Prediction of exact boundaries of exons

T. A. Thanaraj and Alan J. Robinson

Date received (in revised form): 24th September 2000

Abstract

It is known that while the programs used to predict genes are good at determining coding nucleotides, there are considerable inaccuracies in the determination of the gene structural elements. Among them, the most notable is that of the exact boundaries of exons. In order to assess this, we had earlier reviewed various programs that predict potential splice sites and exons. The results led to the following two observations: (i) a high proportion of false positive splice sites from computational predictions occur in the vicinity of real splice sites; and (ii) current algorithms are misled to predict wrong splice sites more often when the coding potential ends within ±25 nucleotides from real sites than when it ends at farther positions. In this report, we review decision tree models for human splice sites and the resultant software tool, namely SpliceProximalCheck, that discriminates such ‘proximal’ false positives from real splice sites. Further presented is an integrated system (MZF-SPC) with Splice ProximalCheck (SPC) as a front-end tool operating on the results of Michael Zhang’s exon finder program. Examination of the output of the integrated program on an illustrative gene set revealed that as much as 61 of 93 MZF-predicted false positive exons could be eliminated by SPC for a loss of only 3 out of 33 MZF-predicted true positive exons.

INTRODUCTION

Systematic analysis of the performance of publicly available programs for gene prediction highlighted that while the current programs perform well at predicting coding nucleotides, exon boundaries are predicted with lower accuracy levels.1 Specificity of the models for splice signals used currently are only 35 per cent at a sensitivity threshold of 0.5. Various publicly available computational tools that can predict human splice sites were recently benchmarked.2 The programs differed from one another in the degree of discriminatory information used for prediction. A clean data set of EST-confirmed (expressed sequence tag) human splice sites3 was used in the benchmarking studies. Results of benchmarking revealed that one in every three false positives of predicted splice sites (as obtained by programs that use coding potential information in addition to splice signals for predicting splice sites) is located in the vicinity of a real splice site (ie within a distance of ±50 nucleotides). Such an observation

Keywords: gene prediction, exon boundaries, decision trees, splice sites, benchmarking, scoring and test data sets, proximal false positives, SpliceProximalCheck

T. A. Thanaraj

European Bioinformatics Institute,
Wellcome Trust Genome Campus,
Hinxton CB10 1SD, UK

Tel: +44 (0)1223 494430
Fax: +44 (0)1223 494468
E-mail: thanaraj@ebi.ac.uk

Alan J. Robinson

run the EBI Industry Programme with research in visualization, advanced IT, gene prediction, gene expression and protein interactions.

Keywords: gene prediction, exon boundaries, decision trees, splice sites, benchmarking, scoring and test data sets, proximal false positives, SpliceProximalCheck

T. A. Thanaraj

European Bioinformatics Institute,
Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK

Tel: +44 (0)1223 494430
Fax: +44 (0)1223 494468
E-mail: thanaraj@ebi.ac.uk

© HEINER STEWARD PUBLICATIONS 1947-5443. BRIEFS IN BIOINFORMATICS. VOL. 1 404-406. NOVEMBER 2000 403

04-thanaraj.pdf

11:24:55, 8:31 AM
persisted with programs that can predict all the potential exons (including optimal and suboptimal). In a high proportion (greater than 50 per cent) of the partially correct predicted exons, the incorrect ends were located in the vicinity of the real splice sites. Further analysis of the distribution of proximal false positives (in comparison with that of GT/AG dinucleotides, which could act as cryptic splice sites) indicated that the splice signals used by the algorithms are not strong enough to discriminate particularly those false predictions that occur within ±25 nucleotides around the real sites. Thus the programs tend to pick up the exon boundaries in the regions where the coding characteristics disappear. Small shifts due to false predictions around real sites do not greatly change the characteristics that are normally associated with real splice site sequences. Current programs are not sensitive to such subtle changes. It is therefore suggested that specialised statistics that can discriminate real splice sites from such proximal false positives be additionally incorporated in gene prediction programs.

In this paper, use of decision trees to build models that help to discriminate such proximal fake splice sites from real splice sites is demonstrated. Decision trees provide an automated means of segmenting a data set according to a user-specified ‘objective’. The resultant segments are then subsequently used for predictions. The real splice sites of the learning data set were taken from a clean data set of EST-confirmed human splice sites. The proximal false splice sites of the learning data set were generated from regions ±50 to ±50 nucleotides (around real splice sites) that have a high confidence of not containing any other functional splice sites. Decision trees were then built based on only the features local to these splice sites. The decision trees yielded a small set of validation rules that can be used to distinguish proximal false splice sites from real splice sites. The quality of the decision trees built with the learning data set was evaluated using a test data set comprising different EST-confirmed real and false splice sites. The computer program with the implementation of the reported decision tree model is available. An integrated system with SpliceProximalCheck operating on the results of Michael Zhang’s exon finder program, MZEF, is available.

MATERIALS AND METHODS

Derivation of the learning data set
The set of EST-confirmed splice sites from our data set published previously was used as the source of real splice sites. Regions 30 nucleotides in length both upstream and downstream of a subset of these sites were used to generate a control set of false splice sites (the subset included those sites that did not possess potential alternative splice sites in their vicinity). All occurrences of the dinucleotide GT (or AG) in the 30 nucleotide length regions from donor (or acceptor) sites were considered as proximal false donor (or false acceptor) sites.

