NAME

SplicePredictorLL − Splice Site Prediction with Logitlinear Models

SYNOPSIS

SplicePredictorLL [**−m** *model*] [**−s** *species*] **−c** *cutoff*] [**−t** *pval*] [**−T** *sval*] [**−oO**] [**−n** *topN*] [**−p** *pstyle*] [**−w** *[nsites sscwl]*] [**−x**] [**−eE** *estdbn*] [**−i** *prmfile*] [**−qQ** *qpfname*] [**−I** *matname*] [**−a** *from*] [**−b** *to*] [**−rR**] [**−lL** *libname*] [**−g** *gbfname(s)*]

DESCRIPTION

The basic version of *SplicePredictorLL* implements logitlinear models for splice site prediction trained on reliable sets of maize and *Arabidopsis thaliana* genomic sequences as described in Reference 1. The predictions are based on the two variables of (i) degree of matching to the splice site consensus and (ii) local compositional contrast. The models assign a *P-value* between 0 and 1 to each potential splice site such that true sites mostly score high and non-sites mostly score low. The *P-value*s represent intrinsic splice site quality. In otherwise constant context, sites with increased *P-value* are predicted to result in more efficient splicing (see Reference 2). Improvements to the basic model include the context-dependent scores *rho* and *gamma* (Reference 3). The *rho-value* of a given site is calculated as a weighted product of its P-value times the P-value of its best potential intron-forming complementary splice site; 0 < *rho* < 1. The *gamma-value* of a site reflects how well this site fits in the locally predicted splicing pattern. If the given site is in a context that suggests preferred usage of nearby sites as splicing partners to the exclusion of the given site, its *gamma-value* will be zero. Otherwise it will be a positive value less or equal to 2; high values of *gamma* would strongly suggest actual usage of the site.

To quickly assess the overall quality of a site we implemented a * grading system: the values of *P*, *rho*, and *gamma* are labeled 5*, 4*, 3*, or 2* if they match or exceed the threshold values for 90%, 80%, 65%, and 50% prediction specificity on the training set, 1* otherwise. The sum of the **-values* (attaining values between 3* and 15*) serves as a simple combined measure. For example, sites scoring 14* or 15* are highly reliable (estimated specificity > 90%).

Minimal input to the program consists of a genomic sequence for which potential splice sites are to be listed. Optionally, the user may also supply cDNA/ESTs or "target proteins" which are known or suspected to significantly match the genomic sequence or its translation into encoded amino acids chains. If supplied, the algorithm will return optimal *spliced alignments* which "thread" the targets into the genomic DNA by scoring for splice sites and sequence similarity in potential exons while allowing for introns as long gaps in the alignment (References 4 and 5).

The **SplicePredictorLL** program was developed in the group of Prof. V. Brendel and is freely available under the GNU General Public Licence at http://brendelgroup.org/bioinformatics2go/SplicePredictor.php/. Correspondence relating to **SplicePredictor** should be addressed to

Volker Brendel
Indiana University
Department of Biology
212 South Hawthorne Drive
Simon Hall 205C
Bloomington, IN 47405
U.S.A.
phone: (812) 855-7074
email: vbrendel@indiana.edu

REFERENCES

1. Kleffe, J., Hermann, K., Vahrson, W., Wittig, B. & Brendel, V. (1996)
Logitlinear models for the prediction of splice sites in plant pre-mRNA sequences.
Nucl. Acids Res. 24, 4709-4718.

2. Brendel, V., Kleffe, J., Carle-Urioste, J.C. & Walbot, V. (1998)
Prediction of splice sites in plant pre-mRNA from sequence properties.
J. Mol. Biol. 276, 85-104.

3. Brendel, V. & Kleffe, J. (1998)
Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in Arabidopsis thaliana genomic DNA.
Nucl. Acids Res. 26, 4748-4757.

4. Usuka, J., Zhu, W. & Brendel, V. (2000)
Optimal spliced alignment of homologous cDNA to a genomic DNA template.
Bioinformatics 16, 203-211.

5. Usuka, J. & Brendel, V. (2000)
Gene structure prediction by spliced alignment of genomic DNA with protein sequences: Increased accuracy by differential splice site scoring.
J. Mol. Biol. 297, 1075-1085.

6. Brendel, V., Xing, L. & Zhu, W. (2004)
Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus.
Bioinformatics 20, 1157-1169.

