Improved spliced alignment from an information theoretic approach

Miao Zhang1,2 and Warren Gish1,*

1Department of Genetics, School of Medicine, Washington University in St Louis, 4566 Scott Avenue, St Louis, MO 63110, USA and 2Department of Biomedical Engineering, School of Engineering, Washington University in St Louis, 1 Brookings Drive, St Louis, MO 63130, USA

ABSTRACT
Motivation: mRNA sequences and expressed sequence tags represent some of the most abundant experimental data for identifying genes and alternatively spliced products in metazoans. These transcript sequences are frequently studied by aligning them to a genomic sequence template. For existing programs, error-prone, polymorphic and cross-species data, as well as non-canonical splice sites, still present significant barriers to producing accurate, complete alignments.

Results: We took a novel approach to spliced alignment that meaningfully combined information from sequence similarity with that obtained from PSSM splice site models. Scoring systems were chosen to maximize their power of discrimination, and dynamic programming (DP) was employed to guarantee optimal solutions would be found. The resultant program, EXALIN, performed better than other popular tools tested under a wide range of conditions that included detection of micro-exons and human–mouse cross-species comparisons. For improved speed with only a marginal decrease in splice site prediction accuracy, EXALIN could perform limited DP guided by a result from BLASTN.

Availability: The source code, binaries, scripts, scoring matrices and splice site models for human, mouse, rice and Caenorhabditis elegans utilized in this study are posted at http://blast.wustl.edu/exalin. The software (scripts, source code and binaries) is copyrighted but free for all to use.

Contact: gish@blast.wustl.edu

Supplementary Information: http://blast.wustl.edu/exalin/exalin-supplement.pdf

INTRODUCTION

Alignment of expressed sequence tags (ESTs) and full-length mRNAs to a genomic sequence template is an important technique for discovering new genes (Bailey et al., 1998), delineating intron/exon boundaries, identifying putative protein-coding regions, finding single-nucleotide polymorphisms (Buetow et al., 1999; Marth et al., 1999; Picoul-Newberg et al., 1999), identifying alternatively spliced forms (Brett et al., 2000; Kan et al., 2001; Ladd and Cooper, 2002; Mironov et al., 1999; Modrek et al., 2001), and making inferences of UTRs (Kan et al., 2000) and potential regulatory control regions. Erroneous alignments not only undermine our current understanding of genome biology, such as our knowledge of the physical structure of genes, gene products and the patterns of their expression, but also can mislead and confound subsequent wet laboratory and in silico experiments.

Many factors, including the presence of repetitive elements, duplicated exons, paralogs, pseudogenes, sequencing errors and polymorphisms, may influence an alignment result. While accurate results may be obtained for most genes by requiring canonical GT..AG splice junctions to help guide the alignment, non-canonical splice sites are utilized at low but significant frequency and affect the expression of hundreds of human genes (Burset et al., 2000). Exons shorter than a few dozen nucleotides are not uncommon, as well (Black, 2000), and may either go undetected or render fast, heuristic methods inaccurate. Sequencing errors and mutations may also appear at exon/intron boundaries and confound the accurate localization of splice junctions.

Many tools have been developed specifically for spliced sequence alignment, each taking different approaches. For example, Sim4 (Florea et al., 1998) finds identical seed words of default length 12 in a manner similar to BLAST (Altschul et al., 1990), extends alignments in both 5′ and 3′ directions using a greedy algorithm and attempts to fill unaligned regions using shorter seeds. BLAT (Kent, 2002) functions somewhat similarly but offers its own distinctive options that govern speed and accuracy. Both programs prefer GT at donor sites and AG at acceptor sites. BLAT provides pair-wise alignment information but no explicit predictions of splice junctions: splice junctions may or may not coincident with longer gaps in the reported alignments, and their likely locations must be inferred by post-processing the output. A full dynamic programming (DP) approach is taken with EST_GENOME (Mott, 1997), accompanied by a strong preference in its scoring for GT..AG splice sites and a weaker preference for GC..AG. Other spliced alignment programs include Spidey (Wheelen et al., 2001), which uses a greedy algorithm to choose an optimal subset of HSPs (high-scoring segment pairs; Altschul et al., 1990) to work with from a BLAST search (Altschul et al., 1997) and GeneSequer (Usuka et al., 2000), which...
Recursion

uses a hidden Markov model. Each of the spliced alignment programs has been observed to produce disparate results at significant frequency (e.g. see Volfovsky et al., 2003).

To improve the accuracy of spliced alignment in a variety of situations, we developed a tool named EXALIN that utilized information obtained from two orthogonal sources: sequence similarity and splice site models. An information theoretic approach to parameterization was taken, to maximize the power of discrimination of each source. Donor and acceptor splice sites were scored using sums of log-odds ratios obtained from position-specific scoring matrices or PSSMs (Berg and von Hippel, 1987; Stormo, 1988; Stormo and Hartzell, 1989). These models were simple in their design but more complex and information-based, compared with the canonical splicing rules employed by other tools. Sequence similarity scores, which may also be interpreted as sums of log-odds ratios (Altschul and Gish, 1996; Karlin and Altschul, 1990), were targeted to characteristics expected of correct alignments, such as their percent identity (Altschul, 1991; Karlin and Altschul, 1990; States et al., 1991).

