Computational Discovery of Internal Micro-Exons

Natalia Volfovsky, Brian J. Haas and Steven L. Salzberg

Genome Res. 2003 13: 1216-1221
Access the most recent version at doi:10.1101/gr.677503

Computational Discovery of Internal Micro-Exons

Natalia Volfovsky,[1,2,3] Brian J. Haas,[1] and Steven L. Salzberg[1]

[1] *The Institute for Genomic Research, Rockville, Maryland 20850, USA*

Very short exons, also known as micro-exons, occur in large numbers in some eukaryotic genomes. Existing annotation tools have a limited ability to recognize these short sequences, which range in length up to 25 bp. Here, we describe a computational method for the identification of micro-exons using near-perfect alignments between cDNA and genomic DNA sequences. Using this method, we detected 319 micro-exons in 4 complete genomes, of which 224 were previously unknown, human (170), the nematode *Caenorhabditis elegans* (4), the fruit fly *Drosophila melanogaster* (14), and the mustard plant *Arabidopsis thaliana* (36). Comparison of our computational method with popular cDNA alignment programs shows that the new algorithm is both efficient and accurate. The algorithm also aids in the discovery of micro-exon-skipping events and cross-species micro-exon conservation.

[Supplementary material available online at www.genome.org.]

With the rapid increase in the generation of eukaryotic genome sequence data, the development of accurate, detailed computational methods for gene structure prediction and verification has become increasingly important. Many computational methods have addressed the difficult problem of exon-intron structure annotation (Florea et al. 1998; Usuka et al. 2000; Kan et al. 2001; Wheelan et al. 2001; Kent 2002). One area that remains particularly difficult is the identification of very short exons, both internal and terminal (Florea et al. 1998; Black 2000). These micro-exons often confound both alignment and ab initio gene prediction programs, because they contain virtually no meaningful statistical signal. Micro-exons with lengths up to 25 bp have been studied experimentally in plants, insects, and vertebrate animals (Reyes et al. 1991; McAlister et al. 1992; Sterner and Berget 1993; Chan and Black 1995; Simpson et al. 2000). These experimental studies support two important features of micro-exons; (1) they sometimes facilitate alternative splicing events, and (2) despite their small size, they are usually conserved between species (Berget 1995; Black 2000). The inclusion of micro-exons is dependent on stages of cell development and cell specificity. It is mediated by several factors, including intronic and exonic splicing enhancer and silencer elements, and SR and hnRNP proteins (Berget 1995; Black 2000; Simpson et al 2000). Experimental and computational studies of micro-exons are recognized as important challenges in the understanding of splicing machinery and genome variability.

This study describes a computational method for finding internal micro-exons in near-perfect alignments of cDNA and genomic DNA sequences. The method provides quick recognition of micro-exons that might have been missed by an alignment program (called the short exon error problem by Florea et al. [1998]). We demonstrate the method on four complete genomes, *C. elegans* (The *C. elegans* Sequencing Consortium 1998), *D. melanogaster* (Adams et al. 2000), *A. thaliana* (The *Arabidopsis* Genome Initiative 2000) and human (Venter et al. 2001; Lander et al. 2001). In each case, significant numbers of micro-exons were identified that were undetected in previous annotation efforts. We discovered previously unreported micro-exons, alternative splicing events, and highly conserved micro-exons across several species.

[2]Present address: Advanced Biomedical Computing Center, National Cancer Institute-Frederick/SAIC, Frederick, Maryland 21702, USA.
[3]Corresponding author.
E-MAIL natalia@ncifcrf.gov; FAX (301) 838-0208.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.677503.

RESULTS AND DISCUSSION

Identification of Micro-Exons

The spliced alignment correction procedure was applied to four eukaryotic genomes, the model plant *Arabidopsis thaliana*, the nematode worm *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster*, and human. The data sets were checked manually and the following questions were addressed: (1) How many micro-exons from these sets are annotated in existing databases? (2) Are there alternative splicing events? (3) Are micro-exons found in these species conserved in related species?

Table 1 summarizes the micro-exons found by our algorithm in each species, including previously annotated as well as newly discovered micro-exons. Table 2 shows the distribution of lengths of all micro-exons found by our method and by Sim4 alone. Supplemental Tables 1–4 contain complete lists of all the micro-exon sequences found in each species. This data substantially expands our recent report on micro-exons in *A. thaliana*, which emerged during the course of aligning a set of 5000 full-length cDNA sequences to the genome (Haas et al. 2002). Here, we applied this method to all cDNA sequences from GenBank (as of March 2002) for *A. thaliana*, plus three additional species.

