Nucleic Acids Research, 1999, Vol. 27, No. 13 **2627–2637**

A clean data set of EST-confirmed splice sites from *Homo sapiens* and standards for clean-up procedures

T. A. Thanaraj*

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

Received April 1, 1999; Revised May 3, 1999; Accepted May 21, 1999

ABSTRACT

A clean data set of verified splice sites from *Homo sapiens* are reported as well as the standards used for the clean-up procedure. The sites were validated by: (i) standard cleaning procedures such as requiring consistency in the annotation of the gene structural elements, completeness of the coding regions and elimination of redundant sequences; (ii) clustering by decision trees coupled with analysis of ClustalW alignments of the translated protein sequence with homologous proteins from SWISS-PROT; (iii) matching against human EST sequences. The sites are categorised as: (i) donor sites, a set of 619 EST-confirmed donor sites, for which 138 are either the sites or the regions around the sites involved in alternative splice events; (ii) acceptor sites, a set of 623 EST-confirmed acceptor sites, for which 144 are either the sites or the regions around the sites are involved in alternative splice events; (iii) genuine splice sites, a set of 392 splice sites wherein both the donor and acceptor sites had EST confirmation and were not involved in any alternative splicing; (iv) alternative splice sites, a set of 209 splice sites wherein both the donor and acceptor sites had EST confirmation and the sites or the regions around them were involved in alternative splicing. A set of nucleotide regions that can be used to generate a control set of false splice sites that have a high confidence of being non-functional are also reported.

INTRODUCTION

The identification of genes is a major objective of genome sequencing efforts. To fulfil this objective, the molecular biology community has turned to computational approaches to help in the identification of genes. Most current computational approaches involve two major steps: (i) identification of gene structural elements (e.g. translation start/stop and splice sites) in an unknown sequence using signals observed in sequences of known structural elements; (ii) identification of potential coding regions from homology searches against databases of protein, EST and cDNA sequences. It is now recognised that accurate prediction of eukaryotic gene structure is dependent largely upon the ability to pinpoint exactly the splice site signals in a sequence (1). This process is complicated by variation in the gene products through 'alternative splicing events' (2).

The above-mentioned procedure to identify the gene structural elements requires an accurate encapsulation of information from known sites. Different models such as patterns, profiles, weight matrices and neural networks have been used to achieve this. The success of these methods depends largely on the quality of the data sets that are used as the training set. Different researchers have created their own data sets in an *ad hoc* manner and until recently there had been no single common data set. A standardised data set is necessary when the prediction accuracy of different programs needs to be assessed. Recently, Burset and Guigo (3) created a common data set to assess the prediction accuracy of these programs and observed that the accuracies they calculated were lower than those reported by the authors originally.

A major observation of this assessment was that while the programs performed well at recognising coding regions, they were poor at predicting the exact boundaries of exons. A reason for this may have been that the training sets used by the authors of these programs had higher quality in terms of identifying coding regions than in terms of identifying the exact position of splice sites. Generally less attention had been paid to validating the exact boundaries of genes. Even the data set that was created by Burset and Guigo (3) for benchmarking did not consider this explicitly. The emphasis in their work was on having a large data set of known gene structures. Consequently, the current programs have drawbacks such as their failure to predict alternative gene products and the accuracy of splice site prediction is better only when it is done 'in context' (1).

The importance of having clean data sets has been described in the literature (4). There is a need for a standardised and clean data set of high integrity wherein much attention is paid to the quality of the splice sites in terms of specifying the boundaries of each exon exactly. In this communication, we report a clean data set of splice sites for human sequences. Matches with human EST sequences at both the donor and acceptor junctions confirm each of the reported sites. The splice sites were categorised further as either genuine or alternative splice sites.

MATERIALS AND METHODS

DNA sequence entries for cleaning were retrieved from the EMBL database (5). The programs that were used in the cleaning processes were FASTA (6), CLUSTALW (7) and Decisionhouse (a commercial decision system from Quadstone Ltd). Human EST sequences were retrieved from the EMBL database

*Tel: +44 1223 494650; Fax: +44 1223 494468; Email: thanaraj@ebi.ac.uk

(5). The SWISS-PROT (8) protein sequence database was used for homology searches to resolve discrepancies. The procedure for data cleaning is described in the following sections.

Standard cleaning procedures

The standard clean-up procedures as adopted by researchers in the field (3,4,9) were followed and improved upon. Human nuclear DNA entries reporting protein coding regions as submitted by individual authors (and not from genome projects) were retrieved from the EMBL database (as of December 1997). The following selection criteria were then applied.

(i) The entries reported genuine human nuclear DNA. Care was taken not to include any synthetic, artificial or foreign genes.

(ii) The entries contained only one CDS or mRNA feature entry.

(iii) The entries did not contain a pseudogene or any alternative gene products, conflicts, variations or mutations in the nucleotide sequence.

(iv) Each entry contained a complete coding region for a gene and had at least one intron.

(v) The description of the gene structure as given in the feature table [mRNA, CDS, exon, intron, 5'- and 3'-UTR, poly(A) signal, etc.] was checked for consistency in annotation.

(vi) Every nucleotide in a region, the end points of which are defined in the CDS or mRNA feature table, had been annotated as from either an exon or an intron.

(vii) Care was taken to ensure that the start of each CDS did not wrongly denote the start of an exon. Similarly, the end of each CDS did not wrongly denote the end of an exon. The gene structure as given by the CDS line was often confirmed by additional feature table information (e.g. mRNA, exon, intron and UTR descriptions).

(viii) The entries were checked to conform to simple sanity checks. For instance: (a) the stop and start codons are standard ones (ATG for a start codon; TAA, TAG or TGA for stop codons); (b) the coding length is a multiple of 3 nt; (c) no in-frame stop codon occurred; (d) the splice sites are marked by the universal consensus dinucleotide sequences, namely GT and AG (with the introns starting with GT and ending with AG).

Eliminating redundant sequences

For each entry, an exonic sequence was constructed by concatenating its constituent exons. The exonic sequences were searched against each other for similarity using FASTA (6). In a similar manner an intronic sequence was constructed for every entry by concatenating its introns. The intronic sequences were also searched for similarity amongst each other. To ensure non-redundancy of the data set, if any group of entries shared >80% identity (either in the exonic or intronic sequences), then only one sequence was retained and the others were discarded.

