*3192–3202* *Nucleic Acids Research, 2007, Vol. 35, No. 10* *Published online 22 April 2007*
*doi:10.1093/nar/gkm187*

# Non-EST-based prediction of novel alternatively spliced cassette exons with cell signaling function in *Caenorhabditis elegans* and human

**German Gaston Leparc and Robi David Mitra***

Department of Genetics and Center for Genome Sciences, Washington University in St Louis, 4444 Forest Park Parkway, Campus Box 8510, St Louis, MO 63108, USA

Received January 3, 2007; Revised March 15, 2007; Accepted March 16, 2007

## ABSTRACT

**To better understand the complex role that alternative splicing plays in intracellular signaling, it is important to catalog the numerous splice variants involved in signal transduction. Therefore, we developed PASE (Prediction of Alternative Signaling Exons), a computational tool to identify novel alternative cassette exons that code for kinase phosphorylation or signaling protein-binding sites. We first applied PASE to the *Caenorhabditis elegans* genome. In this organism, our algorithm had an overall specificity of ≥76.4%, including 33 novel cassette exons that we experimentally verified. We then used PASE to analyze the human genome and made 804 predictions, of which 308 were found as alternative exons in the transcript database. We experimentally tested 384 of the remaining unobserved predictions and discovered 26 novel human exons for a total specificity of ≥41.5% in human. By using a test set of known alternatively spliced signaling exons, we determined that the sensitivity of PASE is ∼70%. GO term analysis revealed that our exon predictions were found in the introns of known signal transduction genes more often than expected by chance, indicating PASE enriches for splice variants that function in signaling pathways. Overall, PASE was able to uncover 59 novel alternative cassette exons in *C. elegans* and humans through a genome-wide *ab initio* prediction method that enriches for exons involved in signaling.**

## INTRODUCTION

One intriguing aspect of signal transduction is that there seems to be a relatively small number of signaling pathways, yet cells are able to generate multiple responses to different signals from the environment. One explanation is that alternative splicing of pre-mRNA generates a much larger number of unique signaling proteins than is generally appreciated, thus enabling a diverse set of responses. Several pieces of evidence support this hypothesis. Bioinformatics analysis and DNA microarray experiments indicate that 59–74% of all human genes are alternative spliced (1,2) and that roughly 75% of alternative splicing events alter the protein-coding region of the transcript (3–5). This suggests that alternative splicing has the potential to produce a large number of different proteins from the surprisingly limited number of genes in multicellular organisms. The connection between alternative splicing and intracellular signaling is further strengthened by studies of glucocorticoid signaling, where it has been shown that alternative splicing is the mechanism by which a single gene (the glucocorticoid receptor) is able to mediate a variety of different responses in different cell types (6). Similar results have been observed in other signaling pathways—both the γ-amino-butyric acid type A receptor gamma-subunit ($GABA_AR$ γ2) gene and the myosin light chain kinase (MLCK) gene contain alternative cassette exons that encode for phosphorylation sites that alter their signaling functions (7–9).

In order to accelerate our efforts to understand the role alternative splicing plays in intracellular signaling pathways, we need to identify and accurately catalog alternatively spliced (AS) isoforms involved in signal transduction. However, there are currently several difficulties to obtaining such a catalog. While most known alternative-splicing events have been discovered through large-scale EST sequencing, this approach has some drawbacks. First, in many model organisms, where intracellular signaling is most easily dissected, relatively few EST sequences have been collected (e.g. *Caenorhabditis elegans* ∼300 000 ESTs). Even in organisms with high EST coverage such as human, many AS isoforms go undetected because EST libraries are biased to highly expressed splice forms as well as to the

*To whom correspondence should be addressed. Tel: +1-314-362-2751; Fax: +1-314-362-2156; Email: rmitra@genetics.wustl.edu

3′ or 5′ ends of genes. Also, many minor splice variants are not captured by this conventional approach because of their specific expression in particular tissues or developmental stages. Indeed, many human tissues have not been adequately sampled—there are ~210 human cell types (10) and many of these have little or no EST coverage. Finally, even when AS transcripts are found, their functions are often unknown. As a result, our ability to detect alternative splice variants and determine their function is limited.

Recognition of these shortcomings has sparked considerable interest in developing computational approaches to splice variant prediction (11). Attempts at *ab initio* prediction of alternative splicing that is based on intronic sequence alone have proven difficult because of the high number of pseudo-splice sites in intronic sequences (12). In addition, many of the known AS cassette exons that affect signaling are quite small, and may be difficult to find by conventional gene prediction methods alone. Recently, progress has been made using species conservation as well as protein domain information (13–15), yet the total number of experimentally validated novel isoforms remains modest compared to the numbers found by traditional EST sequencing. Also, no hypothesis can be made about the function of these observed alternative splice variants.

To complement these approaches and to address the issue of finding signaling exons, we developed prediction of alternative signaling exons (PASE), a computational tool to identify AS cassette exons that are likely to be involved in intracellular signaling. This algorithm uses first-order Markov models and a Bayesian classifier to identify likely donors and acceptor sites in an intron. Using these potential splice sites, it identifies in-frame cassette exons that code for phosphorylation sites or signaling protein-binding sites. As an additional filter, only exons that are conserved across species are kept. Here, we report the results of a genome-wide application of our algorithm and demonstrate that our genome-wide *ab initio* approach enriches for novel exons likely to be involved in cell signaling.

## METHODS

### Preparation of intron, EST/mRNA and species conservation data

All intron data sets were processed from REFSEQ genomic alignment annotations, which were downloaded from the UCSC Genome Browser (May 2004 release). Redundant transcripts and intron-less genes were removed from both human and *C. elegans* data sets. Spliced-ESTs and phastCons 'Most conserved' species conservation blocks were downloaded as chromosome coordinate data from the UCSC Genome Browser (as of January 2, 2005). Two PERL programs (compare_to_expressedseq.pl and compare_to_conserved.pl) were written to compare the overlap of coordinates from these data sets to the predicted exon coordinate data.

### Acceptor and donor splice site scoring

The *C. elegans* and human cassette exon models consist of the pairing of acceptor and donor splice sites with an exon size restriction of 30–330 bp. Both acceptor and donor splice site models use 12-mer first-order Markov chains to capture the over-represented dinucleotide frequencies around the splice sites (34). These models were trained on a randomly sampled set of 5000 REFSEQ internal exons from human and *C. elegans* genome, respectively.

Calculation of the first-order Markov chain model of the splice site:

$$P(\text{splice site}) = P(x_1)P\prod_{i=2}^{12}\cdot P(x_i|x_{i-1}).$$

Log-odds ratio of the splice site versus the background probability $P_b$:

$$\log_2\left(\frac{P(\text{splice site})}{P_b}\right).$$

The background probabilities were calculated by counting the overlapping dinucleotide frequencies of all intronic sequences from the REFSEQ genes of each genome.