Comparison With Existing Alignment Programs

We ran a set of four programs including GeneSeqer (Usuka et al. 2000), Gap2 (Huang et al. 1997), Spidey (Wheelan et al. 2001), and BLAT (Kent 2002) on cDNA sequences containing the micro-exons discovered using the spliced alignment-correction procedure. To minimize memory requirements and increase alignment efficiency, the *Arabidopsis* and worm cDNA alignments were limited to a 100-kbp segment of genomic sequence containing the gene of interest; the human and *Drosophila* cDNAs were searched against a 2-Mbp se-

13:1216–1221 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

Table 1. Micro-Exons Found Using the Spliced Alignment Correction Procedure

Genome	Number ofcDNAs	Micro-exons	Previously annotated micro-exons	Exon skipping
C. elegans	1,356	6	2 (WormBase, March 2002)	3
D. melanogaster	8,214	24	10 (Celera and FlyBase, March 2002)	3
A. thaliana	16,956	66	30 (TIGR, November 2001)	1
Human	59,413	223	53 (Celera, 2001) 67 (Ensembl, March 2002)	53

Of the 319 micro-exons found using the spliced alignment correction procedure, 224 were previously unknown.

quence, because some genes span >100 kbp in the genome (Table 3). For the programs GeneSeqer and gap2, which report all high-scoring alignments, only the single highest scoring alignment was examined, with the score calculated as the sum of the products of the segment lengths and the percent identity for each alignment segment in a single alignment chain. We found that Sim4's performance improved when shorter segments of genomic DNA were used for alignment; we included these results in Table 3.

We previously (Haas et al. 2002) compared the performance and sensitivity of several spliced alignment algorithms on a set of 5000 *Arabidopsis* cDNA sequences. The GeneSeqer program (Usuka et al. 2000) was most accurate at identifying micro-exons in *A. thaliana*. However, it is organism specific, requiring training on validated splice sites for each genome. Consistent with previous results, GeneSeqer performs very well in the identification of micro-exons in *Arabidopsis* cDNA alignments. With the exception of GeneSeqer on *Arabidopsis*, no single program was found capable of identifying all micro-exons for any species. We found micro-exons in both *Drosophila* and human that were not identified by any of the programs evaluated, supporting the need for the correction algorithm described here.

Annotation and Alternative Splicing of Micro-Exons

Many of the micro-exons found in this study (Table 1) were not yet annotated in public genome databases. Exon-skipping events were recently studied computationally in human chromosome 22 (Hide et al. 2001). In that study, special tag sequences for all consecutive and nonconsecutive exon–exon junctions were created and then searched against a comprehensive human EST database. Our method differs in that it uses corrected micro-exon boundaries to check for exon skipping, and it searches against full-length cDNA sequences rather than ESTs. This method identified 68 putative exon-skipping events (Tables 1 and 2).

Alternative splicing of micro-exons was decribed previously in vertebrates (Reyes et al. 1991; McAllister et al. 1992; Inman et al. 1998). The only case of an alternative splicing event in plants involving a micro-exon was reported previously in the invertase gene (Bournary et al. 1996). Within the set of newly discovered micro-exons in *Arabidopsis*, we found one new case of exon skipping, involving a 2-bp micro-exon in the 5′ UTR of accession AF410302.1 (Fig. 1A,B). The homologous cDNA sequence, Ceres17241 (Haas et al. 2002), has 97% identity with AF410302.1, but does not contain the micro-exon. In addition, the near-perfect alignment between these two sequences reveals a mutation within an upstream exon (Fig. 1C). This corroborates findings regarding the troponin gene (Sterner and Berget 1993), in which it was shown that mutations in an upstream exon alter the recognition of the micro-exon.

Our method assumes that introns are flanked by the consensus dinucleotides GT and AG. This choice is supported by data indicating that GT-AG pairs represent >99% of introns (Mount 1982; Burset et al. 2000). However, noncanonical splice sites can be added easily to the recognition procedure. When we searched for noncanonical AT-AC introns in *Arabidopsis*, we found one additional micro-exon (AF255713.1). We also assume that the cDNA sequences represent fully processed mRNAs; it is possible that some of the micro-exons could represent intermediate products that do not appear in the final, mature transcript.

