CABIOS APPLICATIONS NOTE

Vol. 13 no. 4 1997
Pages 477–478

EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA

Richard Mott

This note describes the program EST_GENOME for aligning spliced DNA to unspliced genomic DNA. It is written in ANSI C and has been tested under Digital OSF3.2. The source code and documentation are available from ftp://www.sanger.ac.uk/ ftp/pub/ badger/est_genome.2.tar.Z.

The prediction of genes in uncharacterized genomic DNA sequence is currently one of the main problems facing sequence annotators. Methods based on *de novo* prediction, e.g. searching for motifs like the splice-site consensus, or on statistical properties such as biased codon usage, etc. (Solovyev *et al.*, 1994; Hebsgaard *et al.*, 1996) have been only partially successful, and investigators have often found that the surest way of predicting a gene is by alignment with a homologous protein sequence (Birney *et al.*, 1996; Gelfand *et al.*, 1996; Huang and Zhang, 1996), or a spliced gene product [an expressed sequence tag (EST), mRNA or cDNA], particularly now that a large number of ESTs are available (Hillier *et al.*, 1996).

Standard alignment tools are not ideal for finding the correct alignment of a spliced product to genomic DNA, because of the large introns which can occur in the genomic sequence and because the programs ignore the conserved sequences found at donor/acceptor splice sites (intron/exon boundaries). In addition, very large genomic DNA sequences can be hard to align using quadratic-space dynamic programming because they require too much memory.

The program EST_GENOME addresses this problem. It allows large introns, can recognize splice sites and uses limited memory. This combination of features makes a powerful and useful tool. EST_GENOME is used routinely at the Sanger Centre to help annotate human genomic sequence. As it is slow compared with search methods like BLAST (Altschul *et al.*, 1990), we first screen genomic DNA against dbEST using BLASTN. Any matching ESTs are realigned using EST_GENOME.

The algorithm uses a modification of Smith and Waterman (1981). The penalty structure used to score an alignment is as follows (defaults are in parentheses). Aligned bases score $+match$ (1) or cost $-mismatch$ (1) as appropriate. An indel in either sequence outside of an intron costs $-gap$ (2) (there is no gap initiation cost), and an intron (gap of arbitrary length in the genomic sequence only) costs $-intron$ (40), unless it starts with GT and ends with AG (or CT and AC if the splicing direction is reversed) when it costs $-splice$ (20). Thus, a gap of length L costs $L.gap$ in the spliced sequence and either $\min\{L.gap, intron\}$ or $\min\{L.gap, splice\}$ in the genome.

The numerical difference between *intron* and *splice* allows some slack in marking intron end-points. Sometimes the choice of boundaries which minimize indel and mismatch costs does not coincide exactly with the splice consensus, but, provided $intron - splice$ exceeds the extra mismatch/indel costs incurred, the alignment will respect the proper boundaries. If the alignment's introns still do not start/end with GT/AG (or CT/AC), then this may indicate errors in the sequences. The default parameters generally work well except that exons shorter than *splice* may be skipped. Intron penalties should always be greater than the longest expected random match (typically 10–15 bp) to avoid spurious matches.

The details of the algorithm are as follows. Let $X(i,j)$ be the score of the best local similarity ending at base i in the spliced sequence and j in the genomic sequence. Let $B(i)$ be the score of the best local alignment found so far that ends at i in the spliced sequence. Let $C(i)$ be the genome coordinate to which $B(i)$ refers. Let $S(i)$ and $G(j)$ be the nucleotides at positions i in the spliced and j in the genomic sequences, respectively. Then we have:

$$X(i,j) \leftarrow \max\begin{cases} X(i-1,j) & -gap \\ X(i-1,j-1) & +D \\ X(i,j-1) & -gap \\ B & \\ 0 & \end{cases}$$

$$D \leftarrow \begin{cases} match & \text{if } S(i) = G(j) \\ -mismatch & \text{otherwise} \end{cases}$$

$$B \leftarrow \begin{cases} B(i) - splice & \text{if } C(i),\ j \text{ are a donor–acceptor pair} \\ B(i) - intron & \text{otherwise} \end{cases}$$

$$(B(i), C(i)) \leftarrow \begin{cases} (X(i,j), j) & \text{if } X(i,j) > B(i) \\ (B(i), C(i)) & \text{otherwise} \end{cases}$$