### Bayesian classification of an acceptor–donor pair exon model

A Bayesian classifier was trained and tested with sets of real and pseudo-exon acceptor–donor pair scores. Pseudo-exons in this case are any pair of AG–GT dinucleotides within range of the exon size limits taken from randomly generated intronic sequences based on species-specific dinucleotide distribution. For the human test set, a ratio of 1:10 real exons to pseudo-exons was used (1000 real exons versus 10 000 pseudo-exons), reflecting the order of magnitude number of pseudo-exons relative to real exons typically found in the human genome (12). The *C. elegans* test set had a 1:4 ratio (1000 real exons versus 4000 pseudo-exons) of real exons to pseudo-exons. Training of the Bayesian classifier parameters for both the real and pseudo-exon acceptor–donor pair score distributions were calculated by using multivariate Gaussian distributions with the prior probabilities based on the ratios of real exon to pseudo-exons.

The function for the Multivariate Gaussian distribution model for exon acceptor–donor pair scores is the following:

Let $x_i = \{\text{acceptor bit score, donor bit score}\}$, $\sigma = \text{covariance matrix of acceptor and donor bit score distributions and det } (\sigma) = \text{determinant of the covariance matrix}$

$$f = \frac{1}{2\pi\sqrt{\det(\sigma)}}e^{2(x_i-\bar{x})\sigma^{-1}(x_i-\bar{x})^T}.$$

The calculation of the probability that an exon is real, given its acceptor–donor pair scores is the following:

Let $P(\text{R})$ and $P(\text{P})$ be the prior probability of Real Exons and Pseudo-Exons, respectively

$$P(\text{Real}|x_i) = \frac{f_{\text{real}} \times P(\text{R})}{[f_{\text{real}} \times P(\text{R}) + f_{\text{pseudo}} \times P(\text{P})]}.$$

### Exon translations

A PERL program (intron_phase.pl) was written to determine the reading frames of exons and the phase

of each intron as 0, 1 or 2 depending on where the codon was spliced. When predicting exons, we also translated the five flanking amino acids from both of the adjacent constitutive exons (See Supplementary Data).

### PASE-sensitivity test

A splicing events data file 'AltSplice-rel3.events.txt' and gene sequence file 'AltSplice-rel3.genes.txt.gz' were downloaded from Alternative Splicing Database (http://http://www.ebi.ac.uk/asd/) (24). Single cassette exon splicing events and their sequences were extracted using 'get_single_cassette.pl' PERL program. WU-BLAST was used for BLASTX comparison with the Phospho.ELM database (25) and only exons with 100% identity were selected. PASE was then applied to corresponding REFSEQ gene structures that lack the cassette exons.

### Creation of Scansite log-odds matrices

The current release of Scansite (version 2.0) includes 63 motifs characterizing the binding and/or substrate specificities of many families of Ser/Thr- or Tyr-kinases, SH2, SH3, PDZ, 14-3-3 and PTB domains, together with signature motifs for PtdIns(3,4,5)P3-specific PH domains (35). PDZ-binding motifs were excluded due to the COOH-terminal sequence requirement. We modified the Scansite matrices to be log-odds scoring matrices to take into account the background distribution of the proteomes for human and *C. elegans*. This modification allowed us to compare the information content of the matrices (Supplementary Table A) and to rank the scores of predicted functional exons independent from percentile ranking (See Supplementary Data).

Let $S_{ij}$ be the Scansite selectivity value of amino acid $i$ in position $j$:

$$P(A_{ij}|\mathrm{PSSM}) = \frac{S_{ij}}{\sum S_j}.$$

The calculation of the log-odds scoring matrix is the following:

$$M_{ij} = \log_2(P(A_{ij}|\mathrm{PSSM})/P(A_i|\mathrm{Background})).$$

Using these matrices, a PERL script was written (find_scansite.pl) to search the putative exons for signaling motifs using bit score thresholds of 10 and 6 bits.

### Primer design of selected candidates

Batch processing of the primer design for all candidate exons was done with a PERL program (prediction_to_primer3.pl) in combination with the PRIMER3 software (36). Primers were designed using the following PRIMER3 settings: primer length minimum, 19 nt, desired; 25 nt and maximum, 32 nt; melting temperature minimum, 64°C, desired length, 70°C and maximum length, 73°C; minimum GC content of 45, and maximum of 80; product length, 150–700 nt; and pre-filtering of potentially mispriming sequences with the provided library of human repeats. Figure 3 illustrates the primer design and the expected PCR products. Primer sequences were ordered from Integrated DNA Technologies, Coralville, IN, USA (list of primers available in Supplementary Data).

### Semi-nested RT-PCR experiments

Pooled total RNA samples from 18 different tissues types were used for semi-nested PCR validation in human (Supplementary Table D). Total RNA from whole worms were used for PCR validation in *C. elegans*. Superscript II Reverse Transcription (Invitrogen, Carlsbad, CA, USA) was used to create cDNA with candidate gene-specific reverse primers. All cDNA samples were tested for the presence of RNA Polymerase II transcript as a control. The cDNA from all tissues was then pooled together and a 1:10 dilution was made to be used as a template for the first round of semi-nested PCR. PCR was carried out with the Sigma Jumpstart Taq DNA polymerase kit on an MJ Research PTC-200 (Bio-Rad Laboratories, Mississauga, ON, Canada), with first round of 25 cycles (45 s at 94°C), annealing (30 s at 56°C) and extension (1 min at 72°C) and second round for 35 cycles using the same program using 1:100 dilution of first round reaction as template. PCR products were separated in 2% agarose gels supplemented with ethidium bromide, under a UV light. High-throughput analysis using Phoretix 1D (Nonlinear Dynamics, Newcastle upon Tyne, UK) facilitated multiple lane band size determinations for all PCR experiments.

### Cloning and sequencing

Second round PCR products of the expected predicted size were then ligated into pGEM-T Easy TA cloning vectors (Promega, Madison, WI, USA) and transformed into GeneChoice High Efficiency GC10 chemically competent cells (Cat No. D-1). Bacterial clones were plated on LB/X-gal/IPTG agar plates and grown overnight at 37°C. A maximum of 24 colonies were picked from each plated transformation and used for colony PCR with standard M13 primers. Two microliters of this colony PCR product was then used as the template for cycle sequencing using Applied Biotech, Inc. BigDye® Terminator v3.1 Cycle Sequencing Kit (cat no. 4336917) and then run on an ABI 3700.

### GO term over-representation

The gene ontology (GO) terms were taken from the non-redundant ermineDB GO database (37). In this database, 3080 genes are annotated with the signal transduction GO term ID 0007165 out of a total 18 506 annotated human gene products. Sampling of the genes was done without replacement, therefore we calculated the probability of sampling $r$ genes annotated to a given GO term by using the hypergeometric distribution:

$$h_{k,N,M}(r) = \frac{\binom{M}{r}\binom{N}{k-r}}{\binom{N+M}{k}},$$

where $k$ is the number of genes in our study set, $M$ the total number of genes in the population annotated to the

term in question, while $N$ is the total number of genes in the population *not* annotated to the GO term. The probability of seeing $r$ or more annotations in our study set using the sum over the upper tail of the hypergeometric distribution is the following:

$$\sum_{i=r}^{k} h_{k,N,M}\ (i).$$

### Accession numbers

The RT-PCR verified sequences that were sequenced were deposited in Genbank, under accession numbers EF491733–EF491822.