Comparative Analysis

We next looked for cross-species conservation of micro-exons. Target sequences were constructed and searched against cDNA sequences from rice (1302 sequences), potato (464), maize (1193), rat (8621), and mouse (38,263), all obtained from GenBank. In plants, we confirmed the conservation of a 9-bp exon in the invertase gene, which was previously established experimentally (Simpson et al. 2000). We found identical micro-exons in *Arabidopsis*, rice, potato, and maize invertase cDNAs. This exon encodes three amino-acids (DPN), which might have an important function in enzyme conformation or catalytic activity. We also found a conserved 5-bp micro-exon in the D-ribulose-5-phosphate-3-epimerase in both *Arabidopsis* and rice.

In animal species, we found a conserved 9-bp micro-exon in the *unc-13* gene of *D. melanogaster* and *C. elegans*,

Table 2. Distribution of Micro-Exon Lengths

Genome	2–5 bp	6–10 bp	11–15 bp	16–20 bp	21–25 bp	All micro-exons 2–25 bp	Exon skipping
C. elegans	0	2	1	3	4	10	3
D. melanogaster	1	10	6	5	12	34	4
A. thaliana	7	19	13	28	74	141	1
Human	5	59	77	90	194	425	60

A total of 610 internal micro-exons were identified using the spliced alignment correction procedure and with sim4 alone.

Table 3. Micro-Exon Finding by Popular cDNA Alignment Programs

Program	*C. elegans*	*D. melanogaster*	*A. thaliana*	Human
Total micro-exons	6	24	66	223
GeneSeqer	5	14	66	75
dds/gap2	5	16	51	165
Spidey	4	8	32	101
Sim4	2	1	29	25
BLAT	5	20	59	172
All programs failed	0	2	0	22

The alignment program comparison was restricted to the 319 micro-exons identified using the corrected spliced alignment procedure. For gap2 and GeneSeqer, only the top scoring alignment was examined; micro-exons identified within lower scoring alignments were not described here.

but this particular micro-exon does not seem to be conserved in vertebrates. This exon encodes a VLK tripeptide upstream of the C1 domain reported recently in *C. elegans* (Kohn et al. 2000), but it has not been reported previously or annotated in *Drosophila* (Xu et al. 1998). A large set of conserved micro-exons were found in the mammalian cDNA collection, 64 sequences conserved between mouse and human, 37 between human and rat, and 20 conserved between all three species.

Statistics of Micro-Exons

Table 2 summarizes the total of 610 micro-exons found in the 4 species in this study. Of these, 319 were found by our correction procedure, whereas the rest were found by the Sim4 alignment program alone. The cDNA sequences corresponding to these exons have nearly 100% identity to genomic DNA, and the introns have consensus splice site boundaries (Fig. 2A). Each of the micro-exons identified in this study corresponds to a specific cDNA accession. Due to the biases and redundancy in the cDNA data set, the number of micro-exons does not reflect the number of genes containing micro-exons. To investigate the frequency of micro-exons in genes supported by complete cDNA sequences, the cDNA data set was collapsed into clusters using an all-vs-all BLASTN search, followed by single-linkage clustering of match pairs containing at least 98% identity across at least 50 % of the cDNA length (Table 4). Clusters of cDNA sequences containing micro-exons represent only 1% of all clusters of cDNA sequences in *C. elegans*, 0.5% in *D. melanogaster*, 1.3% in *A. thaliana*, and 1.6% in human, indicating that micro-exons are quite rare in the population of genes supported by cDNA sequences.

Because they are so short, micro-exons might have a stronger tendency than other exons to contain exon-splicing enhancers (ESEs). We checked for some of the most common enhancers – [A/G]AAGAA, TGAAGA, CAACAA – and found 33 instances of micro-exons with these sequences. A larger study would be required to determine whether this represents a statistically significant tendency.

One other interesting feature of the micro-exons are their length distributions, shown in Table 2. The preferred length of micro-exons is a multiple of three; this describes 90% of the micro-exons in *C. elegans*, 85% of the micro-exons in *D. melanogaster*, 64% of the micro-exons in human, and 34% in *A. thaliana*. Obviously, skipping (or including) a micro-exon whose length is a multiple of three will not change the reading frame. The excess of such exons in animal species suggests that the skipping of micro-exons may be a commonly used mechanism of generating alternative gene products in animals.