Similarity and splice model scores were scaled equivalently using a single logarithmic base (base e). This permitted the information from each source to be meaningfully added, compared and maximized. DP was chosen for the alignment algorithm, due to its guaranteed behavior even in difficult situations and the separation it affords between scoring and alignment (Sankoff and Kruskal, 1983). While DP assured the maximum log-likelihood (MLL) spliced alignment would be found, the method is also slow. For increased speed with only a marginal decrease in accuracy of its splice site predictions, EXALIN could optionally perform restricted DP, guided by the output from a BLASTN search.

EXALIN and most of the other tools mentioned above were evaluated on both idealized and real-world datasets, with EXALIN generally having performed best.

METHODS

Recursion

The DP recursion in EXALIN utilized four discrete states (Fig. 1): Paired (M), two Delete states (X and Y, one for the genome and one for the transcript) and Intron (I). The Intron state could be entered or exited directly from any of the other states. The Paired state included cases of matching and mismatching. The score for aligning the nucleotide at position i in the genomic sequence with the nucleotide at position j in the transcript sequence was $S(i,j)$. Unlike EST_GENOME, which uses linear gap penalties, affine gap penalties were implemented in EXALIN via the q (gap initiation) and r (gap extension) parameters. The penalty for a gap of length n was computed as $q + (n - 1)r$.

For all i along the genomic sequence and all j along the transcript sequence, initialization for the default DP recursion (see EXALIN-DPS below) was as follows:

$$M(0, 0) = M(i, 0) = X(i, 0) = Y(i, 0) = I(i, 0) = 0$$

The recursion itself was then as follows:

$$M(i, j) = \max \begin{cases} M(i - 1, j - 1) + S(i, j), \\ X(i - 1, j - 1) + A(i - y - q), \\ Y(i - 1, j - 1) + S(i, j), \\ I(i - 1, j - 1) + A(i - 1) + A(i - y - q). \end{cases}$$

The use of 0 in the initialization and recursion was in accordance with finding locally optimal alignments (Smith and Waterman, 1981).

SPLICE SITE MODELS

$D(i)$ in the recursion was the donor splice model score associated with position i in the genome being the 5’ end of an intron, while $A(i)$ was the acceptor splice model score associated with position i being the 5’ end of an exon. Splice models were implemented as PSSMs, with PSSM column scores $P(c, n)$ summed to calculate $D(i)$ and $A(i)$. Using collections of genomic sequences aligned at splice junctions, the column scores were computed as the log-odds ratios:

$$P(c, n) = \frac{1}{\lambda} \log \frac{q(c, n)}{p(n)},$$

where $q(c, n)$ was the relative frequency observed for nucleotide n at column c, $p(n)$ was the background frequency of the nucleotide in the genomic sequence (determined from the strand-independent A + T composition), $n \in \{A,C,G,T\}$, $\lambda$ was the scale of the scoring system in units of nats (Karlin and Altschul, 1990, Shannon, 1948) and log represents the natural logarithm. The resultant splice model PSSMs were stored in files along with indicators of the donor and acceptor splice junction columns. For simplicity, EXALIN expected splice site model scores to be expressed in units of nats (i.e. using $\lambda = 1$).

Scores x and y were penalties for, respectively, entering and exiting the Intron state, which reduced the frequency of false positive splice site predictions. Without these penalties, the resultant alignments were typically riddled with small introns and exons. A good choice for x and y is one that satisfactorily balances the frequency of false intron predictions against the possibility of missing real splice sites. Default values were arrived at by manual inspection of results obtained with a variety of settings for these penalties, with $x = 0$ and $y = 0.5$.
parameters, using a training set of roughly 80 randomly selected (out of 10,149 available) human RefSeq mRNAs. In units of nats, the default values for $x$ and $y$ used by EXALIN were 15 and 5, respectively.

### Splice site model training

Models trained for human splice site prediction were produced by an iterative refinement process. A bootstrap dataset of 1754 canonical donor and acceptor splice site junctions of human origin (Ian Korf, personal communication; see Supplementary Materials, Table 6) was first used. The bootstrap PSSMs each covered seven positions surrounding the splice junctions: −3 through +5 for donors and −5 through +1 for acceptors. The 59.06% background A + T composition for the unmasked human genome was used in computing the PSSM scores.

In each round of training, 21,730 human RefSeq mRNA sequences were aligned with EXALIN to the unmasked human genome, using the bootstrap PSSMs in the first iteration and refined PSSMs computed from the previous iteration in subsequent rounds. To construct the refined PSSMs, nucleotide counts were accumulated for 20 nt upstream and downstream of donor and acceptor splice junctions. To decide which contiguous positions should be included in the next donor and acceptor PSSMs, the relative entropy (Karlin and Altschul, 1990) was calculated at each column position $c$ using:

$$H(c) = \sum_{c \in \{A,C,T,G\}} q(c,n) \log \frac{q(c,n)}{p(n)},$$

where $q(c,n)$ was the observed frequency of nucleotide $n$ at column position $c$ and $p(n)$ was the strand-independent background composition of nucleotide $n$ in the genome. Extending upstream and downstream of the splice junction, a position $c$ was included in the PSSM only if $H(c)$ was greater than the threshold value of 0.1 nats, where 1 nat = $\log 2$ bits). This threshold was chosen somewhat arbitrarily, in response to the diminishing returns to be had from models spanning more positions. By this criterion, human donor site models were confined to the 9 positions from −3 through +6 and acceptor site models spanned the 19 positions from −18 through +1.