Informatics Group, Sanger Centre, Welcome Trust Genome Campus, Hinxton Hall, Cambridge CB10 1SA, UK
Present address: Smith Kline Beecham Research & Development, Bioinformatics, New Frontiers Science Park (North), Third Avenue, Harlow, Essex CM19 5AW, UK. E-mail: richard_mott-1@sbphrd.com

```
     CNFG9  25669 ATCAGCGCTGCGGCCGCCCGGAAGCTCATCTTGGCCACCGACTCTCGCTTGCGCCGCCGCGGGAGCCGGTGGAAACCTGAGCGGGAGCTG  25758
                  || |||  |||| ||||||||||||||||||||||||||||||||||  ||.| ||||| ||||||| ||||||||||||||||||  ||
yo13c02.s1     20 ATAAGC-TTGCGACCGCCCGGAAGCTCATCTTGGCCACCGACTCTCG-ATG-GTCGCCG-GGGAGCC-GTGGAAACCTGAGCGGGACGTG    104

     CNFG9  25759 GAGAAGGAGCAGAGGGAGGCAGCACCCGGCGTGACGGGAGTGTGTGGGGCACTCAGGCCTTCCGCAGTGTCATCTGCCACACGGAAGGCA  25848
                  ||||||||||||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||
yo13c02.s1    105 GAGAAGGAGCAGAGGGAGGCAGCACCCGGCGTGAC-GGAGTGTGTGGGGCACTCAGGCCTTCCGCAGTGTCATCTGCCACACGGAAGGCA    193

     CNFG9  25849 CGGCCACGGGC-CAGGGGGTCTATGATctgga.....catacCTTCTGCATGCCCAGCTGGCATGGCCCCACGTA-GAGTGGGGGTGGCG  26325
                  |||||||||||   |||||||||||||<<<<< 404 <<<<<||||||||||||||||||||||||||||||||| |||||||| |||||
yo13c02.s1    194 CGGCCACGGGCAGGGGGGGTCTATGAT...............CTTCTGCATGCCCAGCTGGCATGGCCCCACGTAGGAGTGGGGTTGGCG    268

     CNFG9  26326 TCTCGGTGCTGGTCAGCGACACGTTGTCCTGGCTGGGCAGGTCCAGCTCCCGGAGGACCTGGGGCTTCAGCTTCCCGTAGCGCTGGCTGC  26415
                  ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
yo13c02.s1    269 TCTCGGTGCTGGTCAGCGACACGTTGTCCTGGCTGGGCAGGTCCAGCTCCCGGAGGACCTGGGGCTTCAGCTTCCCGTAGCGCTGGCTGC    358

     CNFG9  26416 AGTGACGGATGCTCTTGCGCTGCCATTTCTGGGTGCTGTCACTGTCCTTGCTCACTCCAAACCAGTCGGCGGTCCCCctggc.....ggt  26492
                  |||||||||||||||||||||||||||||||||||||||||||||||||||||| ||||||||||||||||||||||<<<<< 898 <<<
yo13c02.s1    359 AGTGACGGATGCTCTTGCGCTGCCATTTCTGGGTGCTGTCACTGTCCTTGCTCATTCCAAACCAGTCGGCGGTCCCC.............    435

     CNFG9  26492 acCTGCGGATGGTCTGTGTGATGGACGTCTGGCGTTGCAGCACCGGCCGCCGGAGCTCATGGTGGGGTGAAGAGATGTGGGCTGTCTCG  27477
                  <<|||||||||| ||||||||||||||| |||  ||||||||||||||||| |||  |||||| |||||||||||| |||| | |||||
yo13c02.s1    435 ..CTGCGGATGGTTTGTGTGATGGACGT-TGGGCTTGCAGCACCGGCCGCC-GAGTCCATGGTNGGGTGAAGAGAT-TGGG-TTTCTCG    518
```

Fig. 1. Example alignment produced by EST_GENOME. Introns are indicated by <<<<. The direction of gene splicing is reversed.