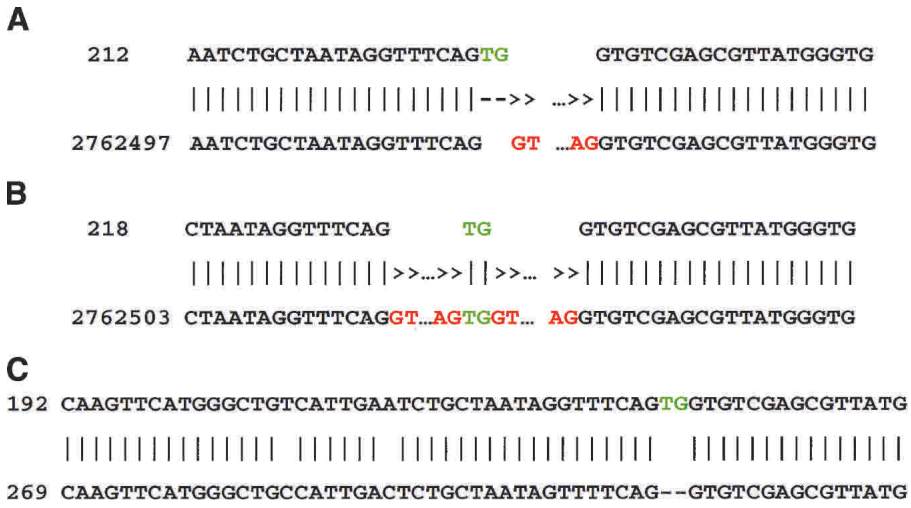

Figure 1 Example of alternative splicing of micro-exon in *A. thaliana*. (*A*) Fragment of Sim4 alignment of AF410302.1 sequence (*top*) with genomic DNA (*bottom*). Sim4 incorrectly aligns the TG-dinucleotide micro-exon against gaps to yield a GT donor splice site. (*B*) Fragment of alignment after the correction procedure in which the TG-dinucleotide micro-exon is positioned adjacent to the consensus splice sites. (*C*) Fragment of alignment between AF410302.1 (*top*) and Ceres_ 17241 (*bottom*) cDNA sequences. The micro-exon is colored green, whereas the dinucleotide consensus splice sites are colored red.

METHODS

The micro-exon search procedure consists of three steps, mapping, recognition, and correction. We will detail these steps below.

Mapping cDNA Sequences to a Genomic Scaffold

The purpose of this step is to reduce the amount of sequences for searching in the following step. All cDNA sequences are aligned against genomic scaffolds and chromosomes using MUMmer (Delcher et al. 1999, 2002) with a minimum match length of 60 bp. This provides an extremely rapid way to

Figure 2 Definition of short exon error. (*A*) Correct exon boundaries. (*B*) Suspicious right exon boundary lacks the consensus AG acceptor splice site flanked by a short region of identically aligned nucleotides. (*C*) Both exons lack suitable boundaries.

find potential cDNA alignments to any size genome. Results are filtered and cDNA sequences are retained only if they have more than two matches to a genome or chromosome. These sequences are then used as the subject for the following, more detailed alignments.

Recognition of Micro-Exons

In spliced alignments of cDNA to genomic DNA, the nucleotides from small exons are sometimes aligned incorrectly to introns or to neighboring exons (Florea et al. 1998). To find micro-exons, these mismatches must be recognized, and then the micro-exons need to be localized correctly to the genomic DNA.

We define a near-perfect alignment as one in which most exons have 100% identity with the genomic DNA, and the remainder have at least 90% identity. All cDNA alignments examined were at least 90% identical with the genomic sequence. In such alignments, a correct exon boundary can be described as a short sequence at the end of an exon in which the cDNA and genomic DNA match, and in which the genomic DNA contains a consensus splice site adjacent to the aligned exon (Fig. 2A). In the figure, short sequences (5 bp) at the ends of exons are shown in the same colors in the cDNA and the aligned genomic DNA. If one or both exons flanking an intron are incorrect—due to mismatches at the ends of the exons or to a non-consensus splice site—the algorithm reports this as a potential case of short exon error (Fig. 2B,C).