All told, three rounds of alignment with EXALIN and splice site model building were performed on the human data, at which point little change was observed in the models. Three rounds of training were similarly conducted in building the mouse splice site models, starting with the same human bootstrap PSSMs and using 17,222 RefSeq mRNA sequences from mouse aligned to the mouse genome sequence, which had an unmasked background composition of 58.32% A + T + C. By G-test (likelihood ratio test; Sokal and Rohlf, 1995), the resultant human and mouse splice site models were significantly different (see Supplementary Materials, Table 7).

### Similarity scoring

The default similarity scoring system utilized by EXALIN was $+5$ for a match, $-11$ for a mismatch, $-11$ for the first residue in a gap and $-11$ for each additional residue in a gap. For future reference, we shall call this the $(-5, -11, -11, -11)$ scoring system. To determine an appropriate value for the scaling factor $\lambda$ to use with gapped alignment scoring systems, Monte Carlo simulations were performed using the method of Altschul and Gish (1996) trivially adapted to nucleotide sequences. Using a model sequence length of 5 kb, $10^3$ pairs of sequences were randomly generated, with one member of each pair drawn from the background $A + T$ composition of 59.06% for the unmasked human genome and the other member drawn from the plus strand composition of 21,730 human RefSeq mRNA sequences ($A = 26.2\%, C = 24.6\%, G = 25.0\%, T = 24.2\%$). The mean percent identity in the maximal scoring alignments was 90.0% overall and 96.3% in aligned segments (excluding gaps). A maximum likelihood estimate of 0.245 nats was obtained for $\lambda$. The default value for $\lambda$ used in EXALIN was the marginally higher value of 0.25 nats. By internally multiplying similarity scores by $\lambda$, with $\lambda$ expressed in nats, EXALIN effectively normalized the scores to the same scale as the splice model scores.

### Genomic sequence preparation

Interspersed repeats were masked by RepeatMasker release 09212003 (A.Smit, unpublished data) with the ‘-w -s -no_is -xsmal’ options, which utilized MaskerAid (Bedell et al., 2000) and WU-BLAST 2.0. Repbase update release 08132003 (Jurka, 2000) was searched in this process. Simple repeats were masked using RepeatMasker with the ‘-s -noint -xsmal’ options. Some low-complexity regions not found by RepeatMasker were masked using SEG (Wootton and Federhen, 1996) with the options ‘-o 1.0 1.0 -z 1 -x’ (the options WU-BLASTN uses when invoking SEG). Any nucleotide masked by any of these steps was soft-masked in the final output. A soft-masked mouse genome sequence was produced similarly from the unmasked Build 33, May 2004.

### Scoring systems for human–mouse alignments

Two scoring matrices were constructed, one targeted at aligning human transcripts to the mouse genome and the other at aligning mouse transcripts to the human genome. The log-odds score of aligning two nucleotides $i$ and $j$ in these matrices was calculated according to Karlin and Altschul (1990) as follows:

$$S(i,j) = \frac{1}{\lambda} \log \frac{q(i,j)}{p(i)p(j)},$$

where $q(i,j)$ was the observed (or target) frequency of nucleotide $i$ paired with nucleotide $j$ in orthologous alignments, $p(i)$ was the background frequency of nucleotide $i$ in the genomic sequence, $p(j)$ was the frequency of nucleotide $j$ in the coding strand of the transcript sequences and $\lambda$ is the scale of the scoring system in ungapped alignments.

Target frequencies were obtained from the orthologous CDS alignments of human and mouse provided by Makalowski and Boguski (1998). Background frequencies $p(i)$ were obtained from the genomic $A + T$ composition. The $p'(j)$ were obtained from the entire plus strand (CDS and UTRs included) of the 21,730 human RefSeq mRNAs (see above) and 17,222 mouse RefSeq mRNAs ($A = 25.8\%, C = 24.9\%, G = 25.1\%, T = 24.2\%$).

While $\lambda = 1$ was used in computing the original similarity scores, $S(i,j)$, gapped alignments were permitted with penalties of 2.80 for the first residue and 2.50 for each additional residue. Effective values for $\lambda$ when using such affine scoring systems were estimated through Monte Carlo simulations as described above, but using $10^3$ pairs of model sequences each 2,500 nt in length. For the case of mouse transcripts aligned to the human genome, an estimate of $\lambda = 0.901$ was obtained. For human transcripts aligned to the mouse genome, an estimate of $\lambda = 0.907$ was obtained.

### Three analysis methods in EXALIN

EXALIN provided three basic modes of operation that differed in their accuracy and speed.

**Method 1:** full DP with splice site information (EXALIN-DPS)

Full DP was the default method chosen for EXALIN, because it was guaranteed to provide an optimal answer (including the complete optimal alignment and an accurate alignment score), its behavior was governed by the fewest number of parameters, and only two user-supplied inputs were required: a genomic sequence and a transcript sequence. After completing the recursion, the traceback procedure to produce the alignment was initiated from the highest scoring point in state space. If an approximate location for an mRNA was known in advance (perhaps through a preliminary BLASTN search), the DP could optionally be confined to the relevant segment of the genomic sequence, using the ‘-oA’ or ‘-oB’ options, which would speed up the analysis.