## RESULTS

### Overview of PASE

Our goal was to develop an algorithm that would identify novel alternative cassette exons involved in intracellular signaling. We focused on single cassette exons because they make up the significant portion (53–61%) of alternative splicing events in most species (16,17). Our algorithm can be summarized in the following steps (Figure 1):

(1) Identify cassette exon splicing events using splice-site Markov models and a Bayesian classifier.
(2) Translate candidate exons and remove those that have premature stops or altered reading frames.
(3) Select exons that are in conserved sequence regions between species.
(4) Select exons with predicted phosphorylation or protein-binding motifs known to be involved in intracellular signaling.

We discuss each of these steps below.

*Identification of cassette exons.* PASE first scans intronic sequences of REFSEQ gene structures for candidate cassette exons. Candidate cassette exons must have pairs of acceptor and donor splice sites within 30–330 bp of each other. Splice site models were generated by training 12-position first-order Markov chains of over-represented dinucleotides for both acceptor and donor splice sites from a randomly sampled training set of real exons from human REFSEQ and *C. elegans* WORMBASE annotations. We then trained a Bayesian classifier to discriminate acceptor–donor pair scores of real exons from pseudo-exons of similar length. Pseudo-exons are any pair of acceptors and donors generated from a random background intronic sequence distribution. This classifier achieved an average of 96% accuracy on multiple training set runs for both *C. elegans* and human test sets (See Methods section).

*Exon amino acid translations.* Alternative exons that introduce frameshifts or premature stop codons are under strong negative selection and are not likely to be functional because they disrupt protein structure (18). Therefore, PASE filters out predicted exons with in-frame stop codons or whose length in base pairs was not a multiple of three, as these cause frameshifts.

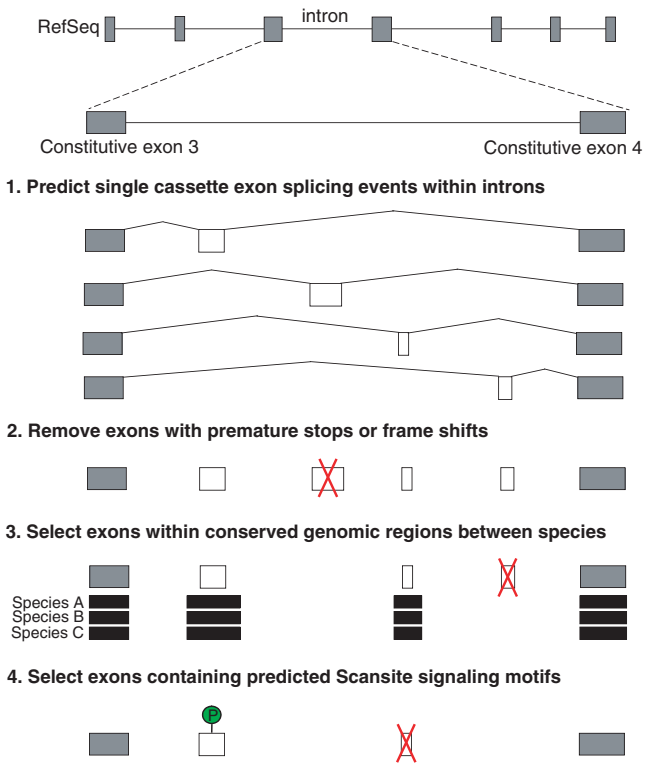


**Figure 1.** Overview of the PASE algorithm. (1) PASE first scans intronic sequences of RefSeq genes for candidate cassette exons. (2) Predicted exons that introduce frameshifts or premature stop codons are removed. (3) Exons that overlap blocks of highly conserved genomic sequence elements are then selected. (4) Finally, the remaining cassette exons are searched with Scansite motifs to identify candidate-signaling exons.

*Species conservation of exons.* Sequences encoding alternative exons are significantly more conserved than neutral sequences (19,20). Furthermore, orthologous exons that are alternative in other species are often found to be more conserved than orthologous constitutive exons and they often have conserved intronic sequences flanking them (14,21). PASE requires predicted exons to overlap sequences that are identified as conserved by the PhastCons phylo-HMM program. This algorithm identifies blocks of highly conserved genomic sequence elements using results from multiple sequence alignments of up to 17 different vertebrate species when compared to human (22). For *C. elegans*, only *C. briggsae* was used for comparative genomics.

*Signaling interaction motifs.* To find cassette exons that encode for signaling motifs, we modified the Scansite 2.0 algorithm (http://scansite.mit.edu) (23) and used it to identify cassette exons with intracellular signaling motifs. Scansite uses a database of experimentally generated position-specific scoring matrices (PSSMs) to identify signaling motifs such as kinase phosphorylation sites, SH2 and SH3 domains (Figure 2). We modified this algorithm to use an information-content-based

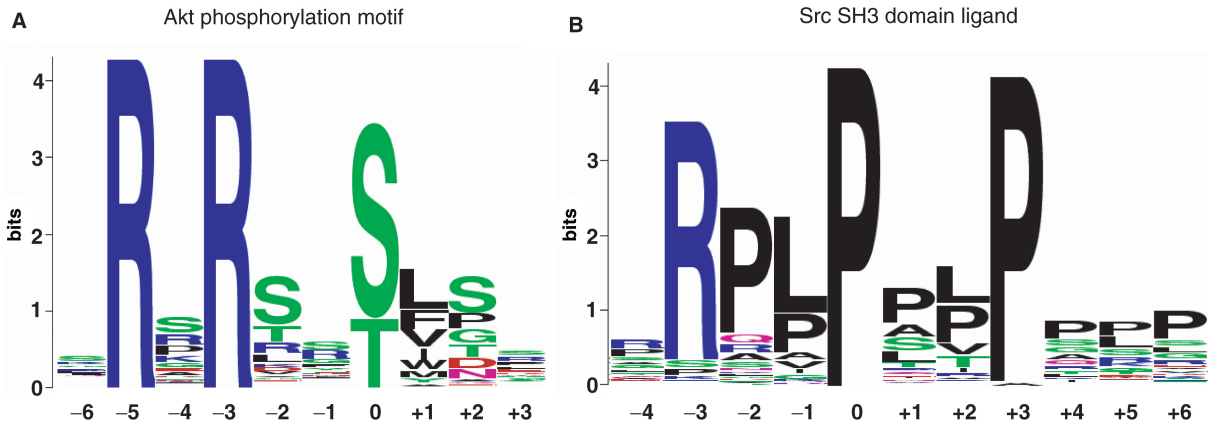


**Figure 2.** Examples of Scansite motifs used in PASE. (**A**) Akt kinase (also known as Protein Kinase B) is a member of the basophilic serine/threonine-specific protein kinase family that selectively phosphorylates the serine/threonine residue (position 0) of protein sequences resembling the linear motif R-X-R-X-X-S/T. (**B**) The non-catalytic Src Homology 3 (SH3) domain of the tyrosine kinase Src mediates specific protein–protein interactions by binding to ligands with the linear motif resembling R-X-X-P-X-X-P.

scoring system, and to account for the species-specific background frequencies of the different amino acids. These modifications were important for improving motif score thresholds in our application (see Supplementary Data). We then scored all putative exons against PSSMs for 59 cell signaling motifs (Supplementary Table A).