- Real donor sites – the data set provided 619 real donor sites.
- Proximal false donor sites – both the upstream and downstream 30 nucleotide length regions around 569 real donor sites were found not to possess any alternative functional donor sites. Some 2,650 occurrences of GT (other than the functional GT) were observed in these regions and were considered as false donor sites. A randomly chosen subset of these false splice sites was used in the learning data set and the remaining splice sites were used in the test data set.
- Real acceptor sites – the data set provided 623 real acceptor sites.
- Proximal false acceptor sites – both the upstream and downstream 30 nucleotide
length regions around 573 real acceptor sites were found not to possess any alternative functional acceptor sites. Some 3,206 occurrences of AG (other than the functional AG) were observed in these regions and they were considered as proximal false acceptor sites. A randomly chosen subset of the false splice sites was used in the learning data set and the remaining splice sites were used in the test data set.

Derivation of test data set
Carrying out the clean-up procedures, as described in the earlier work, on new human gene entries not published when building the earlier clean data set generated a test data set of EST-confirmed splice sites. Such a test set included a total of 229 donor and 236 acceptor sites. The false splice sites of this test data set were generated as described above.

Sizes of the learning and test data sets
- **Donor sites.** The learning data set of 2,520 donor sites comprised 619 real splice sites and 1,901 false splice sites. The test data set of 978 donor sites comprised 229 real splice sites and 749 proximal false splice sites.

- **Acceptor sites.** The learning data set of 2,960 acceptor sites comprised 623 real splice sites and 2,337 false splice sites. The test data set of 1,105 acceptor sites comprised 236 real splice sites and 869 proximal false splice sites.

The decision tree approach
A decision tree finds rules that recursively bifurcate a data set in order to produce subsets that contain homogeneous data within subsets and heterogeneous data between subsets. These sets of rules can then be used to classify other data sets. The objective of the decision tree was to discriminate true splice sites from proximal false ones in the learning data set using properties of the nucleotide sequence around the splice sites. The decision tree implemented in the commercial decision support system Decisionhouse was used to classify the learning data set of splice sites. There are other publicly available as well as commercial decision tree systems. The approach used in this review is general in nature and is applicable with other decision tree systems.

Determination of local features characterising splice sites
Different properties of the nucleotide sequence local to the splice sites were evaluated to determine if they could be used to differentiate between real and fake splice sites. These properties were used as input data to the decision tree as candidates for the generation of the rules to distinguish between the real and false splice sites in the learning data set.

Positional nucleotide frequencies
Nucleotide frequencies at every position in the range of −20 to +20 nucleotides around the splice sites were calculated for both real and false splice sites (data not shown). The labelling scheme used for nucleotide positions around the splice junction is as shown in Figure 1.

Comparison of nucleotide frequency distributions in real splice sites with those in false sites revealed the following observations:

- **Donor sites:** the frequency of thymine differed considerably between real and fake splice sites at positions −4 to +6; that of cytosine, consistently from −4 to +7; that of adenine from −4 to +5; and
that of guanine from –5 to +7. Marked
differences could be observed at a
number of other positions as well
(more prominently in the cases of
adenine, guanine and cytosine).

splice acceptor sites

● Acceptor sites: the frequencies of each
of the four nucleotides differed
consistently at positions −1 to −20
(except at the −4 position for adenine
and guanine) between real and fake
splice sites. In addition, frequencies of
thymine, cytosine and guanine differed
considerably at the +1 position.

information content

● The information content at positions
−20 to +20 around the splice sites was
calculated as follows:

\[ R_i = \text{Frequency of base } b \text{ at position } i \times \log_2(\text{Frequency of base } b \text{ at position } i) \]

where \( b \) denotes frequency of base
(b = T, C, A or G) at position \( i \) and \( g_b \)
denotes frequency of base \( b \) in the
regions –20 to –50 and +20 to +50. It
was observed that the nucleotide
positions –3 to +6 across donor sites
and positions –20 to –1 across acceptor
sites carry significant information.

The above results suggest the
appropriateness of using nucleotide
positions around the splice sites as analysis
candidates to build decision trees.

Dinucleotide positions

Preference/avoidance of nucleotides at
certain positions would imply a similar
pattern with regard to the occurrence of
particular dinucleotides at these positions.
Hence, the sequential dinucleotides
involving adjacent nucleotide positions
were used as additional analysis
candidates. Preference/avoidance of
nucleotides at certain positions located
on either side of the splice junction could
imply a potential pattern in the
occurrence of interacting pairs of
mononucleotides across the splice
junction. Twenty-five such long-range
base pairs involving five bases on either
side of the site (excluding the GT/AG
positions) were considered.

Contrast across the acceptor sites

Introns very often end with a
poly-pyrimidine tract.2 Nucleotides
thymine and cytosine occur more
frequently than adenine and guanine in
the –20 to –3 intronic region, while the
sequence is largely random in the +1 to
+20 exonic region. Consequently, the
ratio of the compositional sum of
thymine and cytosine to that of adenine
and guanine is higher in the intronic
region than in exonic region. This was
not observed with the false splice sites,
irrespective of whether the sites were
derived from real introns or from exons
of a gene. The values for contrast for regions
of different nucleotide ranges around
acceptor sites are shown in Table 1.