Data reduction so far

We started with 4300 entries extracted from the EMBL database (as of December 1997). The stringent standard cleaning procedures that we adopted reduced this to 400 entries. Removing the redundant sequences brought this number down to 310. These remaining entries were cleaned further as described below.

Decision trees to identify unusual splice sites

The commercial decision support system Decisionhouse (Quadstone Ltd) was used to classify the splice sites into groups that shared common characteristics using decision trees. Unusual splice sites that do not fall into common groups can be taken for further examination. A decision tree finds rules that recursively bifurcate the data set in order to produce subsets that are homogeneous within subsets and heterogeneous between subsets. The contents of these subsets can be described by this set of rules that use one or more data fields (termed as analysis candidates) of the data. In situations where the incoming data are of uncertain quality, the unusual data are often highlighted when they do not comply with the rules for groups found with the training set. Thus a decision tree is useful for identifying the unusual splice sites that warrant additional scrutiny.

To use such a decision support system, we generated a control set of false donor sites (from the GT dinucleotides that occurred in a region of –20 to +20 nt around the annotated ones) and a set of false acceptor sites (from the AG dinucleotides that occurred in a region of –20 to +20 nt around the annotated ones). A 20 nt region on either side of these sites was extracted and the nucleotides at these positions were used as the analysis candidates in a decision tree for distinguishing false splice sites from true ones. Description of analysis candidates and generation of the false sites are shown in Figure 1a and b.

Illustration of the decision tree approach with donor junctions. A total of 1550 annotated exon–intron sites were mixed with 2912 false sites (generated as described above) and this population was used as input to the decision tree to find rules that could differentiate true exon–intron sites from false ones on the basis of the nucleotides around the junction. A four-layer decision tree (shown in Fig. 2a) produced eight leaf nodes of different match rates. Each of these nodes is described by a simple rule involving a few of the analysis candidates. For example, exon–intron sites belonging to node 7 (Fig. 2a) follow the rule '(li5! = G) AND (le1! = G) AND (li4! = A)', which characterises a population of 1271 sites of which only two are true and the rest are false exon–intron sites. Similarly, node 8 contains a population of 307 sites of which only nine are true exon–intron sites. The 11 true exon–intron sites that are identified with 1567 false exon–intron sites in nodes 7 and 8 may be interpreted either as the results of incorrect annotation or as exceptions and may require further study. The other six nodes (9–14 in Fig. 2a) required further splitting, as the rules generated thus far have not been able to differentiate adequately between true and false exon–intron sites. Figure 2b gives further splitting of one of these nodes (node 14 from Fig 2a). Node 7 (Fig. 2b) shows two true exon–intron sites combined with 30 false exon–intron sites. Nodes 9, 10, 12 and 13 indicate collectively 23 false mixed with 452 real exon–intron sites. Such unusual sites were studied further by examining the information in the corresponding SWISS-PROT (8) entries, by matching with the corresponding mRNA entries (if available) and by studying the CLUSTALW alignment of the translated protein sequence with homologous proteins from SWISS-PROT.

Many of the supposed true exon–intron sites that were identified by the decision tree as being false turned out to indeed be wrong, as indicated by the following observations:

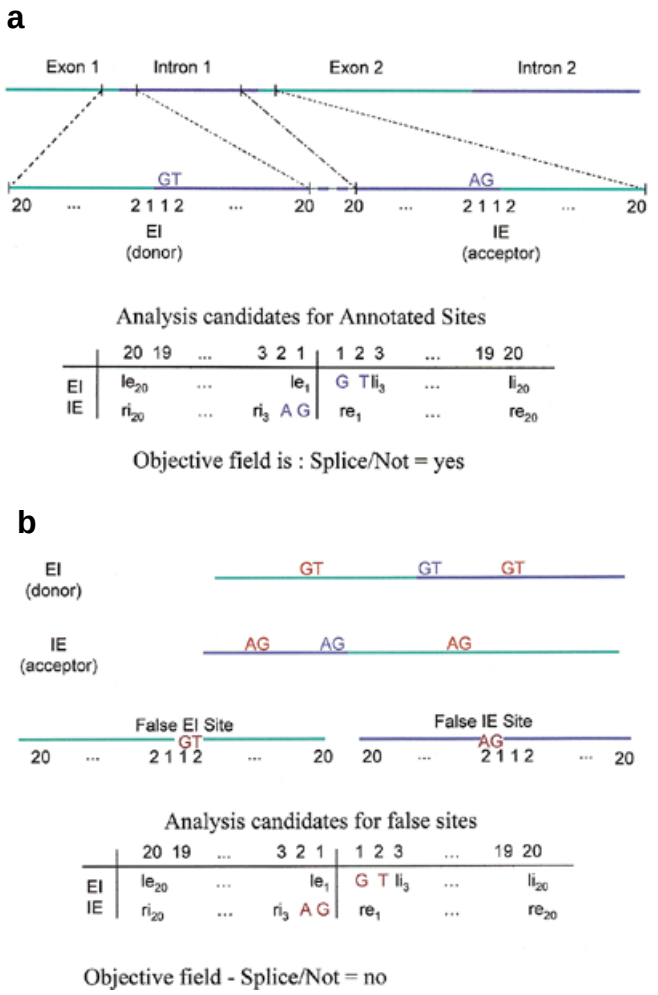

Figure 1. (**a**) Annotated sites for decision trees. Analysis candidates for annotated splice sites. EI represents exon–intron (donor) and IE represents intron–exon (acceptor) junctions. (**b**) Generation of false splice sites and analysis candidates. GT or AG in blue corresponds to annotated donor or acceptor sites; in red corresponds to false sites.

(i) the corresponding mRNA sequence entry did not match at these sites; (ii) mismatches or gaps occurred at corresponding positions in the CLUSTALW alignment; (iii) SWISS-PROT entries report conflicts at these junctions or alternative events.

Similarly, many of the false exon–intron sites identified by the decision tree as being true turned out to be possible correct junctions. As a general rule, we removed such entries from the data set. The exercise was repeated with all the other nodes that needed further splitting (8 and 11 of Fig. 2b) and with all nodes of Figure 2a.

A similar exercise was carried out in the case of intron–exon junctions. At this stage of cleaning the data, we were left with 925 splice sites (pairs of donor and acceptor junctions) derived from 219 entries.