The memory required for the recursion was linearly dependent on the lengths of the two sequences, while the traceback procedure required memory proportional to the product of the sequence lengths. When free memory was determined in advance by the program to be unable to accommodate traceback data for the entire search space, a three-pass procedure was automatically engaged. The first and second passes identified the start and end

---

**Title:** Improved spliced alignment

**Authors:** [List of authors]

**Publication:** [Journal name], [Volume], [Issue], [Year], pp. [Start Page]–[End Page]

**Abstract:** [Abstract text]

---

**Downloaded from [URL]**

**Accessed:** [Date]

**By:** [User name]
points of best-scoring alignments by full DP in the forward and reverse directions, with no traceback information recorded at this time. The third pass then performed DP restricted to the region defined by these start and end points, this time with traceback information maintained. When long genomic contig sequences were involved, the three-pass procedure generally decreased memory requirements below 512 MB for human ESTs aligned to human genes and did not severely degrade speed.

**Method 2: BLASTN-guided DP (EXALIN-BLAST)** The second method utilized results from a preliminary BLASTN search to constrain the DP to regions spanning the ends of HSPs (see Supplementary materials, Fig. 2). Initiation and termination (‘anchor’) points for DP were localized a distance g (default value 40 and user-modifiable with the \(-g\) parameter) from the start and end of each HSP. If an HSP was shorter than g, the distal end was used as the anchor. Except for the 5'- and 3'-terminal HSPs, the method used a global approach with DP, in which \(M(i, j)\) was not compared to 0 in the recursion (Needleman and Wunsch, 1970). For most regions of DP, traceback began at the DP termination point. For the 3'-terminal HSP, traceback began at the highest scoring point in state space. EXALIN-BLAST only reported alignment information in the vicinity of its splice junction predictions, which could limit its utility. Drawbacks to the approach were that the path of the MLL spliced alignment identified by EXALIN-DPS might not be accurately traversed here; and the overall alignment score was inaccurate because it only examined regions local to splice junctions and could double count some portion of the similarity score for exons shorter than 2g. For same-species alignments, EXALIN-BLAST identified splice junctions only marginally less accurately than EXALIN-DPS (data not shown). Besides its speed, an advantage of the method was that the multiple ‘topcombo’ group output of BLASTN could be utilized: when multiple non-intersecting sets of consistent HSPs were reported by BLASTN, as a potential consequence of paralogous gene duplications or pseudogenes being present in the genomic template, EXALIN-BLAST could provide a spliced alignment for each group (see Supplementary materials, Fig. 3).

**Method 3: full DP with a simplified splice site model (EXALIN-SSM)** This method scored splice sites in a manner similar to EST_GENOME: a modest penalty was applied for GT..AG splice sites and a severe penalty was applied for non-canonical sites. The EST_GENOME default scoring parameters were

\[
\begin{align*}
-mismatch & 1 \\
-match & 12 \\
-gap&W & 11 \\
-splice&penalty & 240 \\
-intron&penalty & 100
\end{align*}
\]

This method scored splice sites in a manner similar to EST_GENOME: a modest penalty was applied for GT..AG splice sites and a severe penalty was applied for non-canonical sites.

**Parameters settings**

In preparation for running EXALIN-BLAST, the BLASTN parameters used for aligning transcripts to genomic sequence from the same species were

\[
\begin{align*}
-mismatch & 1 \\
-match & 12 \\
-gap&W & 11 \\
-oneOff & 1
\end{align*}
\]

Unless otherwise indicated, the command line options used with BLAT were

\[
\begin{align*}
-q&=rna \\
-mask&=lower \\
-minIdentity&=1 \\
tileSize & = 8
\end{align*}
\]

When processing full-length RefSeq mRNAs, such as in the micro-exon experiment, the \(-\text{fine} \) option was included. The scoring parameters used by BLAT were unknown to us.

For cross-species alignments, EXALIN used the scoring matrices optimized for comparison of human genomic sequence with mouse transcripts and vice versa, with the corresponding \(A\) values set via the \(-l\) option. For EST_GENOME, the parameters used were

\[
\begin{align*}
-mismatch & 14 \\
-gap&penalty & 25 \\
-slip&penalty & 240 \\
intron&penalty & 380
\end{align*}
\]

For Sim4, a seed size of 8 was used. For Spidey, the \(-st\) parameter was added. For BLAT, the \(-\text{oneOff} = 1 \) option was added.

**Program versions**

EXALIN speed tests were conducted with a hand-optimized version dated May 6, 2005.

**Implementation**

EXALIN was developed in the C programming language. To facilitate EXALIN-BLAST, included PERL scripts named \(\text{Bd} \text{\_deluxe.pm} \) and \(\text{blast} \text{\_deluxe.pl} \) were used to parse BLASTN output into a simpler form.

**RESULTS**

**Idealized alignment of mRNAs to genomic contigs**

A total of 352 idealized mRNA sequences were constructed from the annotations in the NCBI flat file entry for human chromosome 1 (NT_19273.12). For each annotated mRNA feature, the exon segments were extracted from the genomic sequence and spliced together in silico, to produce a virtual mRNA with no discrepancies (polymorphisms or sequencing errors) with respect to the genomic template. A higher proportion of full-length exons would also be present in this dataset than is typical for ESTs, but genomic repeats remained a potentially confounding factor. The resultant mRNAs averaged 1776 nt in length with ~8 exons each, for a mean exon length of 228 nt. Altogether 2739 exons and 2386 pairs of donor and acceptor splice sites were present in the dataset. The mRNAs were typical in their frequency of use of canonical splice junctions, with 2352 (98.6%) canonical donors and 2361 (99.0%) canonical acceptors present. Some overlap may exist between the 352 genes involved in this test set and the genes represented by the 17,230 RefSeq mRNAs used in training the splice site models. Even if a 2-fold redundancy existed in the RefSeq set and its overlap with the test set was complete, the likelihood of a significant bias toward the 60-fold smaller test set would seem to be small.