Correction Procedure

To correct the alignment, we look for potential acceptor and donor splice sites in the genomic DNA within a short distance (30 bp) of each alignment boundary. Next, all combinations of these splice sites are used to construct potential micro-exons within the cDNA sequence, as shown in Figure 3. Each potential micro-exon is then used to search the genomic DNA annotated previously as an intron. If a match to the genomic DNA is found, and if the match produces a new alignment with 100% identity to the cDNA and with consensus splice sites, it is reported as a micro-exon.

Implementation, Performance, and Parameters of the Method

The core algorithm is implemented in a Perl script that calls both the MUMmer (Delcher et al. 2002) and Sim4 (Florea et al. 1998) programs, both of which are freely available. The inputs to the micro-exon program are a set of cDNA sequences plus a genome sequence (either a single sequence or multiple chromosomes). The output is a list of micro-exons with the identifiers of the cDNA, the genome sequence, and the micro-exon locations. The current implementation reports micro-exons of 3–25 bp in length. Shorter exons (1–2 bp) may be found, but these require manual curation. The program parameters include the maximal length of micro-exons, average percent identity, and the length of the short sequences adjacent to splice sites. In all of our applications, we used only full-length cDNA sequences and near-identical alignments.

The system takes <10 sec per cDNA to identify micro-exons. The method is easily parallelized, and we applied a parallel version of the method that uses the Condor system (University of Wisconsin-Madison) to the data in this study. It searches all human genome scaffolds longer than 10 kbp in ~48 h, using a cluster of 10 HP/Compaq Alpha computers.

Table 4. Clusters of cDNA Sequences and Distribution of Micro-Exons

Genome	Number of cDNAs	Number of cDNA clusters	Distribution of cluster sizes (cDNAs/cluster) with numbers of micro-exons			
			1	2–5	6–20	>20
C. elegans	1,356	1,105	934 (5)	165 (5)	6 (0)	—
D. melanogaster	8,214	6,484	5,232 (20)	1,220 (13)	31 (1)	1 (0)
A. thaliana	16,956	10,768	6,729 (82)	3,976 (58)	62 (1)	1 (0)
Human	59,413	26,391	15,212 (111)	9,753 (236)	1,359 (72)	67 (6)

The columns labeled Distribution of cluster sizes show, in each table entry, the number of cDNA clusters in the given size range, and in parentheses the number of micro-exons found within that set of clusters.