Validation by match with human EST sequences

Every splice site was confirmed by comparing sequence fragments encompassing the sites with human EST sequences, thus to obtain experimental proof for the annotation of the splice sites. While carrying out this exercise, it was decided to examine whether the regions proximal to the sites (and the sites as well) were involved in alternative splice events. This exercise would enable us to select nucleotide regions that could be used to generate false splice sites. Such a set of false sites could be used as a negative control set in any work that uses the data set presented in this report. 5' and 3' EST sequence data from human, as found in the EMBL database, were used to validate these 925 splice sites.

A splice site is characterised by its donor (exon–intron, **EI**) and acceptor (intron–exon, **IE**) junctions (Fig. 3). We generated four characteristic fragments for every splice site of length 50 nt:
exon_EI, the region 50 nt upstream of the **EI** junction;
intron_EI, the region 50 nt downstream of the **EI** junction;
intron_IE, the region 50 nt upstream of the **IE** junction;
exon_IE, the region 50 nt downstream of the **IE** junction.
For every splice site, seven types of query sequence (Fig. 3) were defined that were used to query the EST sequences based on similarity.

Query sequence to confirm both the 5' (donor) and the 3' (acceptor) splice sites. In this query sequence, labelled as the **exon_EI-exon_IE** sequence, the **exon_EI** and **exon_IE** fragments were concatenated together. Nucleotides 1–50 of the query constitute the end of an exon at a donor and nucleotides 50–100 constitute the start of an exon at an acceptor junction. Splicing concatenates these two exons together. This 100 nt length sequence was searched for similarity with EST sequences in EMBL using FASTA. The objective of this search is to identify those splice sites of the query sequence for which a match to at least one human EST exists (ideally along all the 100 nt) and thus to confirm the annotation of both the 5' and 3' splice sites.

Query sequences to confirm donor sites. Three types of query sequence were defined to confirm the donor sites: (i) the **exon_EI** and **intron_EI** fragments at a donor junction were concatenated together to form a query sequence of length 100 nt (labelled **exon_EI-intron_EI** sequence); (ii) the **exon_EI** fragment of a donor junction formed a query sequence; (iii) the **intron_EI** fragment of a donor junction formed a query sequence. The resulting EST hits of the FASTA searches against EMBL using these three types of query sequences were analysed.

Query sequences to confirm acceptor sites. Three types of query sequence were defined to confirm the acceptor sites: (i) the **intron_IE** and **exon_IE** fragments at an acceptor site were concatenated together to form a sequence (labelled **intron_IE-exon_IE)** of 100 nt length; (ii) the **exon_IE** fragment of an acceptor site formed a query sequence; (iii) the **intron_IE** fragment of an acceptor site formed a query sequence. The resulting EST hits of the FASTA searches against EMBL using these three types of query sequences were analysed.

Objectives of EST searches and a typical outcome of a search. The objective of searches using **exon_EI-exon_IE** as query sequence was to obtain confirmation using EST sequences for both the 5' and 3' splice sites. The objectives of the searches using the other six types of query sequences were two-fold:

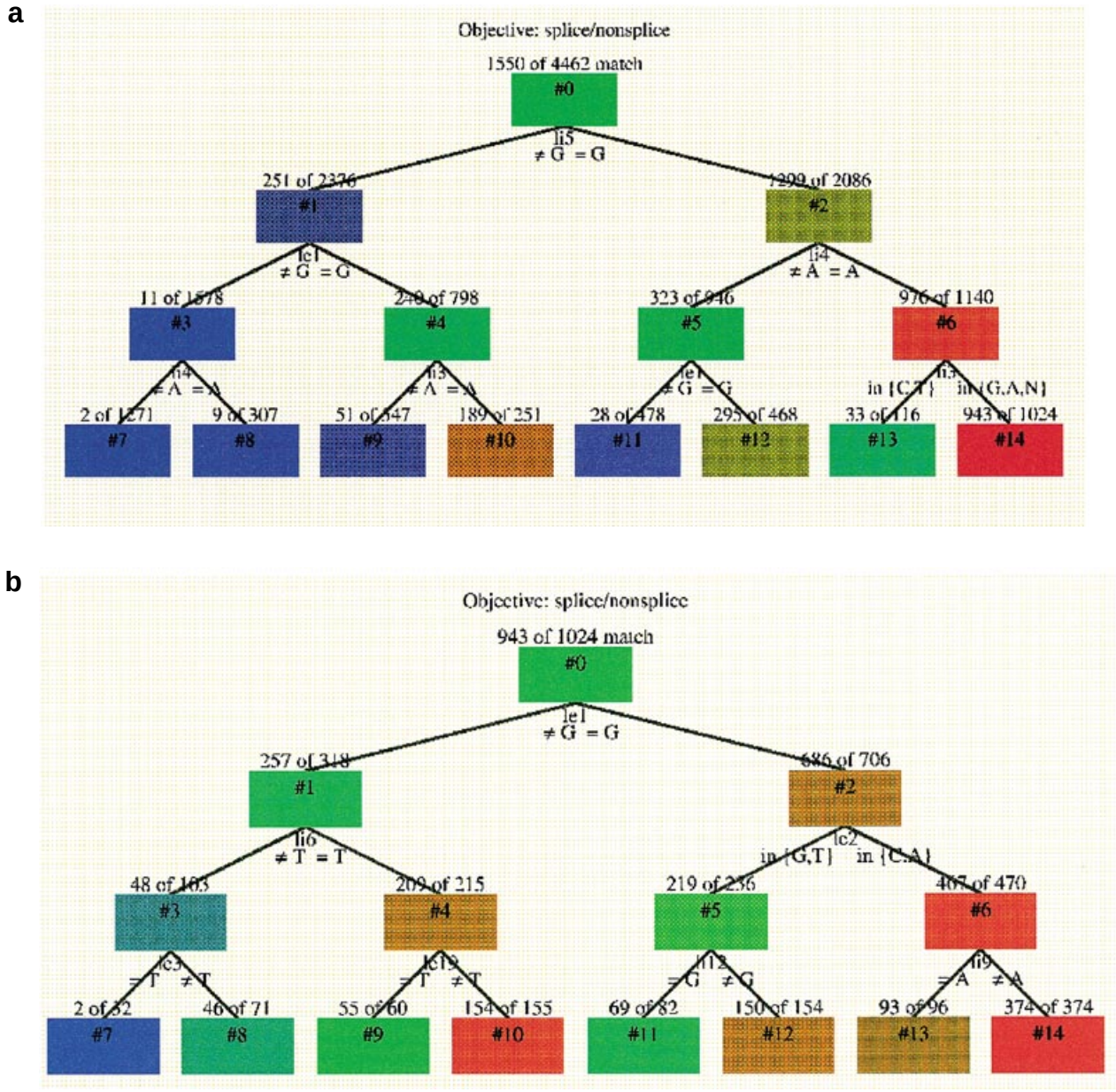

Figure 2. Human exon–intron junctions. (**a**) A four-layer decision tree for human donor junctions. A red node indicates a population rich in real annotated sites; a blue node indicates a population poor in real annotated sites. The two numbers against each node indicate population size of the node and the number of real sites, e.g. the numbers 1550 of 4462 for the root node indicate the population size as 4462 of which 1550 are real annotated sites. (**b**) Further split of a leaf node from the first level four-layer decision tree (from a) for human exon–intron junctions. In this case, node 14 from (a) is split further.