The *B* term is the cost of the best local alignment ending with an intron at *i,j*, so *X*(*i,j*) is the cost of the best overall alignment ending at *i,j*. If the alignment path as well as the score is required, then each time *B*(*i*) changes the previous pair (*B*(*i*), *C*(*i*)) is pushed onto a stack in case it is required during backtracking.

The program uses a linear-space divide-and-conquer strategy (Myers and Miller, 1988; Huang, 1994) to limit memory use:

1. A first-pass Smith–Waterman scan is done to find the start and end of the maximal-scoring segments. Subsequences corresponding to these segments are extracted.
2. If the product of the subsequences' lengths is less than a user-defined threshold, the segments are realigned using the Needleman–Wunsch algorithm, which will give the same result as the Smith–Waterman since they are guaranteed to align end to end.
3. If the product exceeds the threshold, the alignment is made recursively by splitting the spliced sequence in half and finding the genome position which aligns with the mid-point. This process is repeated until the product of lengths is less than the threshold. The divided sequences are aligned separately and then merged.
4. The genome sequence is searched against forward and reverse strands of the spliced sequence, assuming a forward gene splicing direction (i.e. GT/AG consensus). Then the best-scoring orientation is realigned assuming reverse splicing (CT/AC consensus). The overall best alignment is reported.

EST_GENOME displays its results both as an alignment and as a list of matching segments like those produced by MSPcrunch (Sonnhammer and Durbin, 1994). The latter format is easy to parse into other software. Figure 1 shows the alignment EST_GENOME made between the 519 bp EST yo13c02.s1 (Hillier *et al.*, 1996) and the cosmid cNFG9 (33 760 bp) from human chromosome 16 (Higgs, 1997). The program (i.e. all three comparisons) took 11 CPU s on a Ditigal Alpha 255/233. The alignment contains two introns, of lengths 404 and 898 bp, indicated by the <<< symbols. The output truncates the intron sequences. Note that both introns have CT/AG boundaries indicating that the direction of splicing is reversed.

I would like to thank Gos Micklem and Steve Jones for discussing this problem with me. This work was supported by a grant from the Welcome Trust.

References

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Birney,E., Thompson,J.D. and Gibson,T.J. (1996) Pair-wise and searchwise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucleic Acids Res.*, **24**, 2730–2739.

Gelfand,M.S., Mironov,A.A. and Pevzner,P.A. (1996) Gene recognition via spliced sequence alignment. *Proc. Natl. Acad. Sci. USA*, **93**, 9061–9066.

Hebsgaard,S.M., Korning,P.G., Tolstrup,N., Engelbrecht,J., Rouze,P. and Brunak,S. (1996) Splice site prediction in *Arabidopsis thaliana* pre-MRNA by combining local and global sequence information. *Nucleic Acids Res.*, **24**, 3439–3452.

Higgs,D. (1997) *Nature Genet.*, in press.

Hilier,L. *et al.* (1996) Generation and analysis of 280000 human expressed sequence tags. *Genome Res.*, **6**, 807–828.

Huang,X. (1994) On global sequence alignment. *Comput. Applic. Biosci.*, **10**, 227–235.

Huang,X. and Zhang,J. (1996) Methods for comparing a DNA sequence with a protein sequence. *Comput. Applic. Biosci.*, **12**, 497–506.

Myers,E.W. and Miller,W. (1988) Optimal alignments in linear space. *Comput. Applic. Biosci.*, **4**, 11–17.

Smith,T.E. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Solovyev, V.V., Salamov, A.A. and Lawrence,C.B. (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.*, **22**, 5156–5163.

Sonnhammer,E.L.L. and Durbin,R. (1994) A workbench for large scale sequence homology analysis. *Comput. Applic. Biosci.*, **10**, 301–307.

Received on January 1, 1997; revised on February 17, 1997; accepted on February 24, 1997