It is observed that even when the range
is extended as far as 40 nucleotides on
either side of the acceptor site, the
contrast value remains higher for true
splice sites than for false sites. It was
decided to use the –20 to –3/+1 to +18
range in further analyses to calculate
contrast values because such a range is of
medium length and shorter ranges may
not adequately represent odd splice sites.
For this range of nucleotide positions, it
was found that while 92 per cent of
proximal false acceptor sites have a
contrast value less than 3.0, only 31 per
cent of real splice sites have such a value.

For reasons mentioned above, the
following were chosen as analysis
candidates. In the case of donor sites, they
are: mononucleotide positions at –7 to –1

Table 1: Values of contrast for different
ranges of regions around the acceptor
sites

<table>
<thead>
<tr>
<th>Region</th>
<th>Real sites</th>
<th>Proximal false sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>–20 to –3 / +1 to +8</td>
<td>5.33</td>
<td>0.91</td>
</tr>
<tr>
<td>–16 to –3 / +1 to +14</td>
<td>4.87</td>
<td>0.92</td>
</tr>
<tr>
<td>–20 to –3 / +1 to +18</td>
<td>4.08</td>
<td>0.92</td>
</tr>
<tr>
<td>–30 to –3 / +1 to +28</td>
<td>2.89</td>
<td>0.94</td>
</tr>
<tr>
<td>–40 to –3 / +1 to +38</td>
<td>2.27</td>
<td>0.98</td>
</tr>
<tr>
<td>–50 to –3 / +1 to +48</td>
<td>1.94</td>
<td>1.05</td>
</tr>
</tbody>
</table>
and +3 to +7; sequential dinucleotides involving these positions; and long-range dinucleotides involving each of the positions at −5 to −1 and each at +3 to +7. Similar candidates for acceptor sites are: mononucleotides at −7 to −3 and +1 to +5, sequential dinucleotides involving these positions; long-range dinucleotides involving each of the positions at −3 to −7 and each at +1 to +5; and the contrast value.

**RESULTS AND DISCUSSION**

The concept of decision trees

We start with a learning data set consisting of records, each of which has been classified as either a real splice site or a proximal false one. The tree builder utility of the Decisionhouse application builds a binary tree by splitting the data set at each node according to the function of a single analysis candidate. At each branch point of the tree, it determines which analysis candidate makes the best split to bifurcate and separate the records of different classes into different groups (while those of the same class remain within the same group). A more general description of decision trees as applied to gene finding algorithms can be seen from the work of Salzberg and coworkers.9

**The Gini value**

The decision tree recursively segments a starting population of mixed real and false splice sites into subpopulations. A Gini value (named after an Italian economist) measures the goodness of segmentation. Traditionally, a Gini value is a measure of inequality among different groups. It has a low value when there is a similar distribution of the records among different groups, and has a high value when there is an inequality. For an illustrative three-layer tree (data not shown) built for acceptor sites with a set of three far-away nucleotide positions namely +18 to +20 as analysis candidates, the Gini value was reported as 11.9 per cent. Such a low value indicated that the match rates at the end nodes (viz., 16.7, 21.1, 21.1 and 26.3 per cent) are not very different from one another and from that (21 per cent) of the starting node. When a set of more relevant nucleotide positions −3 to −5 were specified as analysis candidates, the Gini value was reported as 68 per cent, indicating that the match rates of the end nodes (0.6, 3.8, 12.1 and 50 per cent) are very different from that of the starting node (21 per cent) as well as different from one another. Thus Gini values can be used as a good indication of the validity and quality of the decision tree.

**Decision trees for splice site classification**

Interpretation of decision trees

Figures 2 and 3 show a four-layer decision tree for acceptor sites and donor sites using the learning data set. The decision tree has a Gini value of 89 per cent and has eight end nodes of differing
match rates; a minimum of 0 per cent at node 7 and a maximum of 87.5 per cent at node 14. The decision tree used six of the possible analysis candidates as fields to achieve the best bifurcation at branch points: [−3, +1], [−3, +3], [−5, +3], [−3, +1] and [−3, +3]. Each of the end nodes can be interpreted as a class of splice sites satisfying a specific rule, which is formulated by tracing a path from the root node to the end node. The rule can be termed as a ‘negative rule’ (if the end node is enriched with false splice sites) or as a ‘positive rule’ (if the end node is enriched with real splice sites). Negative rules are of interest in the context of the present work wherein the emphasis is to filter out the proximal false positives during the computational prediction of splice sites.

Inclusiveness of decision tree model

In order to examine how inclusive the tree is, the learning data set was randomly divided into two equally sized sets (subset-1 and subset-2). The decision tree obtained for subset-1 was applied to subset-2 as well as to the full data set. Changes in Gini values were examined in each case. The exercise was repeated with subset-2 being used to generate the decision tree and then applied to the full data set and subset-1. It was found that there was no considerable difference in the Gini values when the tree built for one set was applied to other sets (the maximum difference observed was 5 per cent). It was also observed that the same set of analysis candidates appeared in the decision tree that was built individually for subset-1, subset-2 or for the full data set. This suggests that the decision tree built for a set of records is inclusive.

How meaningful is the segmentation by the decision tree?