(i) to identify alternative splicing events involving the junctions or their proximal regions; (ii) to confirm the annotated donor (or acceptor) site, irrespective of the confirmation status of the corresponding acceptor (or donor) site.

In a search using **intron_IE-exon_IE** as the query sequence, if a region from an EST sequence shows a match with the **exon_IE** fragment, but the upstream region of EST does not match with the **intron_IE** fragment, then the EST sequence is taken to validate the junction.

In a search using **exon_IE** as the query sequence, if a region from an EST sequence shows a match with the query but the upstream region of EST does not match with the corresponding **intron_IE** fragment, then the EST sequence is taken to validate the junction. If the EST shows a match with only a portion of the **exon_IE** sequence, it indicates that either the annotation of the donor junction was doubtful or alternative splicing events take place in the proximity of the junction.

In a search using the **intron_IE** sequence as the query, if a region from an EST sequence shows a match with the query

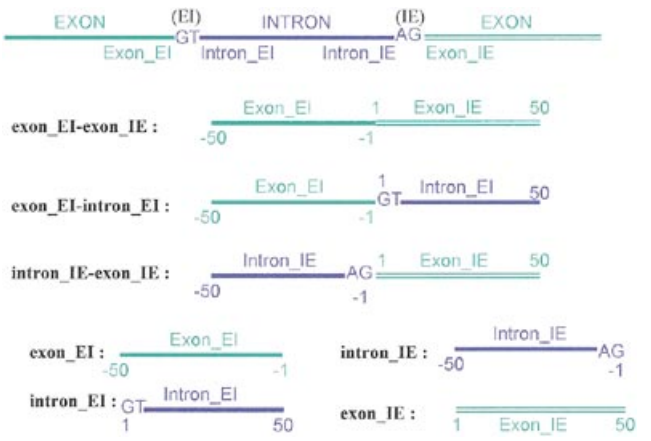

Figure 3. Query constructs involving splice sites for searching EST sequences. Seven types of query constructs. Query order for the first three constructs is 1–100 and for the last four constructs is 1–50 nt positions.

and the downstream region of the EST also matches with the corresponding **exon_EI** fragment, then it indicates that either the junction is doubtful or the annotated junction undergoes alternative splicing events.

A similar procedure was applied to searches pertaining to donor site.

FASTA searches and analysis of the EST hits. We performed FASTA searches (gap penalties = –12/–2, ktup = 6 and *E* value ≤ 1) against EST sequences for all seven types of query sequence defined. An initial analysis of the EST hits as identified by the FASTA searches was performed for every query sequence as described below. It is our observation that hits with a percentage identity (pid) value of <80 were often erroneous matches. EST hits with a pid value in the range 80–90 required careful scrutiny.

For hits with low pid values, the mismatches occurred mostly towards the ends of the sequence and the effective length of the alignment was reduced to lower values in such queries. FASTA had usually reported the same hits in other searches with a lower length of alignment but with a higher pid value. Hence, the effective length of alignment was evaluated in each case.

EST hits with alignment lengths in the range of 20 nt were considered to have occurred by chance and were ignored. Certain intronic and exonic sequences had biased base compositions and in these cases, only those hits with a pid value in the range 95–100 were considered.

For hits of lower pid value, it was checked whether the mismatches were due to undetermined nucleotides in the EST and if the few mismatches were not concentrated at one position in the alignment. Consensus nucleotides as displayed by other EST hits for the same query were used to resolve the nucleotide at the mismatching positions.

RESULTS

The types of EST hits obtained for the seven types of searches are shown in Figure 4a–g. They are discussed below.

Results of searches pertaining to both the 5' and 3' splice sites

The distribution of **exon_EI-exon_IE** query sequences having at least one EST hit for different values of pid is shown in Table 1a. Roughly 50% of the query sequences tested for similarity with EST sequences showed a pid value of 100 and ~70% of queries showed a pid value of ≥80. As the requirement on pid value reduces, more query sequences showed similarity with EST sequences. It should be mentioned here that ~10% of the query sequences did not show a match with any EST sequence.

We considered only those query sequences that had EST hits with a pid value of >80, of which there were 661 such entries (Table 1b). Each one of these 661 entries was scrutinised against the following criteria: no mismatches or gaps occurred in the vicinity (–20 to +20 nt) of the junctions and the effective alignment length extended to at least a range of –30 to +30 nt. Shortened alignments occurred when either the EST sequence was incomplete (types *b* and *c* in Fig. 4a), there was a potential IE junction in the **exon_EI** fragment (types *d* and *f*) or a potential EI junction in the **exon_IE** fragment (types *d* and *e*). The most common form of EST hit was type *a* (Fig. 4a) wherein both the **exon_EI** and **exon_IE** fragments matched with the EST. When the EST sequence was incomplete, the presence of AG (type *b*) or GT (type *c*) dinucleotides was checked for in the query sequence at the corresponding end positions and it was ensured that the short alignments were not a result of the presence of splice junctions other than those annotated.

Careful analysis of the results described above identified 610 annotated splice sites (out of 925) as being correct at both the constituent donor and acceptor junctions and these sites could be characterised as shown in Table 1b. Except in 10 cases, all 50 of the nucleotides on either side of the intron that is spliced out are coding.

Results of searches pertaining to the donor junction

The types of EST hits against the three types of query sequences (pertaining to donor junction) are as shown in Figure 4b–d.