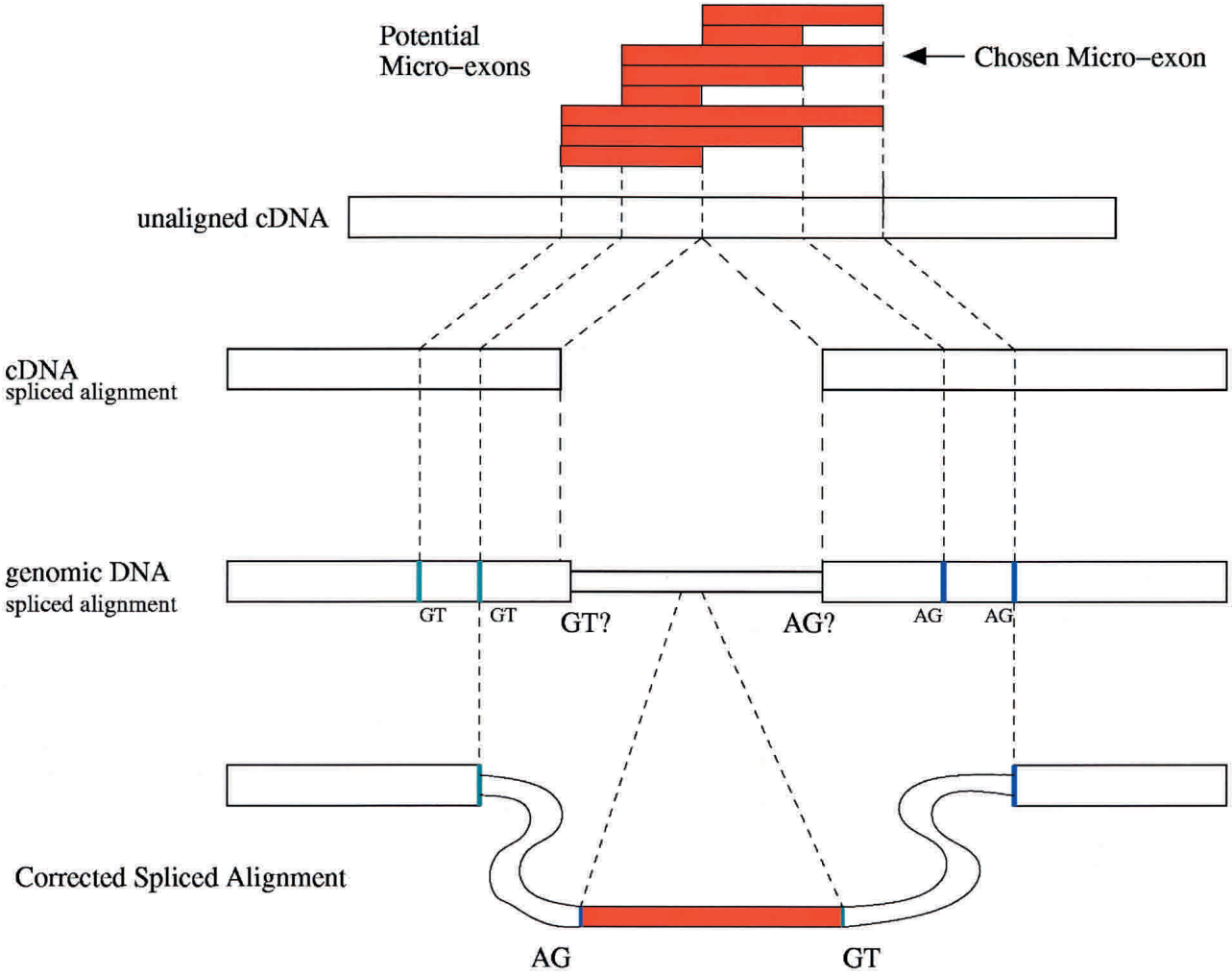

Figure 3 Spliced alignment correction algorithm. Combinations of consensus donor and acceptor splice sites adjacent to the flawed alignment boundaries are used to enumerate all potential micro-exons within the cDNA sequence. One of the potential micro-exons is localized to the genomic DNA intervening the alignment segments, flanked by consensus splice sites, yielding the identity and position of the newly discovered micro-exon.

Testing for Exon Skipping and Cross-Species Conservation

Using the micro-exons discovered in the four species in this study, we searched the original cDNA sequences for exon-skipping events. BLAST (Gish and States 1993) searches revealed that cDNAs with high similarity to the micro-exon-containing query sequences also harbored micro-exons (E-value $<10^{-50}$). For each micro-exon, we created alternative sequences that consisted of the fragments from its flanking exons (10 bp from the ends of the adjacent exons). These 20-bp sequences were searched against the sets of previously selected cDNAs using Sim4. Cases in which at least 18/20 bp aligned with the corresponding cDNAs are reported below as alternative splice variants. These results were confirmed by manual inspection.

A similar procedure was used in the search for exons conserved between related species. After selecting the most similar cDNAs from the other species, we constructed another target sequence, containing both a micro-exon sequence and the ends of flanking exons. These sequences were searched against the sets of cDNAs from related species using Sim4. Matches with >95% identity results are reported as cases of conserved exons after manual inspection.