In order to test whether the segmentation of the decision tree for the learning data set of real and false splice sites may also be brought about in a random data set, the records of the learning data set for acceptor sites were shuffled. Thus, a subset of false acceptor sites from the learning data set were reclassified as real ones and the remaining splice sites were reclassified as false splice sites. The complete set of records in such a shuffled data set was distributed randomly into three subsets. A decision tree of four layers was built for each such subset and then applied to the other two subsets. Gini values were recorded for each. In a similar manner, a subset of real acceptor sites from the learning data set was reclassified as a false site and the exercise was repeated. It was observed that for the randomly classified data sets, the average Gini value was 38 per cent, much lower than that of 89 per cent obtained for the learning data set.
previously. The Gini value was reduced in an average by 19 per cent (as compared with 5 per cent change with the learning data set as noted earlier) when applied to other subsets.

**Extending the four-layer decision trees**

End nodes of the four-layer decision trees (as shown in Figures 2 and 3) are of the following three types:

- Nodes containing a population enriched with proximal false splice sites and the rules describe the characteristics of a proximal non-functional splice site. Such nodes are 7, 8, 11 from Figure 2 and nodes 7, 11, 12 from Figure 3.

- Nodes containing a population enriched with real splice sites and the rules describe the characteristics of a functional splice site. Such nodes are 14 from Figure 2 and nodes 10, 14 from Figure 3.

- Nodes containing a population enriched with real splice sites at a lower specificity. These rules of these nodes describe the characteristics of a functional splice site, but with a low specificity. Such nodes are 9, 10, 12, 13 from Figure 2 and nodes 8, 9, 13 from Figure 3.

Each of the eight end nodes, especially those of the third type, of the four-layer decision tree should be segmented further. This was achieved by extending the decision tree to further layers until the Gini value is maximised (attaining a value close to 100 per cent). However, a termination criterion is needed to assess whether a split at a branch point is reasonable and is not over-fitting the data. Such a criterion is to stop segmenting a node when its population size is less than 16 or when the split leads to a child node of population size less than 16. A value of 16 was chosen because the dinucleotide analysis candidate could assume 16 different values. The final set of decision trees for donor and acceptor sites was built and the rules were extracted. These are shown in Table 2 for donor sites and Table 3 for acceptor sites.

**Validation rules as derived from the extended decision trees**

- **Donor sites.** A set of nine negative rules (shown in italic font in Table 2) accounted collectively for 92 per cent of the false donor sites from the learning data set with an error rate of 0.6 per cent. On applying the decision tree to the test data set, these nine negative rules identified 89 per cent of the false donor sites with an error rate of 1.5 per cent. A set of six positive rules (shown by normal font in Table 2) accounted collectively for 98 per cent of the real donor sites from the learning data set with a specificity of 79 per cent. On applying the decision tree to the test data set, these six positive rules identified 96 per cent of the real donor sites with a specificity of 73 per cent.

- **Acceptor sites.** A set of 12 negative rules (shown in italic font in Table 3) accounted collectively for 86 per cent of the false acceptor sites from the learning data set with an error rate of 0.3 per cent. On applying the decision tree to the test data set, these 12 negative rules identified 83 per cent of the false acceptor sites from the test data set with an error rate of 2.4 per cent. A set of ten positive rules (shown by normal font in Table 3) accounted collectively for 99 per cent of the real acceptor sites from the learning data set with a specificity of 65 per cent. On applying the decision tree to the test data set, these ten positive rules identified 92 per cent of the real acceptor sites with a specificity of 60 per cent.

The set of 9 negative rules was enough to classify 92 per cent of false proximal
Table 2: Validation rules for donor sites and their performances on the test data set

<table>
<thead>
<tr>
<th>End nodes (Figure 3) (and scheme of classification)</th>
<th>Population of end nodes of the tree for the learning set. Given in brackets are those from test set when the tree was applied</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Rules as derived from the four-layer tree (see Figure 3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 – (i)</td>
<td>111162 [40168]</td>
<td>(l+5-x) ⇒ RT (GA) &amp; (l-2-[-1]) ⇒ CC-AQ &amp; (l+3-x) ⇒ AN (GA)</td>
</tr>
<tr>
<td>8 – (i)</td>
<td>11221 [4377]</td>
<td>(l+5-x) ⇒ RT (GA) &amp; (l-2-[-1]) ⇒ CG-AQ &amp; (l+3-x) ⇒ AN (GA)</td>
</tr>
<tr>
<td>9 – (ii)</td>
<td>26166 [757]</td>
<td>(l+5-x) ⇒ GY &amp; (l-2-[-1]) ⇒ CG-AQ &amp; (l+3-x) ⇒ AN (GA)</td>
</tr>
<tr>
<td>10 – (iv)</td>
<td>4154 [21/26]</td>
<td>(l+5-x) ⇒ GY &amp; (l-2-[-1]) ⇒ CG-AQ &amp; (l+3-x) ⇒ AN (GA)</td>
</tr>
<tr>
<td>11 – (v)</td>
<td>01189 [180]</td>
<td>(l-5-x) ⇒ TQ &amp; (l-1-[-5]) ⇒ TN (TCA)(G) &amp; (l+3-x) ⇒ AN (GA)</td>
</tr>
<tr>
<td>12 – (v)</td>
<td>3021 [005]</td>
<td>(l-5-x) ⇒ TQ &amp; (l-1-[-5]) ⇒ TN (TCA)(G) &amp; (l+3-x) ⇒ AN (GA)</td>
</tr>
<tr>
<td>13 – (VII)</td>
<td>14984 [23/4]</td>
<td>(l+2-[-5]) ⇒ NG (A,C,T) &amp; (l-1[-5]) ⇒ TN (TCA)(G) &amp; (l+3-x) ⇒ AN (GA)</td>
</tr>
<tr>
<td>14 – (VII)</td>
<td>532632 [188/233]</td>
<td>(l+2-[-5]) ⇒ NG (A,C,T) &amp; (l-1[-5]) ⇒ TN (TCA)(G) &amp; (l+3-x) ⇒ AN (GA)</td>
</tr>
</tbody>
</table>