## Prediction of alternatively spliced signaling exons in *C. elegans*

*Caenorhabditis elegans* is a model organism that is often used to study signal transduction because it shares many pathways with human and mouse but is more amenable to rapid genetic manipulation. Furthermore, relatively few ESTs have been sequenced from *C. elegans* (300 000 ESTs sequenced from worm versus $\sim 7 \times 10^6$ sequenced from human), so this represents an ideal organism to try an *ab initio* approach to find novel alternative exons involved in signaling. Using the *C. elegans* genome sequence as input, PASE predicted 140 putative alternative exons involved in signaling (Table 1). Seventy-four of these could be identified in the *C. elegans* spliced-ESTs database (Supplementary Table B). The remaining 66 predictions represent either novel alternative exons or false positives and were selected for experimental validation.

## Experimental validation of novel *C. elegans* predictions

We pooled *C. elegans* total RNA from a mixture of developmental stages. Next we performed semi-nested RT-PCR specific to the predicted alternative exons followed by agarose gel electrophoresis (Figure 3). Bands of the predicted size were cloned and sequenced to determine if the predicted exon was present in the product. We evaluated 66 predictions and found 33 novel exons with correctly predicted 5′ splice junctions (50%, see Table 2). These results, together with the 74 predictions already supported by the EST data, indicate the specificity [TP/(TP + FP)] of our algorithm is ≥76.4% (107/140) in *C. elegans*.

**Table 1.** EST and experimentally validated PASE predictions in *C. elegans* and human

| | *C. elegans* | % validated | Human | % validated |
|---|---|---|---|---|
| RefSeq entries | 21 584 | – | 22 615 | – |
| Introns | 118 457 | – | 197 864 | – |
| Translatable exons | 6 008 | 13.1% | 207 176 | 6.2% |
| Conserved exons | 815 | 38.1% | 5 160 | 22.8% |
| Scansite > 10 bits exons | 113 | 37.2% | 5 489 | 19.4% |
| Scansite > 6 bits exons | 823 | 30.8% | 35 190 | 13.7% |
| Conserved & Scansite > 10 bits exons | 20 (18) | 90.0% | 109 (57) | 52.3% |
| Conserved & Scansite > 6 bits exons | 140 (107) | 76.4% | 804 (334) | 41.5% |

The number of validated predictions is in parentheses.

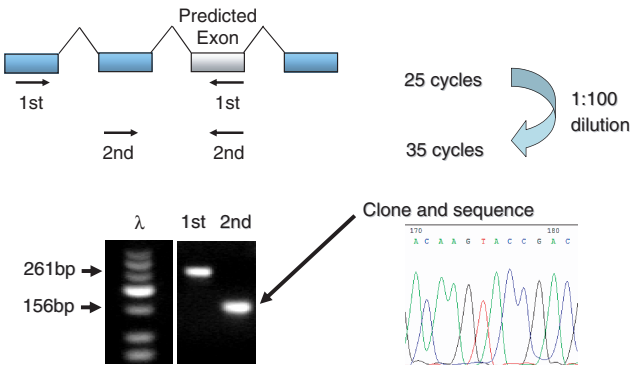


**Figure 3.** Semi-nested RT-PCR approach to detect novel exons from a pool of 18 tissue RNA samples. In the first round of PCR, an external forward primer targeted to a 5′ upstream canonical exonic sequence is used with a reverse primer targeted to the predicted exon in question. A 1:100 dilution of this first round reaction is then used as the template for the second round of PCR. The second round PCR then uses an internal forward primer targeted to an exonic region between the external forward primer and the previously used reverse exon primer. As an example, the novel *C. elegans* exon in gene ZK180.2 is shown with the expected PCR product band sizes (261 bp first round PCR, 156 bp second round PCR). The second round reaction was cloned and sequenced for validation of the prediction (DNA ladder is labeled as λ).

### Analysis of the criteria used to predict exons in *C. elegans*

In order to understand the relative importance of the three filters used by PASE—donor and acceptor site pairs, sequence conservation, Scansite score—we analyzed the enrichment of our validated alternative exons (either experimentally or from EST data) from the set of all candidate exons at different steps in our algorithm. Because our algorithm does not use EST information to make its prediction, this is a reasonable estimate of the specificity of the algorithm. PASE scanned 118 457 introns in 21 584 genes. After applying the acceptor–donor exon model, Bayesian classifier and reading frame filter PASE found 6008 translatable exon predictions for worm (Table 1). Here, ~13% of these exon predictions are observed as splice variants in the EST data (Table 1). After selecting exons that are conserved between *C. elegans* and *C. briggsae*, 815 predictions remained, of which, 38.1% were validated splice forms (Table 1). Thus, a 3- to 4-fold enrichment in documented splice variants is observed when predicted translatable exons are limited to regions of highly conserved sequence. Next, we analyzed the contribution of the intracellular signaling motifs without using a conservation filter. When a 6-bit (or 10-bit) scoring threshold was used, 823 (or 113) of the 6008 exons met this criteria, 30.8% (or 37.2%) of which were in the EST database or experimentally validated. When the conservation filter and the 6-bit threshold were combined, 140 exons passed the filter, 74 (52.8%) of which were validated by spliced-ESTs. Together with the 33 novel validated exons, a total of 107 (76.4%) exons were validated. From this analysis, we conclude that conservation and Scansite score are largely independent filters, and while both contribute significantly to the specificity of the algorithm, conservation plays a slightly larger role.

### Prediction of alternatively spliced signaling exons in human

We next sought to use PASE to find novel signaling exons in humans. We used PASE to scan 197 684 human introns from 22 615 genes (Table 1) and predicted 804 AS exons involved in signaling. Of these, 308 (38.5%) could be found in the human spliced-ESTs library (Supplementary Table C). The remaining 496 predictions are either novel exons or false positives of our algorithm. We ranked these predictions by Scansite bit score and selected the top 384 predictions for experimental validation.

### Experimental validation of novel human predictions

Since testing each of the 384 predictions individually across each tissue would require a large number of experiments, we decided to pool total RNA from 18 human tissues (Supplementary Table D). In order to compensate for the increased dilution of already low expressed splice forms, we performed a sensitive, semi-nested RT-PCR approach specific to the predicted alternative exons followed by agarose gel electrophoresis (Figure 3). Bands of the predicted size were cloned and sequenced. Of the top 384 predictions that were not in the EST database, we found 26 (6.7%) with correctly predicted 5′ splice junctions that were validated using this approach (Table 3). Overall, our algorithm achieved a specificity of ≥41.5% (334/804) in human (Table 1).

### Analysis of the criteria used to predict exons in human

We wanted to understand the role each filter played in distinguishing novel exons from pseudo-exons, so we calculated how each step of our algorithm enriched for *bona fide* exons. PASE scanned 197 684 human introns (~1075 MB) and applied the acceptor–donor exon model, Bayesian classifier and reading frame filter to produce 207 176 predictions (Table 1). Six percent of these predictions are supported by experimental evidence (either present in the EST database or validated by our RT-PCR experiments). We next applied the conservation filter. Of the 207 176 exons, only 5160 (2.4%) were conserved across vertebrate species. Twenty-two percent of these had experimental support, somewhat lower than that observed in *C. elegans*. We also separately searched the 207 176 translatable exons for signaling motifs with log-odds scores greater than 6 bits (or 10 bits) and found that 35 190 (or 5489) exon predictions had one or more motifs that met or exceeded this score threshold (Table 1). In this subset of predictions, the enrichment for matches with expressed sequences was slightly lower than in *C. elegans* with 13.7% (19.4% for 10 bits) of these high-scoring exons matching already observed spliced-EST patterns in human. When both species conservation and high-scoring Scansite motif criteria are combined, there is a significant increase for the enrichment of expressed sequences with 308 of 804 (38.5%) predictions matching with spliced-ESTs, reflecting ~2-fold increase in specificity. Thus, including the 26 experimentally validated exons, the total specificity was ≥41.5%. The results are similar to those observed in *C. elegans*: both the conservation and Scansite filters contribute significantly to our specificity and these filters are largely independent.