The idealized mRNAs were aligned back to the genomic sequence using each program, and the exon predictions were verified against the original annotation. An exon prediction was counted as correct only when both its start and end coordinates matched the annotated coordinates exactly. Sensitivity (Sn) and specificity (Sp) measures summarize each program’s performance (Table 1). EXALIN and BLAT scored higher than other tools, while Spidey produced the lowest figures by both measures. None of the programs produced perfect results, though—even when the mismatch and gap penalties were made effectively infinite—due to portions of three mRNA sequences being 100% identical to nearby paralogs. In these cases all of the programs missed the annotated location of several exons.
Table 1. Accuracy of exon predictions in idealized data

<table>
<thead>
<tr>
<th>Method</th>
<th>Predicted</th>
<th>Correct</th>
<th>Sn (%)</th>
<th>Sp (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spidey</td>
<td>2725</td>
<td>2622</td>
<td>95.72</td>
<td>96.22</td>
</tr>
<tr>
<td>Sim4</td>
<td>2722</td>
<td>2641</td>
<td>96.42</td>
<td>97.02</td>
</tr>
<tr>
<td>BLAT</td>
<td>2730</td>
<td>2694</td>
<td>98.35</td>
<td>98.68</td>
</tr>
<tr>
<td>EST_GENOME</td>
<td>2720</td>
<td>2644</td>
<td>96.53</td>
<td>97.20</td>
</tr>
<tr>
<td>EXALIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1,−3,−9,−2)</td>
<td>2734</td>
<td>2697</td>
<td>98.47</td>
<td>98.65</td>
</tr>
<tr>
<td>(1,−3,−11,−2)</td>
<td>2735</td>
<td>2700</td>
<td>98.58</td>
<td>98.72</td>
</tr>
<tr>
<td>(1,−5,−9,−2)</td>
<td>2734</td>
<td>2703</td>
<td>98.69</td>
<td>98.87</td>
</tr>
<tr>
<td>(1,−9,−11,−2)</td>
<td>2739</td>
<td>2715</td>
<td>99.12</td>
<td>99.12</td>
</tr>
<tr>
<td>(1,−1e4,−1.1e4,−2)</td>
<td>2739</td>
<td>2719</td>
<td>99.26</td>
<td>99.26</td>
</tr>
</tbody>
</table>

Overall, 352 mRNAs containing 2739 exons were analyzed. Sensitivity (Sn) = Correctly predicted exons/2739 total exons. Specificity (Sp) = Correct predictions/total predictions. The notation (a, b, c, d) corresponds to the setting of optional EXALIN command line parameters: –ma–nb–qc–rd, where –ma is the match reward score, –nb is the mismatch penalty, –qc is the gap penalty score and –rd is the incremental gap penalty. When optional scoring parameters were used, –li, .38 was also specified, thereby setting the scale to 1.38 nats or 2 bits, –1e4(–104) in the last experiment acted effectively as an infinite mismatch penalty and –1.1e4 was effectively an infinite gap penalty, to prohibit mismatches and gaps from appearing.

Micro-exon prediction

Volfovsky et al. (2003) predicted several novel ‘micro-exons’ ranging in length 5–25 nt with the additional requirement of being internal. Their method began with an alignment of transcripts to genomic contigs using MUMmer (Delcher et al., 2002), followed by post-processing steps to identify good splice junctions and to extract the micro-exon predictions. A variety of spliced alignment tools were tested by those authors for the ability to confirm the predictions. In an analogous experiment, we focused on just the 223 human micro-exons in the dataset. The two micro-exons reported for transcript AF144094.1 were found to be duplicates of one another, leaving 222 unique predictions.

EXALIN-DPS precisely confirmed 213 of the micro-exons and provided better explanations for 7 more, yielding an overall success rate of 220/222 (99%). (See Supplementary materials for the nine cases where EXALIN-DPS differed). In six of the seven improved cases, EXALIN extended the alignment of an adjacent exon by the precise length of the ‘missing’ micro-exon with 100% identity; non-canonical GC..AG splice junctions were utilized just twice. In the 7th case (M21984.1), EXALIN identified 6 consecutive micro-exons that included the original 9 nt prediction within a longer 15 nucleotide micro-exon, each with 100% identity and flanked by canonical splice sites. One of the two predictions missed by EXALIN was flanked by AC and AT, suggestive of tandem U12-dependent splicing events. These sites are relatively rare and were not explicitly sought in our study as they were by Volfovsky et al. (2003).

When they differed from the original predictions of Volfovsky et al. (2003), the raw results obtained with the other programs were examined for similar exon extensions to those found by EXALIN, with their counts revised upward accordingly in Table 2 (‘Revised’ column).