ACKNOWLEDGMENTS

We thank Razvan Sultana and Svetlana Karamysheva for computational and database support. This work was supported in part by NSF under grant KDI-9980088 and by NIH under grant R01-LM06845.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Berget, S.M. 1995. Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270:** 2411–2414.
Black, D.L. 2000. Protein diversity from alternative splicing: A challenge for bioinformatics and post-genome biology. *Cell* **103** 367–370.
Bournay, A.S., Hedley, P.E., Maddison, A., Waugh, R., and Machray, G.C. 1996. Exon skipping induced by cold stress in a potato invertase gene transcript. *Nucleic Acids Res.* **24:** 2347–2351.
Burset, M., Seledtsov, I.A., and Solovyev, V.V. 2000. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* **28:** 4364–4375.
The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. The *C. elegans* Sequencing Consortium. *Science* **282:** 2012–2018.
Chan, R.C. and Black, D.L. 1995. Conserved intron elements repress splicing of a neuron-specific c-src exon in vitro. *Mol. Cell. Biol.* **15:** 6377–6385.
Delcher, A.L., Kasif, S., Fleischmann, R.D., Peterson, J., White, O., and Salzberg, S.L. 1999. Alignment of whole genomes. *Nucleic Acids Res.* **27:** 2369–2376.
Delcher, A.L., Phillippy, A., Carlton, J., and Salzberg, S.L. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30:** 2478–2483.
Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8:** 967–974.
Gish, W. and States, D.J. 1993. Identification of protein coding regions by database similarity search. *Nat. Genet.* **3:** 266–272. (http://blast.wustl.edu/blast).
Haas, B.J., Volfovsky, N., Town, C.D., Troukhan, M., Alexandrov, N., Feldmann, K.A., Flavell, R.B., White, O., and Salzberg, S.L. 2002. Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biol.* **3:** research0029.1–0029.12.
Hide, W.A., Babenko, V.N., van Heusden, P.A., Seoighe, C., and Kelso, J.F. 2001. The contribution of exon-skipping events on chromosome 22 to protein coding diversity. *Genome Res.* **11:** 1848–1853.
Huang, X., Adams, M.D., Zhou, H., and Kerlavage, A.R. 1997. A tool for analyzing and annotating genomic sequences. *Genomics* **46:** 37–45.
Inman, M.V., Levy, S., Mock, B.A., and Owens, G.C. 1998. Gene organization and chromosome location of the neural-specific RNA binding protein Elavl4. *Gene* **208:** 139–145.
Kan, Z., Rouchka, E.C., Gish, W.R., and States, D.J. 2001. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.* **11:** 889–900.
Kent, W.J. 2002. BLAT—The BLAST-like alignment tool. *Genome Res.* **12:** 656–664.
Kohn, R.E., Duerr, J.S., McManus, J.R., Duke, A., Rakow, T.L., Maruyama, H., Moulder, G., Maruyama, I.N., Barstead, R.J., and Rand, J.B. 2000. Expression of multiple UNC-13 proteins in the *Caenorhabditis elegans* nervous system. *Mol. Biol. Cell* **11:** 3441–3452.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
McAllister, L., Rehm, E.J., Goodman, G.S., and Zinn, K. 1992. Alternative splicing of micro-exons creates multiple forms of the insect cell adhesion molecule fasciclin I. *J. Neurosci.* **12:** 895–905.
Mount, S.M. 1982. A catalogue of splice junction sequences. *Nucleic Acids Res.* **10:** 459–472.
Reyes, A.A., Small, S.J., and Akeson, R. 1991. At least 27 alternatively spliced forms of the neural cell adhesion molecule mRNA are expressed during rat heart development. *Mol. Cell. Biol.* **11:** 1654–1661.
Simpson, C.G., Hedley, P.E., Watters, J.A., Clark, G.P., McQuade, C., Machray, G.C., and Brown, J.W. 2000. Requirements for mini-exon inclusion in potato invertase mRNAs provides evidence for exon-scanning interactions in plants. *RNA* **6:** 422–433.
Sterner, D.A. and Berget, S.M. 1993. In vivo recognition of a vertebrate mini-exon as an exon-intron-exon unit. *Mol. Cell. Biol.* **13:** 2677–2687.
Usuka, J., Zhu, W., and Brendel V. 2000. Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics* **16:** 203–211.
Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.
Wheelan, S.J., Church, D.M., and Ostell, J.M. 2001. Spidey: A tool for mRNA-to-genomic alignments. *Genome Res.* **11:** 1952–1957.
Xu, X.Z., Wes, P.D., Chen, H., Li, H.S., Yu, M., Morgan, S., Liu, Y., and Montell, C. 1998. Retinal targets for calmodulin include proteins implicated in synaptic transmission. *J. Biol. Chem.* **273:** 31297–31307.

WEB SITE REFERENCES

http://www.celeradiscoverysystem.com/; Celera.
http://www.ensembl.org/; Ensembl.
http://www.fruitfly.org/; FlyBase.
http://www.ncbi.nlm.nih.gov/Genbank/; GenBank.
http://www.tigr.org/; TIGR.
http://www.wormbase.org/; WormBase.

Received August 14, 2002; accepted in revised form March 25, 2003.