2. Rules as derived from the extended tree

| I | 111162 [40168] | (l) |
| II,1 | 0013 [002] | (l-4-[-6]) ⇒ GN (Y) & (l) |
| II,2.1 | 0054 [0010] | (l-4-[-3]) ⇒ CT (A,C) & (l+3-x) ⇒ GN (Y) & (l) |
| II,2.2 | 1144 [9/23] | (l-4-[-3]) ⇒ CT (A,C) & (l+3-x) ⇒ GN (Y) & (l) |
| III,1 | 0064 [0020] | (l-2-[-5]) ⇒ TN (CT)(CA) & (l) |
| III,1.1 | 5690 [61/9] | (l-7-[-6]) ⇒ GN (A,C,T) & (l-3-[-5]) ⇒ TN (CT)(CA) & (l) |
| III,1.2 | 2142 [31/8] | (l-7-[-6]) ⇒ GN (A,C,T) & (l-3-[-5]) ⇒ TN (CT)(CA) & (l) |
| IV | 4154 [21/26] | (l) |
| V | 01189 [180] | (l) |
| VI | 3021 [005] | (l) |
| VII,1 | 0066 [117] | (l-4-[-7]) ⇒ RA (G,C) & (l-2-[-6]) ⇒ (l) |
| VII,2 | 1438 [117] | (l-4-[-7]) ⇒ RA (G,C) & (l-2-[-6]) ⇒ (l) |
| VII,1.1 | 1333 [117] | (l-4-[-7]) ⇒ RA (G,C) & (l-2-[-6]) ⇒ (l) |
| VII,1.2 | 1632 [21/8] | (l-3-[-4]) ⇒ CT (G,C,A) & (l-1-[-6]) ⇒ NT (G,C,A) & (l) |
| VII,2 | 506558 [1671/88] | (l-1-[-6]) ⇒ NT (G,C,A) & (l) |

Population of learning set comprises 619 real sites. 1,901 false sites with a total of 2,520 sites. Population of test set comprises 229 real sites, 749 false sites with a total of 978 sites. (i) The classification scheme (column 1) used for the end nodes (in the case of extended tree) indicates the number of layers required to segment further the end nodes of the four-layer tree (eg VII,1.2 indicates that the end node 14 of the four-layer tree was segmented to two further layers). Positive rules are given in italics while positive rules are given in normal font. (ii) The population (column 2) given as XY indicates that of the Y number of sites, X are real sites. The population given in square bracket is that for the test set when the tree derived using the learning set was applied on the test data set. (iii) The symbols used in the last column mean as below: ⇒ indicates the Boolean operator "and" equal to; (l) indicates equal to; and & indicates the Boolean operator "or". A = purine (A, G); T = pyrimidine (T, C); X = any nucleotide (T, C, A, G); (l+5-x) represents the dinucleotides XX and TZ, (l-2-[-1]) represents the dinucleotides XY and TZ, (l+3-x) can be split further but it did not improve the model.

Donor sites from the learning data set and a set of 12 rules was enough to classify 86 per cent of the proximal false acceptor sites from the learning data set. Such a set of rules predicted correctly 89 per cent of the false proximal donor sites and 83 per cent of false acceptor sites in a test data set. Given the earlier observations, that one in every three false positives, as well as that more than half the number of wrong ends from partially correct predicted exons occur in the vicinity of real sites, the rules presented herein can be used to help to improve the prediction.

---

356 © HENRY STEWART PUBLICATIONS I 467-5463, BRIEFINGS IN BIOINFORMATICS, VOL. 1, NO. 4, 343-356, NOVEMBER 2000
### Table 3: Validation rules for acceptor sites and their performances on the test data set

<table>
<thead>
<tr>
<th>End nodes (Figure 2)</th>
<th>Rules</th>
<th>Population of end nodes of the tree for the learning set. Given in brackets are those from test set when the tree was applied</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Rules as derived from the four-layer tree (see Figure 2)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| 2. Rules as derived from the extended tree |
| I | 0.0819 [23/84] | (I) |
| II | 0.0128 [0/104] | ([I – 5 – 5] = G{C.T,Q} & [I = 2 – 3.8]) |
| II.1 | 1.2385 [2/102] | ([I – 7 – 1] = NG, G{T,T,G} & [I = 2 – 3.8]) |
| II.2.1 | 0.5144 [1/405] | ([I – 7 – 1] = NG, G{T,T,G} & [I = 2 – 3.8]) |
| II.2.2 | 0.0527 [1/127] | ([I – 7 – 1] = NG, G{T,T,G} & [I = 2 – 3.8]) |
| II.2.2.2 | 0.0655 [2/222] | ([I – 4 – 2] = G{C.T,Q} & [I = 2 – 3.8]) |
| II.2.2.2 | 0.0104 [2/49] | ([I – 4 – 2] = G{C.T,Q} & [I = 2 – 3.8]) |
| II.2.2.2 | 1.144 [1/11] | ([I – 4 – 2] = G{C.T,Q} & [I = 2 – 3.8]) |
| II.2.2.2 | 0.126 [1/10] | ([I – 4 – 2] = G{C.T,Q} & [I = 2 – 3.8]) |
| II.2.2.2 | 39.696 [7/43] | ([I – 4 – 2] = G{C.T,Q} & [I = 2 – 3.8]) |
| II.2.2.2 | 3.243 [1/121] | ([I – 4 – 2] = G{C.T,Q} & [I = 2 – 3.8]) |
| V | 0.0819 [3/104] | (V) |
| VI.1 | 0.0128 [0/104] | ([I – 5 – 5] = G(A.C.G), A{C.T,Q}) & ([I – 3 – I] = G{C.T,Q} & [I = 2 – 3.8]) |