Table 2. Confirmation of 222 micro-exon predictions

<table>
<thead>
<tr>
<th>Method</th>
<th>Original</th>
<th>Raw</th>
<th>Revised</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spidey</td>
<td>25</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Sim4</td>
<td>101</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>BLAT</td>
<td>172</td>
<td>173</td>
<td>179</td>
</tr>
<tr>
<td>EST_GENOME</td>
<td>ND</td>
<td>105</td>
<td>108</td>
</tr>
<tr>
<td>EXALIN-SSM</td>
<td>ND</td>
<td>129</td>
<td>132</td>
</tr>
<tr>
<td>EXALIN-DPS</td>
<td>(default)</td>
<td>213</td>
<td>220</td>
</tr>
</tbody>
</table>

‘Original’ counts of predictions confirmed by the various methods were taken from Table 3 of Volfovsky et al. (2003). Whether the duplicated micro-exon was counted twice in the original experiment is unknown. ‘Raw’ counts are the number of unique predictions confirmed by each method in our hands. A newer version of BLAT was used here than would have been used by Volfovsky et al. (2003). ‘Revised’ counts are the sum of raw counts and the number of cases where a micro-exon prediction was not confirmed but was equally or better explained by the method (see text). ND = not done.

Distribution of discrepancies relative to splice junctions

ESTs are error-prone and polymorphic, which leads to frequent discrepancies (mismatches and insertions/deletions or ‘indels’) when aligned. Splice junctions may appear potentially anywhere within a transcript, as well. To a first approximation then, sequencing errors and polymorphisms may be characterized as independent and identically distributed events relative to splice junctions. When programs align high-quality data using canonical splice sites but their results differ (such as we saw above with idealized data), it may be difficult to say one program is more accurate than another. A program might engender more confidence than another, though, if the discrepancies in its alignments of imperfect data are distributed more uniformly and independently of the splice junctions it predicts.

For a sampling of EST data selected as follows, we assessed the distribution of discrepancies relative to splice junctions using the various programs and tested whether the discrepancies were uniformly distributed. All human EST sequences from dbEST were first aligned to the human genome sequence using BLASTN. Those ESTs observed to align to six randomly chosen human contigs (NT_004321.12, NT_022139.12, NT_078037.12, NT_023929.12, NT_008818.12 and NT_024477.12) were collected. The highest scoring hit for each EST was identified in the complete genome. An EST was discarded if it hit elsewhere besides one of the six contigs with an equal or higher score. The 16,639 ESTs that remained were then aligned to their respective best-hit contigs using the spliced alignment programs. To speed the processing of EST_GENOME and EXALIN, just the relevant subsequence of a genomic contig was processed for each EST, as indicated by the earlier BLASTN search. An additional 20 kb of flanking sequence was included on each side to capture external exons BLASTN might have missed.

Discrepancies were tallied separately for donors and acceptors, over a distance of 35 nt from splice junctions in seven non-overlapping windows (Table 3). Discrepancy rates are reported as a percentage per nucleotide observed within the window. EXALIN fit our expectations best of a program not particularly influenced by the presence of sequencing errors and polymorphic differences. On both sides of its splice junction predictions,
EXALIN exhibited the smoothest distribution of mismatches and indels across all windows, as evidenced by its lowest-of-all standard deviation ($\sigma$) and $X^2$ values. Even in the best case, though, the $X^2$ value of 80 for five degrees of freedom soundly rejects ($P < 10^{-15}$) the null hypothesis that the discrepancies reported in each window were normally distributed with a common mean and variance.

In addition to its consistency, EXALIN was second only to EST_GENOME (the only other DP method tested) in its mean average upstream and downstream discrepancy rates, while reporting 4% more splice junctions than EST_GENOME and 14% more splice junctions than BLAT (Supplementary materials, Table 8).

**Cross-species alignment**

All of the tools were evaluated for their accuracy in aligning transcript and genomic sequences from human and mouse. For this more difficult situation, in which less information from sequence similarity was expected, precisely targeted scoring matrices were constructed as described in Methods. To improve the accuracy of the test, the original transcript sequences of Makalowski and Boguski (1998) were mapped to contemporary RefSeq entries of potentially higher quality (Supplementary materials).

EXALIN was initially used to align each transcript to its native, same-species locus, from which a reference set of intron start and end coordinates located in CDS regions was obtained. The sensitivity and the specificity at detecting the reference introns in cross-species comparisons were then determined for each program (Table 4). An intron or splice junction pair in the reference set was counted as correctly recovered only when both sides of the intron were correctly localized. The DP methods EST_GENOME and EXALIN performed best in this test by several percentage points, and EXALIN was clearly superior to all.

While this test was circular in its use of EXALIN to generate the reference set of introns, our earlier experiment involving same-species data (Table 1) suggests an accuracy of ~98% for the reference set. The complementary error rate of 2% would be insignificant, compared with the cross-species discrepancy rates seen here for each of the programs, including the sensitivity (but not specificity) results for EXALIN.

**Cross-species analysis of the CFTR locus**

Recently, Florea et al. (2005) described an integrated, multi-faceted approach to cross-species alignment of cDNA sequences and applied their method (named AIR) to the CFTR locus in a variety of species. We applied EXALIN in the same test to map annotated human mRNAs from the CFTR region onto the syntenic region of the mouse. For the nine genes annotated in the CFTR region of both species, ten transcripts comprising 136 exons were aligned. Both prior and new results are summarized in Table 5. EXALIN identified splice junctions and exact exon boundaries more accurately than the three other tools tested, including AIR.