The results of FASTA searches for each of these search types were scrutinised to determine those queries that showed a match with at least one EST with a pid value in the range 80–100 and of any alignment length. The search using the **exon_EI-intron_EI** sequences reported 574 such entries (with 4675 EST hits). The corresponding result in the case of the search with the **exon_EI** sequences was 682 entries (with 6864 EST hits) and for **intron_EI** sequences was 142 entries (with 317 EST hits). This identified 695 donor junctions. The FASTA results for each of these cases were examined. Having already identified the splice sites that had proof from EST sequences at both the constituent donor and acceptor junctions, the emphasis now was to confirm the results, to identify those donor junctions which were missed (because they were overlooked or the corresponding acceptor sites were wrongly annotated) and to identify potential alternatives involving the donor junction.

Of the 695 donor junctions, 659 met the criteria described in Materials and Methods and could be classified as described below.

Direct_EI junctions, i.e. donor junctions for which at least one EST hit existed for a match to the **exon_EI** fragment and

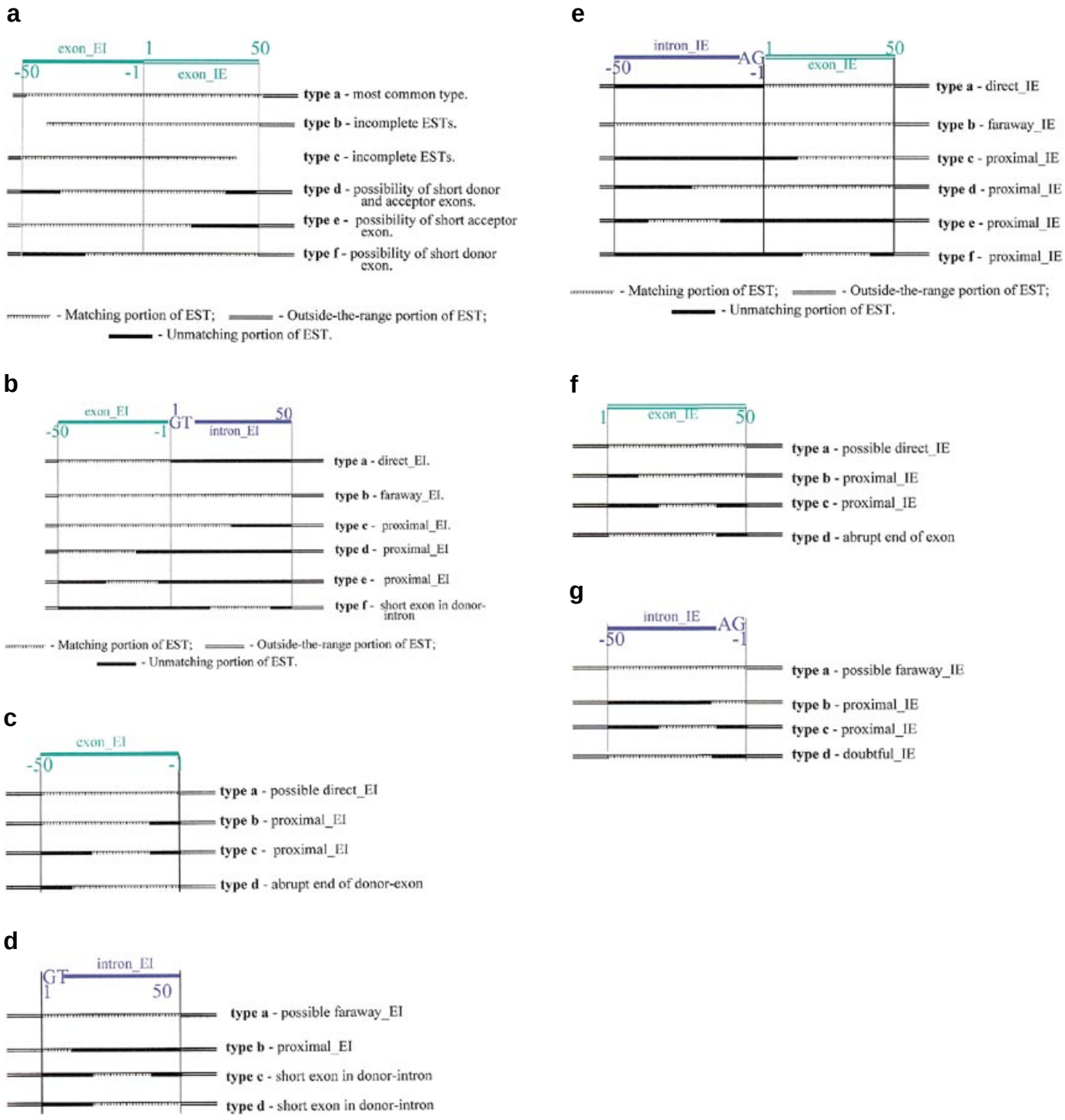

Figure 4. Types of EST hits obtained for the different types of search. (**a**) EST hits for **exon_EI-exon_IE** query sequences. (**b**) EST hits for **exon_EI-intron_EI** query sequences. (**c**) EST hits for **exon_EI** query sequences. (**d**) EST hits for **intron_EI** query sequences. (**e**) EST hits for **intron_IE-exon_IE** query sequences. (**f**) EST hits for **exon_IE** query sequences. (**g**) EST hits for **intron_IE** query sequences.

the downstream region of the EST not matching the **intron_EI** fragment (type *a* in Fig. 4b and c).

A **Direct_EI** junction was reclassified as a **Genuine_EI** junction when the downstream region on the EST matched the corresponding **exon_IE** fragment (and the search using the corresponding **exon_EI-exon_IE** sequence also identified the same EST hit).

If for a **Genuine_EI** junction there existed at least one other EST hit where the downstream region of the EST matched with the **intron_EI** fragment, then the junction was reclassified as a

Table 1a. Distribution of **exon_EI-exon_IE** query sequences having at least one EST hit for different pid values

pid value	No. of query sequences	No. of matching EST sequences
100	430	2579
≥95	591	5168
≥90	619	5776
≥80	661	6656
≥70	737	7912
≥60	800	9561

Table 1b. Results of FASTA search using **exon_EI-exon_IE** sequences against EST sequences

Effective length of alignment	No. of entries	Percentage identity as reported by FASTA
100	502	100 (in the case of 352 queries)
		95–100 (in the case of 133 queries)
		90–95 (in the case of 17 entries)
95–99	49	>90 (in the case of 38 queries)
		>85 (in the case of 11 queries)
90–94	22	>90 (in the case of 19 queries)
		>85 (in the case of 3 queries)
75–90	27[a]	>90
75–90	10[b]	>90

[a]EST sequences were incomplete in these cases.
[b]Alignment did not cover the complete length of these query sequences.