OPTIONS

−m *model*
set model [0= without / 1= with (default) sub-classification]

−s *species*
Specify either "maize" [default] or "Arabidopsis" for splice site model to be used.

−c *cutoff*
set prediction threshold level

```
[0 = all GU (AG) sites with 50 base flanks
 1 = threshold at 100% sensitivity for training set [default]
 2 = threshold at  95% sensitivity for training set
 3 = threshold at maximal Tau value for training set
]
```

−t *pval* set prediction threshold to *pval* [overrides -c option]

−T *sval* set prediction threshold to *sval* [overrides -c option]

−oO -o: order sites by P-value [default: by position] -O: order sites by *-value [default: by position]

−n *topN*
display top N splice sites

−p *pstyle*
1 (terse=WWW); 2 (default); 3 (very terse=EXDOMINO); 4 (verbose); 5 (spreadsheet)

−w [*nsites sscwl*]
report splice site clusters (>= *nsites* in <= *sscwl* bases; default: 4/1500, appropriate for -T 14 option)

−x [*from to*]
LaTex graphical output in *.tex file(s)

−eE *estdbn*
Read EST sequence data from library file *estdbn*; -e: align + strand only, -E: align + and - strands.

−i *prmfile*
Read parameters for EST matching from file *prmfile*.

−qQ *qpfname*
Read target protein sequence data from library (FASTA-format) file *qpfname*.

−I *matname*
Read amino acid substitution scoring matrix from file *matname*.

−a *from*
Analyze genomic sequence from position *from* [default: 1].

−b *to* Analyze genomic sequence up to position *to* [default: end of sequence].

−r Analyze reverse strand.

−R Analyze both strands.

−lL *libname*
Read (multiple) sequence data from library file *libfname* (FASTA-format).

−g *gbfname(s)*
Read nucleic acid sequence data from GenBank file(s) *gbfname(s)*. If specified, the -g option must be last.

USAGE

Input file format

Genomic DNA input: Sequences should be in the one-letter-code ({a,b,c,d,g,h,i,k,m,n,q,r,s,t,u,v,w,y}), upper or lower case; all other characters are ignored during input. Multiple sequence input is accepted in *library (FASTA) file format* or in *GenBank format*.

Library (FASTA) file format refers to raw sequence data separated by identifier lines of the form starting with ">" followed by the sequence name. For options **-e**, **-E**, **-q**, and **-l**, the name of the sequence is taken to be the first string on the ">" line delimited by space, tab, |, or : starting from position 5. For example, ">gi|idnumber|something-else" is given the name "idnumber". For options **-Q** and **-L**, the name of the sequence is taken to be the first string on the ">" line delimited by space, tab, |, or : starting from position 2. In the above example, the name would be "gi". Typically, this option is appropriate for sequences supplied by the user in the format ">my-sequence-name comments".

Examples (**-e**, **-E**, **-q**, and **-l** options):

```
>gi|sequence1 - upper case
ACGATTGGATCAAAATCCATGAAAGAGGGGAATCTATAGGCGGAATTGAG
CGCCAGCGACTGGCTGCCTTGGCGGGGGAGGCCTTGGCGGA

>SQ;sequence2 - upper case with numbering
       1  ACGATTGGAT CAAAATCCAT GAAAGAGGGG AATCTATAGG CGGAATTGAG
```

```
      51  CGCCAGCGAC TGGCTGCCTT GGCGGGGGAG GCCTTGGCGG A

>vb:sequence3 - lower case
acgattggatcaaaatccatgaaagaggggaatctataggcggaattgagcgccagcgac
tggctgccttggcgggggaggccttggcgga

>vb:sequence4 - mixed format
       1  ACGATTGGAT CAAAATCCAT GAAAGAGGGG AATCTATAGG GGGGGGATCT
cgccagcgac
         tggctgcct        tggcggggg        AGGCCTTGGCGGA
```

GenBank format refers to raw sequence data with possible annotations as in standard GenBank files. Minimal requirements are the LOCUS and ORIGIN lines. Multiple sequences must be separated by // lines.

EST sequence input: EST sequences for spliced alignment may be supplied as a sequence file in library format with the *-eE estdbn* options. Spliced alignment will only be performed for genomic DNA sequences of lengths not exceeding the parameter MAXGLGTH (default: 13000).