### Table 3. Distribution of mismatches and insertion/deletions relative to splice junctions

<table>
<thead>
<tr>
<th>Distance from junction (nt)</th>
<th>0–5</th>
<th>6–10</th>
<th>11–15</th>
<th>16–20</th>
<th>21–25</th>
<th>26–30</th>
<th>31–35</th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>$X^2$</th>
<th>$N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Upstream of donor sites</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spidey</td>
<td>4.09</td>
<td>2.13</td>
<td>1.96</td>
<td>2.02</td>
<td>1.83</td>
<td>1.81</td>
<td>1.90</td>
<td>2.26</td>
<td>0.76</td>
<td>2397</td>
<td>131126</td>
</tr>
<tr>
<td>Sim4</td>
<td>2.40</td>
<td>1.41</td>
<td>1.18</td>
<td>1.09</td>
<td>0.94</td>
<td>1.11</td>
<td>1.30</td>
<td>1.35</td>
<td>0.45</td>
<td>1422</td>
<td>329202</td>
</tr>
<tr>
<td>BLAT</td>
<td>0.72</td>
<td>1.69</td>
<td>1.72</td>
<td>1.73</td>
<td>1.21</td>
<td>1.34</td>
<td>1.93</td>
<td>1.47</td>
<td>0.38</td>
<td>825</td>
<td>117114</td>
</tr>
<tr>
<td>EST_GENOME</td>
<td>1.55</td>
<td>1.16</td>
<td>1.10</td>
<td>1.02</td>
<td>0.93</td>
<td>1.06</td>
<td>1.30</td>
<td>1.16</td>
<td>0.19</td>
<td>289</td>
<td>131038</td>
</tr>
<tr>
<td>EXALIN</td>
<td>1.53</td>
<td>1.44</td>
<td>1.34</td>
<td>1.35</td>
<td>1.21</td>
<td>1.28</td>
<td>1.51</td>
<td>1.38</td>
<td>0.11</td>
<td>80</td>
<td>135200</td>
</tr>
<tr>
<td>(B) Downstream of acceptor sites</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spidey</td>
<td>4.07</td>
<td>2.02</td>
<td>1.64</td>
<td>1.82</td>
<td>1.70</td>
<td>1.58</td>
<td>1.62</td>
<td>2.08</td>
<td>0.83</td>
<td>3122</td>
<td>130207</td>
</tr>
<tr>
<td>Sim4</td>
<td>2.57</td>
<td>1.22</td>
<td>1.19</td>
<td>1.53</td>
<td>1.13</td>
<td>1.20</td>
<td>1.36</td>
<td>1.46</td>
<td>0.47</td>
<td>1439</td>
<td>133103</td>
</tr>
<tr>
<td>BLAT</td>
<td>0.79</td>
<td>1.66</td>
<td>1.63</td>
<td>1.80</td>
<td>1.39</td>
<td>1.43</td>
<td>2.11</td>
<td>1.54</td>
<td>0.38</td>
<td>769</td>
<td>117747</td>
</tr>
<tr>
<td>EST_GENOME</td>
<td>1.49</td>
<td>1.09</td>
<td>1.12</td>
<td>1.14</td>
<td>1.19</td>
<td>1.23</td>
<td>1.46</td>
<td>1.25</td>
<td>0.15</td>
<td>169</td>
<td>130874</td>
</tr>
<tr>
<td>EXALIN</td>
<td>1.46</td>
<td>1.38</td>
<td>1.33</td>
<td>1.31</td>
<td>1.18</td>
<td>1.34</td>
<td>1.59</td>
<td>1.37</td>
<td>0.12</td>
<td>95</td>
<td>135448</td>
</tr>
</tbody>
</table>

Using the indicated spliced alignment programs with a collection of 16639 ESTs, the frequency of discrepancies (mismatches plus insertions/deletions) was determined for 7 non-overlapping, 5 nt windows at increasing distances from predicted splice junctions. Discrepancy rates are expressed as a percentage of the nucleotides observed in each window. $\mu$ is the average discrepancy rate (total discrepancies divided by total nucleotides). $\sigma$ is the standard deviation about the mean of the individual rates reported for the 7 windows. $X^2$ is the continuity-corrected chi-square value for 5 degrees of freedom ($7$ windows – 1 for population totals – 1 for intrinsically determined mean) (Sokal and Rohlf, 1995). $N$ is the mean number of nucleotides per window.

### Table 4. Accuracy of cross-species splice junction predictions

<table>
<thead>
<tr>
<th></th>
<th>Predicted</th>
<th>Correct</th>
<th>Sn(%)</th>
<th>Sp(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Mouse transcripts aligned to the human genome</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spidey</td>
<td>2201</td>
<td>1636</td>
<td>69.5</td>
<td>74.3</td>
</tr>
<tr>
<td>Sim4</td>
<td>2052</td>
<td>1160</td>
<td>49.3</td>
<td>56.5</td>
</tr>
<tr>
<td>BLAT</td>
<td>1890</td>
<td>956</td>
<td>25.3</td>
<td>31.5</td>
</tr>
<tr>
<td>EST_GENOME</td>
<td>2197</td>
<td>1942</td>
<td>82.5</td>
<td>88.3</td>
</tr>
<tr>
<td>EXALIN</td>
<td>2203</td>
<td>2114</td>
<td>89.8</td>
<td>95.9</td>
</tr>
<tr>
<td>(B) Human transcripts aligned to the mouse genome</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spidey</td>
<td>2339</td>
<td>1701</td>
<td>71.3</td>
<td>72.7</td>
</tr>
<tr>
<td>Sim4</td>
<td>2249</td>
<td>1210</td>
<td>50.7</td>
<td>53.8</td>
</tr>
<tr>
<td>BLAT</td>
<td>1960</td>
<td>604</td>
<td>25.3</td>
<td>30.8</td>
</tr>
<tr>
<td>EST_GENOME</td>
<td>2165</td>
<td>1905</td>
<td>79.8</td>
<td>87.9</td>
</tr>
<tr>
<td>EXALIN</td>
<td>2270</td>
<td>2109</td>
<td>88.5</td>
<td>92.9</td>
</tr>
</tbody>
</table>