Population of learning set comprises 623 real sites, 2,337 fake sites with a total of 2,960 sites. Population of test set comprises 226 real sites, 869 fake sites with a total of 1,105 sites. (a) The classification scheme (column 1) indicates the number of layers required to segment further the end nodes of the four-layer tree (io, VIII, I) indicates that this end node 14 of the four-layer tree was segmented to two further layers. Negative rules are given in italic font while positive rules are given in normal font. (b) The population (column 2) given as X/Y indicates that of the Y number of sites, X are real sites. The population given in square brackets is that for the test set when the tree derived using the learning set was applied on the test data set. (c) Symbols from last column indicate the Brooker operator (not equal to); (d) equal to; and equals R. The output of (A.G) = Peptide (T.C); N = any nucleotide (T.C.A.G); X.Z represents the dinucleotides XZ and Y.Z; X.Y.Z represents the dinucleotides XY and XZ. (v) VII.2 and VIII could be split further but it did not improve the model.
deterministic model

splice signals

accuracy of the current programs. Thus these rules can either form an integral part of the splice site prediction programs to affect the scoring system or act as filters on the predicted splice sites. The rules are deterministic and they are simple to code in computer programs. It is to be noted that the decision tree models reported herein could identify only 92 per cent of the false donor sites and 86 per cent of the false acceptor sites. The reasons for the low values are probably that the decision tree algorithm implemented in Decisionhouse may not be the most efficient one and the analysis candidates that are used (such as the sequential, long-range dinucleotides and contrast) do not adequately describe biologically relevant information that is part of the splicing mechanism. Characteristics delineating branch points in the case of acceptor sites might have served as another set of analysis candidates, but they could not be used because annotation of them is lacking in the databases.

Splice signals

Examination of the rules (as given in Tables 2 and 3) gave a strong indication of the possibility of interactions involving nucleotide positions that arise from either side of the splice sites, in addition to interactions among bases within either the upstream or downstream regions from the site. In 82 per cent of the cases of real donor sites (as seen from rule VIII.2 in Table 2), the signals were in the form of long-range dinucleotide positions involving –2 to –1 and +5 to +6 and of the sequential dinucleotide positions at (+3, +4). In 72 per cent of the proximal false donor sites (as seen from the rules I and II in Table 2), the signals were in the form of only the sequential dinucleotide positions (–2, –1) and (+3, +4). In 65 per cent of the real acceptor site cases (as seen from rule VIII in Table 3), the signals were in the form of long-range dinucleotide positions (–3, +3) and the sequential dinucleotide position (–7, –6). In 53 per cent of the false acceptor site cases (as seen from rules I and II in Table 3), the signals were in the form of only the long-range dinucleotide position (–3, +1). The significant positional interactions that carry splice signals are enumerated below.

- **Donor sites.** The rules (Table 2) revealed that only 4 of the considered 12 sequential dinucleotide positions carry splice signals. The most prominent dinucleotides are (+3, +4), (–2, –1), (–3, +5, +6) which occurred in the four-layer tree. The fourth one, namely (–7, –6), occurred in the extended tree. The patterns also revealed that only 6 of the possible 25 long-range dinucleotide positions carried splice signals. The most prominent long-range dinucleotides are (–1, +5) and (–2, +5). They occurred in the four-layer tree. The remaining four, namely (–3, +5), (–4, +7), (–3, +4), (–1, +6), occurred in the extended tree. The overall scenario for donor sites is as follows: (i) there are strong long-range interactions from position +5 to positions –1, –3 and to –2; (ii) the sequential dinucleotides at (–2, –1), (–3, +4) and (+5, +6) act as distinct motifs.

- **Acceptor sites.** Contrast has been identified as the primary split candidate emphasizing that the poly-pyrimidine tract present at the end of introns acts as a primary splice signal for acceptor sites. The rules revealed that 6 of the considered 10 sequential dinucleotide positions carry splice signals. The most prominent sequential dinucleotides are (–7, –6) and (–6, –5) and they occurred in the four-layer tree. The remaining four, namely (+1, +2), (+2, +3), (+3, +4), (+5, +6), occurred in the extended tree. The rules also revealed that 8 of the possible 25 long-range dinucleotide positions carried splice signals. The most prominent long-range dinucleotides are (–3, +1), (–3, +3), (–5, +3) and they occurred in the four-layer tree. The remaining six, namely (–4, +1), (–4, +8), (–5, +2), (–5, +3), (–7, +1)
snRNAs

and (–7, +3), occurred in the extended tree. These 14 dinucleotide units were enough to discriminate real acceptor sites from the proximal false sites. The overall scenario for acceptor sites is as follows: (i) the poly-pyrimidine tract is a strong splice signal; (ii) there are strong long-range interactions from position –3 to positions +1 and +3; (iii) the sequential dinucleotides at (–7, –6) and (–6, –5) act as distinct motifs.