### Determining the sensitivity of the PASE algorithm

A test set was created using splicing event data from the Alternative Splicing Database (ASD) (24) and a list of experimentally verified phosphorylation sites in eukaryotic proteins from PhosphoELM (25). A set of 3797 single cassette exon splicing events were extracted from the ASD database and used in a BLASTX comparison with the PhosphoELM database. A total of 20 PhosphoELM sites had 100% sequence identity matched with a single cassette exon (Supplementary Table E). We then applied PASE to the corresponding REFSEQ gene structures that lack the signaling cassette exons. Fourteen of the exons were correctly predicted on both acceptor and donor splice sites as well as correctly identifying the phosphorylation site (total sensitivity $(TP/TP+FN) = 70\%$). Of the false negatives, two cassette exons were correctly predicted, but PASE missed the phosphorylation site. Three exons had incorrect donor sites predicted, of which one had the phosphorylation site predicted correctly. One exon in the test set was not predicted as well as having a missed phosphorylation site.

**Table 2.** Thirty-three novel *C. elegans* exons with predicted interaction sites confirmed by cloning and sequencing

| Wormbase ID | Gene Description | Intron | Best hit (human homolog) | Score | Motif |
|---|---|---|---|---|---|
| F29C4.8 | Collagens (type IV and type XIII), and related proteins (COLlagen) | 3 | p85_SH3_m1 | 12.6825 | PPGLRGSPGWPGLPG |
| ZK180.2 | GABA-B ion channel receptor subunit GABABR1 homolog | 3 | PKC_common | 10.8279 | FGWKRVGTVKQNDQP |
| K11C4.5 | $Ca^{2+}$ release channel (ryanodine receptor) | 9 | Casn_Kin2 | 10.2192 | KDVLEEETEEQEPIW |
| K02H8.1 | It encodes a muscleblind-like. Would have nucleic-acid-binding activity | 3 | Grb2_SH2 | 10.023 | NTPIYPPYYNGMMYP |
| C14A11.3a | Guanine nucleotide exchange factor for Rho and Rac GTPases | 1 | PKC_zeta | 8.8355 | KKYGFWGSVFSKYCF |
| T13H5.1 | Protein tyrosine phosphatase | 13 | PLCg_SH3 | 8.831 | KRPHQVPPMKVDPEG |
| C29A12.4 | Neurexin III-alpha/GLIoTactin (*Drosophila* neuroligin-like) homolog | 23 | Casn_Kin2 | 8.6202 | QITDGDESEDEFDGS |
| H16O14.2 | Encodes a putative nuclear protein | 1 | ErkDD | 8.4135 | KKPPNMHINIPTDEG |
| C44C10.9 | Encodes a Collagen structural gene | 3 | Src_Kin | 8.3384 | YSTPDDIYSAYEKFI |
| F36H1.4c | Abnormal cell LINeage | 5 | GSK3_Kin | 8.2932 | MVHHPNQTISTTPSS |
| T22B2.4 | Predicted RNA-binding protein SEB4 (RRM superfamily) (SUPpressor) | 4 | PIP3_PH | 8.1225 | MGTKKSEFLS RMCV F |
| F10C1.7c | Nuclear envelope protein lamin, intermediate filament superfamily | 2 | PKA_Kin | 8.0228 | RKEFKRETENGTDKM |
| C02F12.1 | Tetraspanin family integral membrane protein | 3 | PDGFR_Kin | 7.5919 | KDKFSNNYMGVYLKN |
| C23F12.1a | Actin-binding cytoskeleton protein, filamin | 3 | p85_SH3_m1 | 7.5172 | EPLGGGVPKQPVQFY |
| Y49F6B.9 | Predicted E3 ubiquitin ligase | 6 | Cam_Kin2 | 7.3624 | HPCPRCKTLIVKEND |
| F33D4.2e | Inositol 1,4,5-trisphosphate receptor (Inositol Triphosphate Receptor) | 19 | DNA_PK | 7.1926 | IGKMSQDSQSDYDSD |
| T12A7.6 | Putative protein, unknown function | 3 | Src_Kin | 7.1049 | YAKTDLIYDDWKFDN |
| Y67D8C.10b | Calcium transporting ATPase (Membrane Calcium ATPase) | 5 | PKA_Kin | 7.0479 | HHREHRDSHHQAQNQ |
| H10D18.5 | Encodes a putative nuclear protein | 2 | p85_SH3_m2 | 6.9925 | IDNKPLFPYMHFAQF |
| T02G6.2 | It encodes a putative nuclear protein. predicted to localize in the nucleus | 4 | Clk2_Kin | 6.9219 | RVEYRYHSETLLYDF |
| C32C4.1 | Voltage-gated $K^+$ channel KCNB/KCNC | 5 | Casn_Kin1 | 6.8928 | KEFTGITSGWPFLGA |
| ZC518.1a | Bestrophin (Best vitelliform macular dystrophy-associated protein) | 11 | GSK3_Kin | 6.8602 | RADSPDDSSHDSCSH |
| C06H5.6 | Synaptic vesicle transporter SVOP and related transporters | 6 | PKC_zeta | 6.8268 | IKKYINESVAFNKQT |
| F28B4.2 | Guanine-nucleotide releasing factor | 1 | Fgr_Kin | 6.7201 | VPQYHMQYFTFDKIT |
| F08A10.1a | Ca2+-activated K+ channel proteins (intermediate conductance classes) | 3 | Src_SH2 | 6.6538 | PRRKRVDYDQISMNW |
| F53A3.4 | Prion-like-(Q/N-rich)-domain-bearing protein | 1 | ATM_Kin | 6.5667 | HQQSIQFSQFPPPQL |
| H14N18.4a | Gamma-glutamyltransferase | 7 | Casn_Kin1 | 6.4654 | KDMPDSETINKAPDH |
| Y67D8C.9 | Puromycin-sensitive aminopeptidase and related aminopeptidases | 2 | Itk_SH3 | 6.4406 | QSKKKTPPRVVERLI |
| C14F11.5 | Alpha crystallins (heat shock protein) | 4 | Cdk5_Kin | 6.3331 | TVTPEQRSPGRKAFE |
| C36F7.4a | Immunoglobulin C-2 Type/fibronectin type III domains (neuRonal IGCAM) | 3 | Erk1_Kin | 6.2825 | CKADGNPTPTVIWRR |
| T23E1.1 | It encodes a putative membrane protein family member of bilaterial origin. | 2 | p38_Kin | 6.2769 | PMIFSCNSPMGNWAN |
| ZK180.2 | GABA-B ion channel receptor subunit GABABR1 homolog | 3 | GSK3_Kin | 6.2566 | THAKFASSDSHEPHE |
| K05B2.5 | Monocarboxylate transporter | 10 | Nck_2nd_SH3 | 6.0161 | ADTADGMPQLQDQDN |