Query protein input: Query protein sequences for spliced alignment may be supplied with the *-qQ qpfname* option, where *qpfname* is a sequence file in library format. Spliced alignment will only be performed for genomic DNA sequences of lengths not exceeding the parameter MAXGLGTH (default: 13000).

Parameters

There always is a trade-off between *sensitivity* ("How many true sites will be correctly predicted?") versus *specificity* ("How large is the number of presumably false positive predictions?"). Four settings are optional: "all GU and AG sites" prints out the donor and acceptor model scores at each GU or AG, respectively, in the sequence; "100% learning set" (default) sets the printing threshold at a level that includes all sites that were in our learning sets; "95% learning set" sets the printing threshold at a level that includes 95% of the sites that were in our learning sets; "maximal tau" represents the best compromise between sensitivity and specificity.

Output format

Output is directed to standard output.

Potential splice sites (example):

```
t    q      loc     sequence          P      rho   gamma   *  P*R*G*        parse
  .......
D --->    35713          ccgGTttgt   0.206  0.100  0.191  10 (3 4 3)  IAEEEEE-D-IIIAEED
D ->      35734          tctGTaatt   0.015  0.001  0.000   3 (1 1 1)  AEEEEED-I-IIAEEDI
D -->     35774          atgGTaact   0.223  0.001  0.000   6 (3 2 1)  IIAEEDI-I-IAEEDIA
D ->      35799          ttgGTgtgt   0.008  0.000  0.000   3 (1 1 1)  IAEEDII-I-AEEDIAE
A  <----  35819 ttattaattgcgtAGgt    0.618  0.112  0.538  13 (4 4 5)  AEEDIII-A-EEDIAED
D ->      35820          tagGTtcat   0.005  0.000  0.000   3 (1 1 1)  EEDIIIA-E-EDIAEDA
A     <-  35838 atttcctatacaaAGgg    0.062  0.001  0.000   3 (1 1 1)  EDIIIAE-E-DIAEDIA
D ->      35890          tatGTgatt   0.006  0.000  0.001   3 (1 1 1)  DIIIAEE-D-IAEDIAE
A     <-  35929 tgtgattccttcaAGtt    0.001  0.000  0.000   3 (1 1 1)  DIIAEED-I-AEDIAEE
A     <-  35959 gaatattatcctcAGtt    0.011  0.000  0.008   4 (1 1 2)  IIAEEDI-A-EDIAEEE
```

```
A     <-  36011 accccaaatttaaAGgt     0.003  0.000  0.000   3 (1 1 1)  IAEEDIA-E-DIAEEEE
D ----->  36012           aagGTacga   0.922  0.494  0.933  15 (5 5 5)  AEEDIAE-D-IAEEEEE
A     <-  36076 atatattccttgtAGgc     0.084  0.004  0.000   4 (1 2 1)  IADIAED-I-AEEEEED
A <-----  36100 tcgtgttcattgcAGat     0.816  0.345  0.732  15 (5 5 5)  ADIAEDI-A-EEEEEDI
A     <-  36122 tgttacctgagatAGta     0.003  0.000  0.000   3 (1 1 1)  DIAEDIA-E-EEEEDIA
A     <-  36125 tacctgagatagtAGaa     0.007  0.000  0.000   3 (1 1 1)  IAEDIAE-E-EEEDIIA
A     <-  36128 ctgagatagtagaAGct     0.003  0.000  0.000   3 (1 1 1)  AEDIAEE-E-EEDIIAE
A     <-  36148 tgtatcctttctgAGgt     0.001  0.000  0.000   3 (1 1 1)  ADIAEEE-E-EDIIAEE
A     <-  36166 gatgctgcgctaaAGgc     0.001  0.000  0.000   3 (1 1 1)  DIAEEEE-E-DIIAEEE
D ----->  36206           acgGTaatg   0.494  0.398  1.266  14 (4 5 5)  IAEEEEE-D-IIAEEED
D ->      36250           ttgGTattc   0.006  0.000  0.000   3 (1 1 1)  AEEEEED-I-IAEEEDI
A     <-  36271 tgagattatatcaAGag     0.002  0.000  0.000   3 (1 1 1)  IAEEEDI-I-AEEEDII
A <-----  36296 ataatttttctgcAGtc     0.805  0.371  0.778  15 (5 5 5)  AEEEDII-A-EEEDIIA
  .......
```