FarawayAlternative_EI junction (type *b* in Fig. 4b; type *a* in Fig. 4d).

If the match of the EST with the **intron_EI** fragment did not cover the complete region of 50 nt (i.e. the alternative functional donor junction was within the 50 nt range of the **intron_EI** fragment), then the junction was reclassified as a **ProximalAlternative_EI** junction (type *c* in Fig. 4b; type *b* in Fig. 4d). If an EST hit exhibited a match with only a portion of the **exon_EI** fragment (types *d* and *e* in Fig. 4b; types *b* and *c* in Fig. 4c), then there was probably an alternative functional donor junction upstream of the annotated one and such a junction was also classified as **ProximalAlternative_EI**.

There were instances of EST hits where an EST matched with a short region in the **intron_EI** fragment (type *f* in Fig. 4b and types *c* and *d* in Fig. 4d) and if the junction was otherwise genuine, then the site was classified as **GenuineEI_WithShort-ExonInDonorIntron**.

If the above-mentioned alternative events were observed for a junction that did not have proof as a **Direct_EI** junction, then the site was classified as a **FarawayAndWrong_EI** junction.

If an EST hit exhibited a match to the **exon_EI** fragment but the downstream region of the EST pointed to a nucleotide region other than the **exon_IE** fragment, then the junction was classified as a **Direct_EI_DifferentAcceptorExon** junction or as a **Genuine_EI_AlternativeInAcceptorExon** junction (if the junction was otherwise a **Genuine_EI**).

If a **Genuine_EI** junction exhibited more than one type of alternative event, then it was classified as a **Multiple-Alternative_EI** junction.

The junctions that could not be classified into any of the above categories were classified as **Doubtful_EI** junctions.

The results of the search using **exon_EI-exon_IE** sequences had earlier identified a set of 10 splice sites (see Table 1b and Fig. 4a, types *d–f*, and Fig. 4c, type *d*) where the alignment did not cover the complete length of the query sequences. Many of these donor junctions are classified as **Doubtful_EI** junctions.

Terminal_EI Junctions. The donor junctions checked thus far had the corresponding acceptor junctions available in the EMBL entries. There were 13 terminal donor junctions (for which we could get the 50 nt regions on either side) that had not been checked so far. Upon checking for EST validation, four of them had confirmation from EST sequences. These four were categorised as **Terminal_EI** junctions.

Donor junctions with EST confirmation. The results of the searches pertaining to the donor junctions are shown in Table 2. The 619 donor junctions that had validation from EST matches were categorised into eight types. Donor junctions from four categories (4–7 in Table 2) involving alternative events may be treated as a separate group because the sites may carry multiple codes. The 50 nt regions around the junctions from categories 1–3, 5 and 8 (see Table 2) could be used for

generating false donor sites. The other categories could not be used for generating false sites because the 50 nt regions would carry EST-confirmed alternative donor junctions.

Table 2. Categorisation of the EST-confirmed donor junctions

Category	Type of Junction[a]	No. of entries
1	Genuine_EI	463
2	Genuine_EI_AlternativeInAcceptorExon	20
3	Direct_EI_DifferentAcceptorExon	14
4	Genuine_EI_WithShortExonInDonorIntron	13
5	FarawayAlternative_EI	68
6	ProximalAlternative_EI	29
7	MultipleAlternative_EI	8
8	Terminal_EI	4
	Total number of EST-confirmed donor sites	619
	WrongAndFaraway_EI	23
	Doubtful_EI	17

[a]Fifty nucleotide regions around the junctions from categories 1–3, 5 and 8 could be used to generate false donor sites.

Results of searches pertaining to the acceptor site

The results of the FASTA searches with each of the three types of query sequences (pertaining to the acceptor site) were analysed to select those entries for which at least one EST hit was reported with a pid score in the range 80–100. The search with **intron_IE-exon_IE** sequences reported 601 entries (with 4986 EST hits). The corresponding figures in the cases of searches using **exon_IE** and **intron_IE** sequences were 689 (with 6995 EST hits) and 145 (with 359 EST hits), respectively. Together these entries corresponded to a total of 701 acceptor junctions. In each of these cases, the EST hits were scrutinised using the criteria described in Materials and Methods. The acceptor junctions were classified in a manner similar to that for the donor junctions.

An acceptor junction for which the **exon_IE** fragment matched with a region of the EST sequence and the immediate upstream region of the EST did not match with the **intron_IE** fragment was classified as a **Direct_IE** (type *a* in Fig. 4e and f) junction.

If the **Direct_IE** had also been confirmed earlier by the search using the **exon_EI-exon_IE** sequences with the same EST hit (i.e. if the immediate upstream region in the EST matched with the corresponding **exon_EI** fragment) then the junction was reclassified as a **Genuine_IE** junction; otherwise the junction was renamed as a **Direct_IE_DifferentDonorExon.**

If for a **Genuine_IE** junction, there existed another EST hit where the immediate upstream region corresponded to a nucleotide region other than the **exon_EI** fragment, then such a junction was renamed as a **Genuine_IE_AlternativeInDonorExon** junction.

If for an annotated acceptor junction, the immediate upstream region in the EST sequence showed a match with the **intron_IE** fragment (types *b* and *d* in Fig. 4e or types *a* and *b* in Fig. 4g), it indicates that the junction is non-functional in the particular transcript from which the EST was derived. If the junction had other EST hits making it a **Genuine_IE** junction, then the junction was implicated as being involved in alternative splicing events with the **intron_IE** fragment (the complete or a portion of the 50 nt) acting as the exonic region in alternative transcripts. Examination of the length of alignment ascertained whether the alternative acceptor junction occurred upstream of the complete **intron_IE** fragment (such an event was named a **FarawayAlternative_IE** junction) or within the **intron_IE** fragment (such an event was named a **ProximalAlternative_IE** junction). If the junction was otherwise not a **Direct** or **Genuine_IE** junction, then the junction was considered as not proven by EST matches and was classified as a **WrongAndFaraway_IE** junction. An EST hit showing a match with only a portion of **intron_IE** (type *e* in Fig. 4e and type *c* in Fig. 4g) indicated a proximal functional acceptor junction; if the junction possessed other hits indicating a **Genuine_IE** junction, then such a junction was classified as a **ProximalAlternative_IE** junction; otherwise it was assumed that the annotation was wrong and the site was classified as a **WrongAndFaraway_IE** junction. There were instances when EST sequences showed matches to portions in **exon_IE** fragments (types *c* and *f* in Fig. 4e and types *b* and *c* in Fig. 4f). Such instances pointed to the possibilities of the occurrence of short exons and hence the possibility of additional acceptor junctions. If such a junction was also confirmed to be a **Direct_IE** junction by other EST hits, then the junction was also classified as a **ProximalAlternative_IE** junction. Finally, there were cases that could not be resolved unambiguously into any of the above types and they were classified as **Doubtful_IE** junctions.