**Table 3.** Twenty-six novel human exons with predicted interaction sites confirmed by cloning and sequencing

| Gene symbol | Gene description | Intron | Best Scansite HIT | Score | Motif |
|---|---|---|---|---|---|
| KCNH1 | Potassium voltage-gated channel, subfamily H, | 7 | p85_SH2 | 13.1167 | WEEDPYEYIRMKFDV |
| VPS8 | Vacuolar protein sorting-associated 8 | 4 | p85_SH2 | 11.891 | WEPPVEDYISMTFSE |
| CRYGN | Gamma N-crystallin variant | 3 | Abl_SH2 | 10.5251 | GDGAWVLYEEPNYHG |
| CD58 | Lymphocyte function-associated antigen 3 precursor (Ag3) | 3 | PKC_common | 10.1748 | GKNVTVKTIKKKQKR |
| PLEKHA6 | Phosphoinositol 3-phosphate-binding protein-3 | 7 | Cdc2_Kin | 9.7482 | FPYNYPPSPTVHDKM |
| ANKS1A/ODIN | Ankyrin repeat and sterile alpha motif domain | 4 | PDGFR_Kin | 9.3912 | KYGPFDPYINAKNND |
| PLEKHA6 | Phosphoinositol 3-phosphate-binding protein-3 | 13 | PLCg_SH3 | 8.8428 | ESPPAVPPLPSESRF |
| ESR1 | Estrogen receptor 1 / estrogen receptor alpha | 2 | p85_SH3_m2 | 8.5929 | EKPWQQMPLKGHNDY |
| ITGA6 | Integrin alpha chain, alpha 6 | 5 | Akt_Kin | 8.5524 | PPREQPDTFPDVMMN |
| MGC26733 | Hypothetical protein MGC26733 | 18 | Nck_2nd_SH3 | 8.5351 | KVFDECFPDQPQIGH |
| SNX1 | Sorting nexin 1 isoform c | 11 | PLCg_NSH2 | 8.4544 | RYGQSGNYMELAWHC |
| MTMR6 | Myotubularin-related protein 6 | 6 | PKA_Kin | 8.353 | RPKRRMQSWWATQKD |
| RGPD5 | RANBP2-like and GRIP domain containing 5 isoform | 20 | GSK3_Kin | 8.0857 | FWTSTPSSQPESKEP |
| LRP1 | Low density lipoprotein-related receptor 1 | 19 | Casn_Kin1 | 7.6201 | GDGSDEQTCPEPADN |
| UTY | Tetratricopeptide repeat protein isoform 1 | 12 | Abl_SH3 | 7.3266 | IEEAWSLPIPAELTS |
| PMS1 | Postmeiotic segregation 1 | 4 | PDGFR_Kin | 7.2751 | YMKKSGDYVTVVEDV |
| CREB5 | cAMP responsive element binding protein 5 | 3 | M1433 | 6.7301 | MDFSKGHTWTIVMNA |
| RECQL5 | RecQ protein-like 5 isoform 1 | 6 | Crk_SH3 | 6.5986 | ISTFQSPPPLPSRTL |
| PAK3 | p21-activated kinase 3 | 2 | GSK3_Kin | 6.5144 | FQTSRPVTVASSQSE |
| GTDC1 | Glycosyltransferase-like domain containing 1 | 8 | Nck_2nd_SH3 | 6.3065 | LQEKEREPKMQFNTQ |
| LSAMP | Limbic system-associated membrane protein | 1 | PKC_common | 6.1958 | FKQRKKPTLCRCVVE |
| PPP2R1B | Protein phosphatase 2, regulatory subunit A (PR 65), beta | 15 | p38_Kin | 6.1829 | AAVRDIQSPCRAQGP |
| WDFY3 | WD repeat and FYVE domain containing 3 isoform | 2 | Erk1_Kin | 6.8017 | EKQCALLSPKDFKAT |
| RPS6KC1 | Ribosomal protein S6 kinase, 52kDa, polypeptide | 1 | p38_Kin | 6.619 | PGWWVIT S PNILANQ |
| OSR1/OXSR1 | Oxidative-stress responsive 1 | 3 | GSK3_Kin | 6.5846 | MVGSFANTNHLSRWW |
| RUNX3 | Runt-related transcription factor 3 | 3 | p38_Kin | 6.2715 | SCSCWLPSPHTDFFQ |

### Predicted exons are found in signaling proteins

If our validated alternative exons are functional (i.e. play a role in signal transduction), then we expect to find them predominantly in genes involved in intracellular signaling. On the other hand, if PASE is not finding functional exons, we would expect our predictions to be randomly distributed across all genes (after correcting for differences in the amount of intronic sequence). Therefore, to determine if PASE is finding functional exons, we used a hypergeometric test to calculate *P*-values of the enrichment of our predicted alternative exons to occur in signal transduction genes labeled with term GO:0007165 (See Methods section). The 26 novel human exons discovered in this study mapped to 20 GO annotated genes, and our set of 334 validated exons (covered by ESTs or experimentally validated here) mapped to 218 GO annotated genes. In both cases, a significant enrichment was observed (8 of 20 GO annotated genes, $P = 0.002$, and 62 of 218 GO annotated genes, $P = 3e{-}6$). These results support the conclusion that we are enriching for functional exons. A similar analysis was performed on all the 804 predictions, which mapped to 504 GO annotated genes, 151 of which were signal transduction genes. Interestingly, this complete set of predictions also showed significant enrichment ($P < 4e{-}12$).

### Differential tissue-specific splicing of the novel exons in LRP1 and ESR1

To determine if the inclusion and/or exclusion of the exons occurred in a tissue-specific manner, we performed RT-PCR in each of the 18 human tissues. We focused on two novel exons from two genes, estrogen receptor alpha (ESR1/ERα) and low-density lipoprotein receptor 1 (LRP1), because both genes are involved in important signal transduction pathways with clinical significance (26,27). We used flanking exon primers for the canonical splice junction and semi-nested, exon-specific primers for the novel exon variant. In the case of ERα, we found only breast and liver expressed the novel minor splice variant, but all tissues tested expressed the constitutive splice junction (Figure 4A). This result shows that the novel ERα exon is typically excluded from most tissues, and that its inclusion may be due to tissue-specific splicing mechanisms. In the case of LRP1, we saw a broader distribution of this minor splice variant, with the exception of uterus tissue, which did not show any expression of this novel exon (Figure 4B). The observation that these isoforms are expressed in a tissue-specific fashion lends further support to the idea that these are not stochastic events, but instead regulated to perform a tissue-specific function. Also, these results indicate that the flanking PCR is not sensitive for the detection of these minor splice variants, compared to the exon-specific semi-nested PCR approach.

## DISCUSSION

Our algorithm is the first attempt to predict signaling cassette exons by combining sequence-based exon prediction with additional information from short peptide motifs that are bound or phosphorylated by signaling proteins. This approach, when applied to *C. elegans*, made 140 predictions, 74 of which were present in the *C. elegans* EST database. We experimentally tested the remaining 66 predictions, finding an additional 33 novel isoforms.