Column *t*: type (D, donor, or A, acceptor)
Column *q*: quality. The length of the arrow indicates the site quality measured by the *-value:

```
----- = *value 14-15 = highly likely (estimated specificity   >90%)
----  = *value 11-13 =    likely     (estimated specificity 60-70%)
---   = *value  8-10 =    possible   (estimated specificity 35-45%)
--    = *value  5- 7 =    uncertain  (estimated specificity 10-20%)
-     = *value  3- 4 =    doubtful   (estimated specificity   < 5%)
```

```
The arrow head points into the predicted intron.
```

Column *loc*: site location (position of first or last base of potential intron for D or A, respectively)
Column *sequence*: site sequence
Column *P*: P-value
Column *rho*: rho-value
Column *gamma*: gamma-value
Column *: *-value
Column *P*R*G**: individual *-values for P, rho, and gamma
Column *parse* (not shown): highest scoring assignment of the given site and the seven adjacent sites upstream and downstream as either A (acceptor), D (donor), E (exon), or I (intron)
Note: Spliced alignment with ESTs confirms introns 35713-35819, 36012-36100, and 36206-36296 (see file out.gbA.orig in the GeneSeqer/SplicePredictor distribution data directory).

Spliced alignment: For each significantly matching EST, the predicted gene structure based on an optimal spliced alignment is displayed. The upper line gives the genomic DNA and the lower line gives the EST sequence. Identities are indicated by vertical bars in the center line. Introns are indicated by dots, gaps in the exons by '_'. For protein spliced alignments, the alignment gives the genomic DNA sequence, its inferred protein translation (one-letter-code), and the matching parts of the target protein sequence. Identical residues are linked by "|", positively scoring substitutions by "+", and zero scoring substitutions by "." according to the amino acid substitution scoring matrix used in the alignment. Coordinates for the predicted exons and introns are given in the list preceding the alignment. Exons are assigned a normalized similarity score (1.000 represents 100% identity). For introns, the list gives the P-values of the donor and acceptor sites (Reference 4) as well as a similarity score (s) based on the sequence similarity in the adjacent 50 bases of exon.

Special lines:
MATCH gDNAx cDNAy scr lgth cvrg y

where gDNA = name of genomic DNA sequence; x = + (forward strand) or - (reverse strand); cDNA = name of cDNA sequence; y = + (forward strand) or - (reverse strand); scr = alignment score; lgth = cumulative length of scored exons; cvrg = coverage of genomic DNA segment (y = G) or cDNA (y = C) or target protein (y = P), whichever is highest

PGS_gDNAx_cDNAy (a b,c d, ...)
or
PGS_gDNAx_qp (a b,c d, ...)

where gDNA = name of genomic DNA sequence; x = + (forward strand) or - (reverse strand); cDNA = name of cDNA sequence; y = + (forward strand) or - (reverse strand); qp = name of target protein; a, b, c, d, ... = exon coordinates.

The MATCH and PGS lines are useful for summarizing the search results for an application involving multiple genomic DNA sequences and multiple ESTs or target proteins (use a combination of ’egrep’ and ’sort’). PGS = Predicted Gene Structure (GenBank CDS-styled exon coordinates).

NOTES

The related **GeneSeqer** program implements Bayesian models for splice site prediction (Reference 6). The **SplicePredictor.c** source code includes both the older logitlinear models (compiled by default as **SplicePredictorLL**) and the recent Bayesian models (compiled by default as **SplicePredictor**). The current default **SplicePredictor** will be fully supported and documented as soon as the manuscript is published.

COMPILATION OPTIONS

The following parameters are set in the file
GENESEQER/include/sahmt.h (change and re-compile depending on need and available memory):

MAXGLGTH - maximum length of genomic DNA segment for spliced alignment; default: 15000
MAXCLGTH - maximum length of cDNA/EST for spliced alignment; default: 8000
MAXPLGTH - maximum length of protein sequence for spliced alignment; default: 3000

FILES

GENESEQER/README
GENESEQER/bin
GENESEQER/data (examples)
GENESEQER/doc/SplicePredictorLL.1 (this file)
GENESEQER/doc/SplicePredictor.1 (this file)
GENESEQER/include
GENESEQER/src

SEE ALSO

GeneSeqer(1), SplicePredictor(1).

NOTES

A hardcopy of this manual page is obtained by ‘man -t ./SplicePredictorLL.1 | lpr’.

AUTHOR

Volker Brendel <vbrendel@indiana.edu>