Terminal_IE Junctions. There were 16 terminal acceptor junctions that have not been considered so far. Upon checking for EST validation, it was found that 11 of them could be validated as correct acceptor junctions.

Acceptor junctions with EST confirmation. The results of the searches are shown in Table 3. A total of 623 acceptor junctions could be validated and were categorised onto six groups. Junctions from categories 2, 4 and 5 may be treated separately because the sites may carry multiple codes. Except for **ProximalAlternative_IE** junctions (50 nt regions around such junctions would carry EST-confirmed acceptor sites), all other junctions could be used for generating false acceptor sites by considering the –50 to +50 nt regions.

Combining the results to generate the data set on splice sites

Depending on the types of the constituent donor and acceptor sites, the splice sites were categorised into six types (Table 4). These are: (i) **Genuine**, where both the constituent donor and acceptor junctions had proof from EST confirmation and were not involved in any alternative splice events; (ii) **Genuine-Alternative**, where both the donor and acceptor junctions had proof from EST confirmation but either or both of them were involved in alternative splice events; (iii) **Direct EI**, where only the donor junction had EST confirmation; (iv) **Direct IE**, where only the acceptor junction had EST confirmation; (v) **Terminal EI**; (vi) **Terminal IE** where only the donor or acceptor junction was available for examination.

Table 3. Categorisation of the EST-confirmed acceptor junctions

Category	Type of junction[a]	No. of entries
1	Genuine_IE	457
2	Genuine_IE_AlternativeInDonorExon	23
3	Direct_IE_DifferentDonorExon	11
4	FarawayAlternative_IE	71
5	ProximalAlternative_IE	50
6	Terminal _IE	11
	Total number of EST-confirmed acceptor sites	623
	WrongAndFaraway_IE	9
	Doubtful_IE	33

[a]Fifty nucleotide regions around the junctions from categories 1–4 and 6 could be used to generate false acceptor sites.

Table 4. Categorisation of the EST-confirmed splice sites

Category	Type of junction	No. of entries
1	Genuine	392
2	Genuine-Alternative	209
3	Direct EI	14
4	Direct IE	11
5	Terminal EI	4
6	Terminal IE	11

DISCUSSION

In this work, we have created a standard data set of human splice sites through exhaustive cleaning procedures such as requiring consistency in the annotation and completeness of the coding regions together with elimination of redundant sequences and identification of ambiguous sites which are prone to alternative splicing. The latter was carried out with the use of decision trees applied to a combination of annotated splice sites and generated false sites. The splice sites were further confirmed by comparison with SWISS-PROT and EST data. In the following sections, we discuss the characteristics of the data set.

Contents of the data set

The contents of the data set can be interpreted in the following ways.

(i) The data set contains a set of 619 donor sites confirmed by EST matches. In 463 sites, the donor site always paired with the annotated acceptor site only. In 88 sites, the donor site paired with alternative acceptor sites in addition to the annotated acceptor sites. In 14 cases, the donor site paired with acceptor sites other than the annotated ones. In 50 cases, alternative functional donor sites were observed in the –50 to +50 nt region around the donor sites.

(ii) The data set contains a set of 623 acceptor sites confirmed by EST matches. In 457 of these sites, the acceptor site always paired with the annotated donor site only. In 94 sites, the acceptor site paired with alternative donor sites in addition to the annotated donor site. In 11 cases, the acceptor site paired with donor sites other than the annotated ones. In 50 cases, alternative functional acceptor sites were observed in the –50 to +50 nt region around the donor sites.

(iii) A set of 601 splice sites where both the constituent donor and acceptor sites had proof from EST confirmation and could be categorised unambiguously. Of these, 392 are genuine sites wherein both the constituent donor and acceptor sites were not involved in any alternative splicing events. Such a set is useful to study the interactions between the donor and acceptor sites.

(iv) The data set contains 44 donor and 50 acceptor junctions located in the untranslated regions of genes and the rest from translated regions of the genes.

(v) In the case of 569 donor (categories 1–3, 5 and 8 in Table 2) and 573 acceptor (categories 1–4 and 6 in Table 3) sites, the 50 nt regions on either side of the site could be used to derive either false donor or acceptor sites. Such generated false sites are useful as negative controls when training the gene prediction programs and these false sites have a high probability of being non-functional ones.

Quality of the splice sites in the data set

The different splice sites included in the data set showed a varying number of matching EST sequences. The variation depends upon factors such as the type of the splice junction (one that is involved in alternative splicing will have more EST hits), the expression level of the gene (highly expressed genes may be over-represented in the EST database) and presence or absence of poly(A) tails [genes that lack poly(A) tails may be under-represented in the EST database]. However, it is useful to annotate every splice junction with a confidence value that is derived using the number of observed EST hits with a suitable correction.

The overall average number of unique EST hits for a donor or acceptor site in the data set was found to be 17, while the average value for a junction involved in alternative processing was 25 (higher than the overall average) and that of a junction not involved in alternative processing was 15. The correction for the differential expression of genes was done by assigning a maximum value of 17 for the number of unique EST sequences. The confidence value for a junction was calculated as [number of unique EST sequences/average value for this type of junction]. Values of more than 1.0 were truncated to 1.0. It was found that in the case of both donor and acceptor junctions, 75% of the junctions had a confidence value greater than 0.50 and 60% of the junctions had a confidence value of 1.0.

It must be emphasised that very stringent criteria (such as high percentage identity, low *E* value and not allowing any mismatches in the –20 to +20 nt region around the junction) were used to identify the EST hits and four different types of searches with the EST database were made before a splice junction was confirmed. As a result, even if one EST hit was observed for a splice junction, such a splice junction is to be considered as significant.

Comprehensiveness of the validation procedures

Though the initial data cleaning made sure that the EMBL annotation did not mention any alternative splicing in the sequence entries, the final results identified alternative splicing

in some sequence entries. A discussion on how comprehensive is the validation from EST sequences is in order here. It is a common notion that while the presence of an EST hit supports an event, absence of an EST hit cannot be taken as a proof that the event does not occur. Such an event in our study is the involvement of the splice sites (that were identified as **Genuine_Splice** sites) in alternative splicing. Every splice site that was analysed had on average as many as 17 EST hits and hence we can regard the conclusions with a high level of confidence.