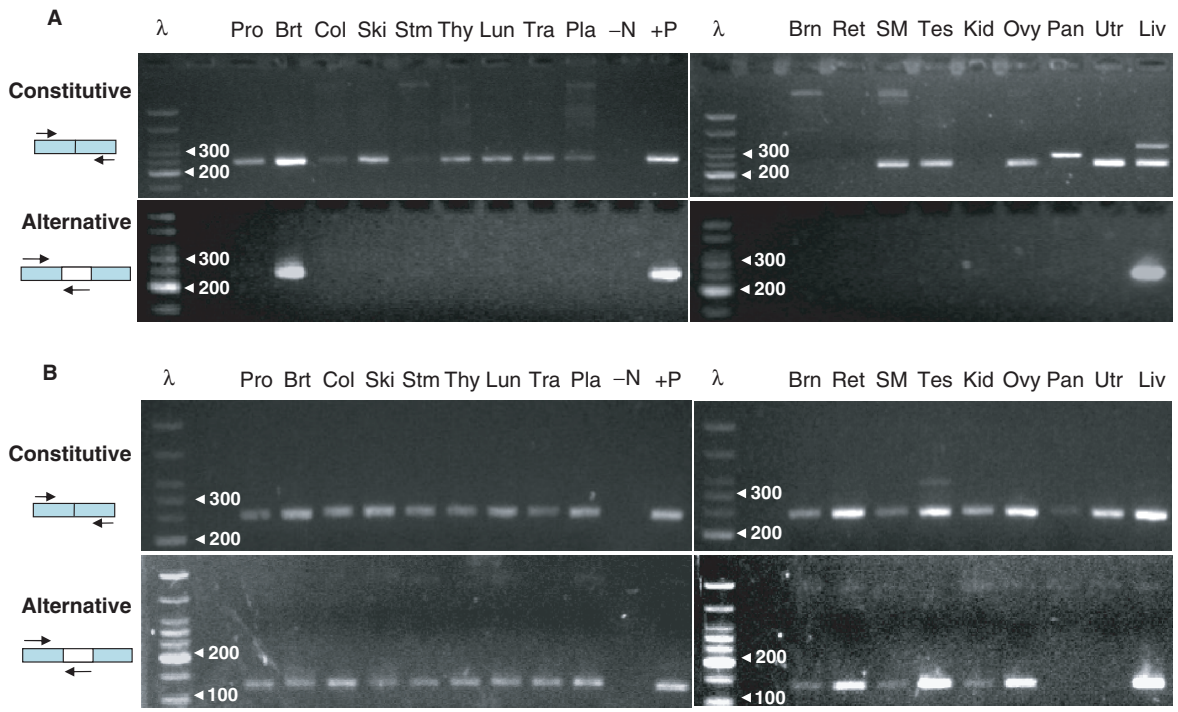


**Figure 4.** Novel exons in ERα and LRP1 exhibit tissue-specific expression. Flanking exon primers were used in RT-PCR tests for expression of the constitutive splice junctions, while exon-specific semi-nested primers were used to test for the expression of the novel alternative exon. (**A**) Expression of the ERα constitutive exon 2–3 splice junction and the novel alternative exon across 18 tissues. The constitutive splice junction is expressed in several tissues, shown as a 234 bp PCR product. The novel exon is shown as a 244 bp PCR product that is exclusively included in breast and liver. (**B**) Expression of the LRP1 constitutive exon 18–19 splice junction and the novel alternative exon across 18 tissues. The constitutive splice junction is expressed in most tissues, showing a 240 bp PCR product. The novel exon is shown as a 128 bp PCR product that is observed in all tissues except uterus. Abbreviations for lanes: DNA ladder (λ), prostate (Pro), breast (Brt), colon (Col), skin (Ski), stomach (Stm), thymus (Thy), lung (Lun), trachea (Tra), placenta (Pla), no template negative control (−N), and pooled tissue control (+P), brain (Brn), retina (Ret), skeletal Muscle (SM), testis (Tes), kidney (Kid), ovary (Ovy), pancreas (Pan), uterus (Utr) and liver (Liv). λ Lanes have size markers spaced at 50-nt intervals up to 350 nt, then the top two bands are for 500 and 766 nt. The strong band corresponds to 200 nt.

Thus, the overall specificity of our algorithm is ≥76.4% (107/140) in *C. elegans*. We also used the algorithm to find human cassette exons, making 804 predictions, of which 308 were found as alternative exons in sequenced ESTs. We experimentally tested 384 of the remaining 496 predictions and discovered an additional 26 novel human exons (total specificity 334/804 ≥ 41.5%). Overall, we discovered 59 novel cassette exons. The human exons that we uncovered are likely to be involved in signal transduction because they were found in the introns of known signal transduction genes more often than expected by chance ($P < 0.003$). Using a test set of known AS phosphorylation sites, we determined that the sensitivity of our algorithm is ~70%.

In both organisms analyzed, a large fraction of predictions were found in the EST database—52.8% in *C. elegans* and 38.5% in humans. Surprisingly, when we experimentally tested the remaining predictions, we saw different discovery rates in the two organisms. In *C. elegans*, 50% of the predicted exons not present in spliced-ESTs were validated; in human, only 6.7% were validated. Why was the discovery rate lower in humans than *C. elegans*? One might hypothesize that because *C. elegans* has shorter introns than humans (~15-fold) and their splice sites have a higher information content (28), PASE makes more accurate predictions in *C. elegans*. However, this seems unlikely to be the main reason because the fraction of predictions covered by ESTs was similar in *C. elegans* and humans. Another possibility is that our experimental approach is not able to detect novel exons in humans as well as it can in *C. elegans*. To validate our *C. elegans* predictions, we used total RNA obtained from whole worms at various stages of their life cycles. This means that RNA from every cell type was present in the sample, which may explain the higher discovery rate. On the other hand, because we analyzed RNA from 18 human tissues, we sampled only a small fraction of human cell types (out of a total possible 210 cell types). Therefore, it is possible that some of our predictions were *bona fide* exons but were not present in our RNA pool, and that our total specificity is at least 41.5%. In fact, the GO term analysis for all 804 predictions ($P < 4e{-}12$) suggests there may be more signaling exons that we have yet to find.

The novel exons found here all encode for phosphorylation or binding sites; this allows one to predict interactions with specific signaling proteins. For example, the novel exon we found in estrogen receptor alpha (ERα) is predicted to have a SH3 class II binding site (K-P-X-X-Q/K-X) targeted by the p85α SH3 domain.

Thus, one would predict that this isoform of ERα interacts with a protein containing a p85α SH3 domain. Indeed, previous work suggests that ERα directly interacts with the p85α SH3 domain of PI3K, but since the canonical ERα protein does not contain a binding site, the mechanism of this interaction is currently unknown (29). The cassette exon found here may explain this interaction. The involvement of the estrogen-signaling pathway in cancer (27), obesity (30,31) and cardiovascular disease (32) makes this a particularly interesting direction for future work. In addition, several other interesting exon predictions coincided with published literature and could be candidates for further investigation (See Supplementary Table C).

Our exon validation approach was designed to facilitate the detection of rare alternative exons in pooled RNA samples, and we found it to be robust and sensitive. One drawback to our validation pipeline was that only the 5′ splice junctions were sequenced—the 3′ splice junctions were not. The correct 5′ junction guarantees the preservation of the reading frame through the novel exon and the presence of the putative interaction site, but the protein might be modified or truncated downstream of the exon. Therefore, for a subset of our predictions, we also confirmed the 3′ splice junctions and found 31 out of 31 of them had the correct predicted 3′ splice junction (See Supplementary Data C and D).