The observed possibility that the nearby or long-range nucleotide positions can form dinucleotide units of splice signals has been discussed previously. This is in good agreement with the splicing mechanism of donor site selection. U1 snRNA first base pairs with a region of –4 to +6 around donor site and the same region upstream and downstream regions are later recognised by U5 and U6 snRNAs. Thus the long-range nucleotide positions might depend on one another. In the case of acceptor site selection, the splicing factor U2AF35 recognises the 3’ splice site AG in a sequence-specific manner.

AVAILABILITY OF THE PROGRAMS AND DATA SETS

The learning data set used in this work is available on the WWW. The test data set of EST-confirmed sites is also available from the same web site.

SpliceProximalCheck program

Computer implementation of the trees is available. As input it takes sequences of length 7 nucleotides (in the case of donor sites) or 20 nucleotides (in the case of acceptor sites) from either side of a putative splice site. The sequences are then scrutinised against the validation rules and the putative site is appropriately marked as either a proximal false site, a real splice site or as undecided. The utility of the program becomes obvious under the following situation: when the putative splice site (as predicted by exon prediction programs that use coding potential among others) is identified as a false proximal site by SpliceProximalCheck, then the user can scrutinise the nearby cryptic sites using the same program; thus improving the predictive ability of the programs.

Integration of SpliceProximalCheck with publicly available tools for exon predictions

As discussed earlier, the motivation behind developing SpliceProximalCheck is to enable the further scrutiny of the exons derived by gene prediction programs. Thus, it is highly desirable to present to the gene annotation community an integrated system with SpliceProximalCheck as a front-end tool to exon prediction programs. For this purpose an integrated system with SpliceProximalCheck as a front-end tool to MZEF, a publicly available program for exon predictions, is available.

ILLUSTRATIVE EXAMPLES FOR MZEF-SPC AND PERFORMANCE TESTS

As discussed so far, the reported program specialises in identifying the proximal false positive splice sites. The program labels a given site as either ‘False’ or ‘Possibly true’ or ‘Undecided’. We demonstrate below two real examples of improvement by SPC program on the results of MZEF predictions. The emphasis is on illustrating how the false positive predictions from MZEF are identified by SPC. For this purpose, two human DNA sequences, namely HSERPG and HSU52852, were randomly chosen from the EMBL nucleotide sequence database. The results from MZEF-SPC are shown in Table 4. The predicted exons are ordered as per their P-value (as determined by MZEF). The following observations were made.

Results with HSU52852

- **Acceptor sites:** of the 11 MZEF-predicted false positive acceptor sites, 9 were correctly identified as false
Table 4: Illustrative examples of the results from MZE-F-SPC program

<table>
<thead>
<tr>
<th>MZF. predicted exons</th>
<th>MZE-F P-value</th>
<th>Acceptor sites</th>
<th>Donor sites</th>
<th>Exact exons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>As per EMBL</td>
<td>Output of SPC</td>
<td>Agree*</td>
</tr>
<tr>
<td>I. EMBL ID : HSUS2852</td>
<td></td>
<td>As per EMBL</td>
<td>Output of SPC</td>
<td>Agree*</td>
</tr>
<tr>
<td>3806 – 3957</td>
<td>1.00</td>
<td>T T Yes</td>
<td>T T Yes</td>
<td>T T Yes</td>
</tr>
<tr>
<td>3849 – 3957</td>
<td>0.99</td>
<td>F F Yes</td>
<td>F F Yes</td>
<td>F F Yes</td>
</tr>
<tr>
<td>3906 – 3957</td>
<td>0.99</td>
<td>F F Yes</td>
<td>F T No</td>
<td>F F Yes</td>
</tr>
<tr>
<td>3806 – 3957</td>
<td>0.97</td>
<td>F F Yes</td>
<td>F T No</td>
<td>F F Yes</td>
</tr>
<tr>
<td>4061 – 4117</td>
<td>0.96</td>
<td>T T Yes</td>
<td>F T No</td>
<td>F T No</td>
</tr>
<tr>
<td>4061 – 4187</td>
<td>0.95</td>
<td>T T Yes</td>
<td>T T Yes</td>
<td>T T Yes</td>
</tr>
<tr>
<td>3886 – 3957</td>
<td>0.94</td>
<td>F T No</td>
<td>F T No</td>
<td>F T No</td>
</tr>
<tr>
<td>4277 – 4274</td>
<td>0.92</td>
<td>T T Yes</td>
<td>T T Yes</td>
<td>T T Yes</td>
</tr>
<tr>
<td>4637 – 4606</td>
<td>0.92</td>
<td>F F Yes</td>
<td>T T Yes</td>
<td>F F Yes</td>
</tr>
<tr>
<td>4626 – 4606</td>
<td>0.91</td>
<td>T T Yes</td>
<td>T T Yes</td>
<td>T T Yes</td>
</tr>
<tr>
<td>5584 – 5710</td>
<td>0.87</td>
<td>T T Yes</td>
<td>T T Yes</td>
<td>T T Yes</td>
</tr>
<tr>
<td>5806 – 5900</td>
<td>0.85</td>
<td>T T Yes</td>
<td>T T Yes</td>
<td>T T Yes</td>
</tr>
<tr>
<td>3876 – 3957</td>
<td>0.82</td>
<td>F F Yes</td>
<td>F F Yes</td>
<td>F F Yes</td>
</tr>
<tr>
<td>6770 – 6637</td>
<td>0.79</td>
<td>F F Yes</td>
<td>F F Yes</td>
<td>F F Yes</td>
</tr>
<tr>
<td>6389 – 6406</td>
<td>0.75</td>
<td>F F Yes</td>
<td>F F Yes</td>
<td>F F Yes</td>
</tr>
<tr>
<td>4061 – 4147</td>
<td>0.69</td>
<td>F F Yes</td>
<td>T T Yes</td>
<td>F F Yes</td>
</tr>
<tr>
<td>2676 – 2734</td>
<td>0.67</td>
<td>F F Yes</td>
<td>T T Yes</td>
<td>F F Yes</td>
</tr>
<tr>
<td>6481 – 6406</td>
<td>0.66</td>
<td>F F Yes</td>
<td>F F Yes</td>
<td>F F Yes</td>
</tr>
<tr>
<td>3816 – 3957</td>
<td>0.66</td>
<td>F F Yes</td>
<td>F F Yes</td>
<td>F F Yes</td>
</tr>
<tr>
<td>5804 – 5652</td>
<td>0.57</td>
<td>F F Yes</td>
<td>F F Yes</td>
<td>F F Yes</td>
</tr>
<tr>
<td>3849 – 3961</td>
<td>0.57</td>
<td>F F Yes</td>
<td>F F Yes</td>
<td>F F Yes</td>
</tr>
<tr>
<td>3915 – 3957</td>
<td>0.57</td>
<td>F T No</td>
<td>F T No</td>
<td>F T No</td>
</tr>
<tr>
<td>6437 – 6620</td>
<td>0.55</td>
<td>F T No</td>
<td>F T No</td>
<td>F T No</td>
</tr>
<tr>
<td>6426 – 6620</td>
<td>0.52</td>
<td>F T No</td>
<td>F T No</td>
<td>F T No</td>
</tr>
</tbody>
</table>

