosomes and even whole genomes
- High-throughput processing of sequences
- Parallel operation in multi-processor environments
- C++ and Perl modules, portability to IBM AIX, HP Tru64, Sun Solaris, FreeBSD, OS-X and Linux

METHOD

ESTmapper generates a spliced alignment between the query and the target genome in three stages. Stage one detects genomic regions potentially containing the query, starting from exactly matching 20-mers and grouping them in chains consistent with the feature model, for instance allowing for introns. Stage two selects a subset of the spanned regions, based on the extent and depth of coverage of the query. Stage three produces nucleotide-level alignments of the query and selected regions, using an optimized version of the sim4 algorithm.

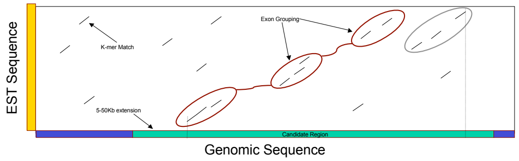
Stage 1: Signal Finding

Hash Table: ESTmapper uses a position index of all words of size k (k-mers) in the genome to quickly locate genomic regions of potential interest. A second table stores a list of k-mers that occur more than T times in the genome. K-mers in this list are ignored, as they lead to numerous false positives. Varying the parameters k and T does not dramatically affect ESTmapper's results, but has a significant effect on memory utilization and CPU time.

Percent mapped, ESTs mapped per second (ESTmapper default)`

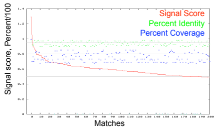
	k=14	k=17	k=20	k=23	k=26
T=100	83.508, 0.857	83.665, 1.688	83.650, 8.041	83.672, 11.002	83.686, 13.160
T=250	83.421, 0.982	83.591, 1.604	83.619, 6.324	83.641, 9.071	83.674, 10.542
T=500	83.425, 1.049	83.575, 1.575	83.608, 5.965	83.650, 8.623	83.688, 10.158
T=750	83.439, 1.078	83.575, 1.602	83.620, 6.114	83.654, 9.090	83.690, 10.500
T=1000	83.449, 1.098	83.583, 1.618	83.624, 6.230	83.657, 9.463	83.670, 11.895
T=2000	83.504, 1.218	83.616, 1.703	83.622, 6.646	83.645, 10.865	83.670, 13.358
T=4000	83.505, 1.170	83.606`, 1.782	83.609, 7.260	83.646, 11.232	83.675, 14.341

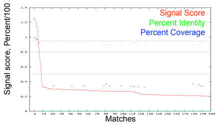
Chaining: For each cDNA and each genomic sequence, ESTmapper produces a list of k-mer matches and their positions in the two sequences. These matches are grouped and chained to identify genomic regions potentially containing the cDNA. Individual regions are then extended by 5 to 50Kb depending on the size of the unmatched fragment at either end of the cDNA sequence. Lastly, overlapping genomic areas are coalesced to create candidate regions.



Stage 2: Signal Filtering

Candidate regions are scored by the portion of the query contained in exact 20-mer matches, and the top scoring regions are selected for alignment in Stage 3. Examples of filtering of candidate regions for two repetitive ESTs are shown below.





Stage 3: Alignment

Spliced alignments between the query and each of the selected genomic regions are obtained using a version of sim4 [1], modified to allow multiple non-overlapping matches, and optimized for processing large sequence databases. The *coverage* (fraction of the query contained in the alignment) and *percent sequence identity* statistics of the spliced alignment are then used to determine the validity of the match. Typically, matches with at least 50% coverage and 95% identity are considered significant and reported.

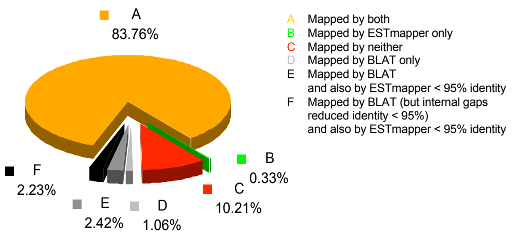
RESULTS

We evaluated the performance of ESTmapper and BLAT [2] by mapping NCBI dbEST (http://www.ncbi.nih.gov/dbEST/) sequences to the human genome assembly Build 34.

While BLAT reports 5% more sequences mapped than ESTmapper, a large portion (~2.2%) of the additional matches is due to BLAT favoring and retaining selected high-similarity sections of the match over the entire alignment, which would have <95% sequence identity. Another significant percentage (~2.4%) is regained when ESTmapper's thresholds are decreased to 90% identity.

	Number of ESTs Mapped At 50% coverage and 95% identity	CPU Time Used IBM p690, 1.3GHz
ESTmapper	4,806,136 (84.8%)	251.5 hours
BLAT	5,093,804 (89.9%)	262.6 hours

We analyzed the nature of differences between the two sets of matches using a set of 100,000 randomly selected EST sequences. Of the 5709 sequences found by BLAT but not by ESTmapper, our method produces more complete alignments that result in sequence identity lower than the 95% in 4652 of the cases. Additionally, each tool has a relatively small set that it maps exclusively (327 for ESTmapper, 1057 for BLAT).



CONCLUSION

Sequence databases are growing at an unprecedented pace, with hundreds of thousands of new sequences added to NCBI's RefSeq, dbEST and dbSNP repositories each month. Fast mapping of these features on a target genome is a critical task for keeping the genome annotations current. Extensively used to produce alignment data for Celera's annotation, SNP and mRNA repositories, ESTmapper can become an essential tool for new large genome sequencing and analysis projects.

REFERENCES

1. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. A computer program for aligning a cDNA sequence with a genomic sequence". Genome Res., 1998 8(9):964-974.
2. Kent WJ. "BLAT—the BLAST-like alignment tool." Genome Res., 2002 12(4):656-665.
3. Mott R. "EST_GENOME: a program to align spliced DNA sequence to unspliced genomic Dna", CABIOS, 1997 13(4):477-478.
4. Wheelan SJ, Church DM, Ostell JM. "Spidey: A tool for mRNA-to-Genomic Alignments", Genome Res., 2001 11(11):1952-1957.