Only 67% of splice sites could be validated with EST sequences and 4% turned out to be annotated wrongly. Of the remaining sites, 8% did not have a match with any EST sequence and 22% did not have significant matches with any EST sequences. The reasons that almost a third of the junctions did not get support from EST sequences might be due to the EST database not being completely representative of all the genes in an organism and the inherently high rate of sequencing errors in EST sequences.

A discussion about **Faraway_Alternative** sites (Tables 2 and 3), which applies to certain annotated intronic sequences as well as exonic sequences, is in order here. Wolfsberg and Landsman (10) noted that there are a significant number of intron sequences reported in EST sequences and thus they are useful to analyse the splicing pattern of the cDNA sequences characterised previously. They further remarked that there is a small possibility of contamination in cytoplasmic mRNA by genomic DNA or partially spliced nuclear hnRNA. However, the alternatively spliced transcripts of the **Faraway_Alternative** sites included in our data set may represent unreported novel transcripts. We support such a hypothesis by: (i) the observation that at least 30% of human genes are spliced alternatively (2); (ii) our observation that such an alternative junction showed an average of 25 EST hits (50% more than that for a splice junction of any type); (iii) these sites were able to be resolved unambiguously. Such sites that could not be resolved in a similar manner were eliminated as **FarawayAndWrong** and **Doubtful** junctions (Tables 2 and 3).

How representative is the data set

EST sequences are usually sequenced from cDNA sequences (derived from the mRNA) using the poly(A) tail of the mRNA as primers. As a result, genes whose mRNA does not possess a poly(A) tail would be under-represented in the data set. In addition, because of the cleaning procedures, genes with non-standard start and stop codons as well as those with non-standard splice junctions (i.e. those that do not have GT…AG as the consensus dinucleotides at the splice junctions) are not represented in the data set. Moreover, since the EST database is biased towards highly expressed genes, genes of low expression levels would be under-represented in this data set.

Because of the above-mentioned method of deriving EST sequences, it is important to examine whether the splice sites from the 5'-end of the genes are disfavoured in the data set. An examination of human EST sequences in the database revealed that the 3' EST sequences outnumbered 5' EST sequences by only 11%. Analysis of the locations of the splice sites along the genes from the data set revealed that a total of 580 (336 donor plus 244 acceptor) splice junctions originated from the 5'-end of the genes. In contrast, 662 (283 donor plus 379 acceptor) splice junctions originated from the 3'-end of the genes. The difference between the 5'- and 3'-ends was only 14%. However, these values as percentages of annotated splice junctions from the respective ends of the genes (the 5'-end had 443 donor and 330 acceptor junctions; the 3'-end had 364 donor and 480 acceptor junctions); the values were 75% for the 5'-end and 78% for the 3'-end. Thus the distribution of EST-confirmed splice sites in the 5'- or 3'-end of the genes reflected the pattern observed for the genes.

Utility of the data set

In summary, the data set classifies splice sites as donors, acceptors, genuine and alternative splice sites. Such a data set has considerable utility both for its own intrinsic interest and for its applicability to techniques such as computational gene finding, whose effectiveness is dependent on appropriate training sets. The data set can also be used as a standardised set for comparing the performance of different gene-finding programs. The data set would be useful for several investigations on splice site predictions and alternative splicing. In addition we have set standards for creating high quality data sets on splice sites.

AVAILABILITY OF THE DATA SET

The data set is available on the World Wide Web at http://www.ebi.ac.uk/~thanaraj/splice.html . The data set of the splice sites contains the EMBL entry ID name, the location of the constituent donor and acceptor junctions, the upstream and downstream regions of 70 nt, the type of splice site, the types of the constituent donor and acceptor junctions (the types are as described earlier) and the confidence values for the splice sites. A region of 70 nt on either side of the junctions is provided; such a longer region might be needed to retrieve signals regarding poly(Y) tracts and branch sites. In addition we have provided the nucleotide sequences of genes (along with gene structure annotation) that resulted after the cleaning using decision trees. Regions of 50 nt that could be used to generate false donor and false acceptor sites are also provided. Such sites have a high probability of being non-functional ones.

CONCLUSIONS

This report presents not only a set of direct donor and acceptor sites validated by EST hits, but also sets of sites that are involved in alternative splicing events. Most importantly, we present a set of splice sites in which both the constituent donor and acceptor sites have EST proof and are not involved in any alternative splicing events. The data set presented herein is clean and has experimental proof. It is our hope that such a data set is useful not only for studies to derive signals for either donor or acceptor junctions, but also for interactions between the donor and acceptor sites. We also report nucleotide regions (around the real junctions) which could be used to generate false control junctions. Such sites have a high confidence level of being non-functional.

ACKNOWLEDGEMENTS

The author wishes to thank Alan Robinson for support during the work and careful reading of the manuscript, Jeroen Coppieters for discussions in the initial stages of the project, Alvis Brazma for

discussions on decision trees and Juha Muilu for providing the human EST data in a flat file format as well as for discussions on EST sequences. The author also wishes to thank the referees for useful comments and suggestions.

REFERENCES

1. Burge,C.B. and Karlin,S. (1998) *Curr. Opin. Struct. Biol.*, **8**, 346–354.
2. Mironov,A.A. and Gelfand,M.S. (1998) In *Proceedings of the First International Conference on Bioinformatics of Genome Regulation and Structure*, Vol. 2. BGRS'98, Institute of Cytology and Genetics, Novosibirsk, Russia, pp. 249–250.
3. Burset,M. and Guigo,R. (1996) *Genomics*, **34**, 353–367.
4. Korning,P.G., Hebsgaard,S.M., Rouze,P. and Brunak,S. (1996) *Nucleic Acids Res.*, **24**, 316–320.
5. Stosser,G., Tuli,M.A., Lopez,R. and Sterk,P. (1999) *Nucleic Acids Res.*, **27**, 18–24.
6. Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
7. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.
8. Bairoch,A. and Apweiler,R. (1998) *Nucleic Acids Res.*, **26**, 38–42.
9. Reese,M. and Kulp,D. (1998) http://www-hgc.lbl.gov/inf/genesets.html
10. Wolfsberg,T.G. and Landsman,D. (1997) *Nucleic Acids Res.*, **25**, 1626–1632.