'T' under the SPC column means 'possibly true.' The results under 'Acceptor' and 'Donor' columns are shown only for unique acceptor and donor sites. The analysis was extended to eight more entries and the consolidated results are discussed in the text. Exact exons indicates the exons for which both the boundaries are correctly predicted.
by SPC; the remaining two sites were labelled as ‘Possibly true’. In a similar manner, all the 9 SPC-labelled false sites were in deed false positives. All the 7 MZEF-predicted real sites were labelled as true by SPC.

- **Donor sites:** of the 6 MZEF-predicted false positive donor sites, 3 were correctly identified as false by SPC. In a similar manner, all the 3 SPC-labelled false sites were in deed false positives. All the 7 MZEF-predicted real sites were also labelled as true by SPC.

- **Exact exons:** exact exons are those for which both the boundaries are correctly predicted. Of the 19 MZEF-predicted false positive exons, 14 were correctly indicated as false by SPC. In a similar manner, all the 14 SPC-indicated false exons were indeed false positives. All the 6 MZEF-predicted true positive exons were also labelled as true by SPC.

### Results with HSERPG

#### Accuracy of Prediction

- **Acceptor sites:** of the 11 MZEF-predicted false positive acceptor sites, 5 were correctly identified as false by SPC. In a similar manner, 5 of the 6 SPC-labelled false sites were indeed false positive predictions by MZEF. Two of the 3 MZEF-predicted true positives were also labelled as true by SPC.

- **Donor sites:** both MZEF and SPC did not predict any false positives. All the MZEF-predicted true positive sites were also labelled as true by SPC.

- **Exact exons:** of the 11 MZEF-predicted false positive exons, 5 were correctly labelled as false by SPC. In a similar manner, 5 of the 6 SPC-indicated false exons were indeed false positive predictions by MZEF. Two of the 3 MZEF-predicted true positive exons were indicated as true by SPC.

Thus, SPC brings about an improvement by helping to eliminate false positive sites. In order to substantiate the observation, we did the analysis for eight more randomly chosen entries (namely AF101475, AF134406, AF135027, AF166330, AF184072, AF189277, AF195808, AF228497) and consolidated the results. The following observations were made:

- **Acceptor sites:** of the 78 MZEF-predicted false positives, 49 (63 per cent) were correctly labelled by SPC as false sites. Of the 34 MZEF-predicted true positives, 31 were correctly labelled as true by SPC.

- **Donor sites:** of the 15 MZEF-predicted false positives, 9 were correctly labelled by SPC as false sites. All the 40 MZEF-predicted true positives were correctly labelled as true by SPC.

- **Exact exons:** of the 93 MZEF-predicted false positive exons (either one or both the boundaries was wrongly predicted), 61 (66 per cent) were correctly labelled as false by SPC. Of the 33 MZEF-predicted true positives, 30 were correctly labelled as true by SPC.

Thus for a small loss in the sensitivity, SPC brings about a substantial improvement in specificity. Some 60–66 per cent of MZEF-predicted false positives can be eliminated for a loss of only 3 out of ~33 true positive predictions.

### Acknowledgements

The authors thank Martin Senger for help with developing the MZEF-SPC webserver and Alvis Brazma for discussions on decision trees.

### References


5. http://www.ebi.ac.uk/~thanaraj/SpliceProximalCheck.html


7. Quadstone Ltd, http://www.quadstone.co.uk