The results presented here demonstrate that PASE is able to find alternative signaling exons with high selectivity. We used PASE to predict exons in humans and *C. elegans* and discovered 59 novel exons, several of which may play important biological roles. Because PASE does not use EST data to predict exons, it may be particularly useful when applied to organisms with low EST coverage (e.g. as was the case with *C. elegans*). Currently, PASE is able to predict alternative single cassette exons—we plan to extend the algorithm to encompass exon extensions, intron retentions and alternative 3′ or 5′ exons and possibly include the prediction of the disruption of putative signaling motifs (33). In addition, we anticipate an improvement in performance when more signaling-related protein features such as sites for acetylation, proteolysis and even protein domains are included in these splicing event predictions.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

## ACKNOWLEDGEMENTS

We would like to thank Gary Stormo, Justin Sonnenburg, Katherine Varley and Jason Gertz for reviewing our manuscript. Tim Schedl and Elaine Mardis for their excellent suggestions concerning our experimental approach. Total RNA samples of *C. elegans* were graciously donated by Jennifer Davila-Aponte from the Sean Eddy lab. We would also like to thank Yun Yue, Lee Tessler and Michael Brooks for helpful discussions. This work and the Open Access publication charges were funded by NIH grant no. 5P50HG003170-03 and GATP training grant no. T32-HG000045, Whitaker Foundation (St. Louis, MO).

*Conflict of interest statement*. None declared.

## REFERENCES

1. Johnson,J.M. *et al*. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141–2144.
2. 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.
3. Okazaki,Y. *et al*. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **420**, 563–573.
4. Zavolan, M., Kondo, S., Schonbach, C., Adachi, J., Hume, D.A., Hayashizaki, Y. and Gaasterland, T. (2003) *Genome Res*, **13**, 1290–1300.
5. Lareau, L.F., Green, R.E., Bhatnagar, R.S. and Brenner, S.E. (2004) *Curr Opin Struct Biol*, **14**, 273–282.
6. Zhou,J. and Cidlowski,J.A. (2005) The human glucocorticoid receptor: one gene, multiple proteins and diverse responses. *Steroids*, **70**, 407–417.
7. Meier,J. and Grantyn,R. (2004) Preferential accumulation of GABAA receptor gamma 2L, not gamma 2S, cytoplasmic loops at rat spinal cord inhibitory synapses. *J. Physiol*., **559**(Pt. 2), 355–365.
8. Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T.A. and Soreq, H. (2005) *Gene*, **344**, 1–20.
9. Birukov,K.G. *et al*. (2001) Differential regulation of alternatively spliced endothelial cell myosin light chain kinase isoforms by p60(Src). *J. Biol. Chem*., **276**, 8567–8573.
10. Freitas,R.A. (1999) Nanomedicine. In: *Appendix C: Catalog of Distinct Cell Types in the Adult Human Body*. Landes Bioscience, Georgetown, TX.
11. Zavolan,M. and van Nimwegen,E. (2006) The types and prevalence of alternative splice forms. *Curr. Opin. Struct. Biol*., **16**, 362–367.
12. Sun,H. and Chasin,L.A. (2000) Multiple splicing defects in an intronic false exon. *Mol. Cell. Biol*., **20**, 6414–6425.
13. Ohler,U., Shomron,N. and Burge,C.B. (2005) Recognition of unknown conserved alternatively spliced exons. *PLoS Comput. Biol*., **1**, 113–122.
14. Yeo, G.W., Van Nostrand, E., Holste, D., Poggio, T. and Burge, C.B. (2005) *Proc Natl Acad Sci U S A*, **102**, 2850–2855.
15. Hiller, M., Huse, K., Platzer, M. and Backofen, R. (2005) *Nucleic Acids Res*, **33**, 5611–5621.
16. Stamm, S., Zhu, J., Nakai, K., Stoilov, P., Stoss, O. and Zhang, M.Q. (2000) *DNA Cell Biol*, 19, 739–756.
17. Mironov,A.A., Fickett,J.W. and Gelfand,M.S. (1999) Frequent alternative splicing of human genes. *Genome. Res*., **9**(12), 1288–1293.
18. Magen,A. and Ast,G. (2005) The importance of being divisible by three in alternative splicing. *Nucleic Acids Res*., **33**, 5574–5582.
19. Modrek,B. and Lee,C.J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet*., **34**, 177–180.
20. Sugnet, C.W., Kent, W.J., Ares, M., Jr. and Haussler, D. (2004) *Pac Symp Biocomput*, 66–77.
21. Sorek,R. and Ast,G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome. Res*., **13**, 1631–1637.
22. Siepel,A. *et al*. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome. Res*., **15**, 1034–1050.
23. Yaffe, M.B., Leparc, G.G., Lai, J., Obata, T., Volinia, S. and Cantley, L.C. (2001) *Nat Biotechnol*, **19**, 348–353.

24. Stamm, S., Riethoven, J.J., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Tang, Y., Barbosa-Morais, N.L. and Thanaraj, T.A. (2006) *Nucleic Acids Res*, **34**, D46–55.
25. Diella,F. *et al*. (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, **5**, 79.
26. Boucher, P., Gotthardt, M., Li, W.P., Anderson, R.G. and Herz, J. (2003) *Science*, **300**, 329–332.
27. Ariazi, E.A., Ariazi, J.L., Cordera, F. and Jordan, V.C. (2006) *Curr Top Med Chem*, **6**, 181–202.
28. Lim,L.P. and Burge,C.B. (2001) A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl Acad. Sci. USA*, **98**, 11193–11198.
29. Simoncini,T., Rabkin,E. and Liao,J.K. (2003) Molecular basis of cell membrane estrogen receptor interaction with phosphatidylinositol 3-kinase in endothelial cells. *Arterioscler. Thromb. Vasc. Biol.*, **23**, 198–203.
30. Mueller,S.O. and Korach,K.S. (2001) Estrogen receptors and endocrine diseases: lessons from estrogen receptor knockout mice. *Curr. Opin. Pharmacol.*, **1**, 613–619.
31. Heine, P.A., Taylor, J.A., Iwamoto, G.A., Lubahn, D.B. and Cooke, P.S. (2000) *Proc Natl Acad Sci U S A*, **97**, 12729–12734.
32. Wang, M., Crisostomo, P., Wairiuko, G.M. and Meldrum, D.R. (2006) *Am J Physiol Heart Circ Physiol*, **290**, H2204–2209.
33. Hiller, M., Huse, K., Platzer, M. and Backofen, R. (2005) Genome Biol, **6**, R58.
34. Burge,C.B. (1998) Modeling dependencies in pre-mRNA splicing signals. In Salzberg,S.L., Searls,D.B. and Kasif,S. (eds), *Computational Methods in Molecular Biology*. Elsevier Science, Amsterdam.
35. Obenauer,J.C., Cantley,L.C. and Yaffe,M.B. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.*, **31**, 3635–3641.
36. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
37. Lee, H.K., Braynen, W., Keshav, K. and Pavlidis, P. (2005) *BMC Bioinformatics*, **6**, 269.