The accuracy of different tools at detecting donor-acceptor pairs of splice junctions located in coding regions: (A) 2354 human splice junction pairs; and (B) 2383 mouse splice junction pairs. Sensitivity (Sn) = Correct predictions/Total predictions. Specificity (Sp) = Correct predictions/Total predictions.
EXALIN 0.853 0.847 0.985 1 0.980 0.973
Sim4a 0.713 0.602 0.985 0.950 0.889 0.742
BLASTZa 0.015 0.015 0.978 0.993 0.040 0.039

including comparison, EST_GENOME aligned the same two sequences in 0.01 s for the preliminary BLASTN search.

For a comparison of EXALIN-BLAST to Sim4, one of the fastest programs for spliced alignment, 5238 human EST sequences were individually aligned with contig NT_004386 (1 177 200 bp) to which each of the ESTs were known in advance to map. EXALIN-BLAST required 3.7 h for this task on a 400 MHz UltraSPARC II processor, including the time to run BLASTN against the contig, for an average of 2.5 s per EST. Sim4 required 1.5 h for the task, for an average of 1.0 s per EST.

**DISCUSSION**

EXALIN was seen to produce more accurate and complete results than the other tools tested under a wide range of conditions. Its superior performance is attributed to the use of targeted information-based measures for sequence similarity and splice site models, the equivalent scaling of these measures, and the guaranteed behavior of the DP employed. The two programs in our study that used full DP, EST_GENOME and EXALIN, were also the slowest. If only splice junction predictions were of interest, approximately two orders of magnitude greater speed could be obtained using the alternate EXALIN-BLAST search mode, for which the accuracy was only marginally reduced from that of EXALIN-DPS.

Another advantage to using EXALIN-BLAST would be to avoid errors that may be induced by the presence of paralogs, pseudogenes and gene duplications. Spidey performs a similar function, but its clustering algorithm differs significantly from that of BLASTN. Spidey uses a greedy approach that initiates clustering with the highest scoring alignment (Wheelan et al., 2001), whereas BLASTN uses DP to identify HSP clusters guaranteed to have the highest sum of scores. The BLASTN approach may be less prone to being sidetracked by high-scoring, intron-less pseudogenes. BLASTN further allows users to limit the distance along the subject sequence between alignments placed in the same group (re: the hspsepSmax parameter), which can avoid clustering HSPs together from different genes.

Even with idealized, perfect data, we saw how different methods and parameter settings can yield different results (Table 1). In some cases the results may be ambiguous as to which alignment is correct, due to the presence of repeated sequences. Given its more complex splice site models, EXALIN could resolve ambiguous situations and statistically discriminate between canonical GT..AG splice junctions that are naively (equally) weighted by other programs. The importance of splice site information increases in cross-species situations, where relatively less information is available from sequence similarity. Since all pair-wise similarity scores were configurable using a scoring matrix with EXALIN, just as they are with BLASTN, the EXALIN-BLAST search mode may be effective in cross-species analyses, too. The increase in speed over EXALIN-DPS may not be as dramatic, however, after BLASTN parameters are adjusted to obtain satisfactory sensitivity in the presence of the more frequent mutational differences expected with increasing evolutionary distance.

Our results reinforce the idea that information theoretic approaches to sequence analysis combined with robust algorithms are desirable (Altschul, 1991; Altschul et al., 1990; States et al., 1991).

**Acknowledgements**

We thank Dr Gary D. Stormo for reviewing an early draft of the manuscript and for instructive suggestions; Dr Ian Korf for providing the bootstrap dataset of human splice junctions; Karl Haller (University of Arizona, Tucson) for providing splice junction data for rice; LaDeana Hillier for splice junction data for C. elegans; John Kloss for programming advice; Raymond Yeh for the optimized parameters used with EST_GENOME; Dr Jarret Glasscock and German Leparc for helpful discussions; and the Washington University Genome Sequencing Center for generously providing the computational resources used to develop the human and mouse splice site models. Funding to pay the Open Access publication charges for this article was provided by Washington University, St. Louis.

**Conflict of Interest:** none declared.

**REFERENCES**


Karlin, S. and Altschul, S.F. (1990) Methods for assessing the statistical significance of
Kan, Z. et al. (2000) UTR reconstruction and analysis using genomically aligned EST
Karlin, S. and Altschul, S.F. (1990) Methods for assessing the statistical significance of
Ladd, A.N. and Cooper, T.A. (2002) Finding signals that regulate alternative splicing in
the post-genomic era. Genome Biol., 3, review0008.