4 Confidence-based prediction of (alternative) splicing

This chapter presents our expressed sequence tag (EST) based approach for prediction and ranking of constitutive as well as alternative splice events (Gupta et al. (2004a)). The naive way to deduce (alternative) splice events from EST data is to map the EST sequences to the genome (Section 3.3.5). However, these predictions also include a large number of potentially non-real instances of splicing. The false predictions are attributed to factors like misalignment caused by bad sequence quality of ESTs and/or genomic contamination of cDNA libraries (Sorek and Safer (2003)). In order to remove these false positives, one requires more stringent approaches when applying EST data for the prediction of splice events.

Alternative Splicing Annotation Project (ASAP) was one of the first attempt of genome-wide prediction of (alternative) splicing based on EST data (Modrek and Lee (2002), Lee et al. (2003), Section 3.3.5). In ASAP the problem of bad sequence quality of EST data is addressed by stringently defining the search parameters e.g. allowing only perfect matches. Additional stringency criterion was applied in ASAP by focusing only on those splice events that are marked by the more frequent *consensus* splice signals. Another large-scale approach used for the construction of Alternative Splicing Database (ASD: Clark and Thanaraj (2002), Thanaraj et al. (2004), Section 3.3.5) relies only on mRNA confirmed splice events. In both the aforementioned databases, the applied criteria implicate the loss of a significant fraction of isoforms for which either the splice signals are non-consensus and/or for which full length mRNAs are not present. Alternatively, we suggest to define a combined measure of stringency so that even those positive evidences which are insufficient alone could lead to a prediction if other evidences are favorable. By definition such a composite score based on various EST-based evidences will also reveal examples of non-consensus splicing if they are represented by a sufficient number of ESTs. In order to compute a composite score, we used a *fuzzy logic* approach.

4.1 Fuzzy logic

The concept of fuzzy logic was conceived by Lotfi Zadeh (Zadeh (1965)). It maps a set of input parameters to certain output parameters based on heuristically defined

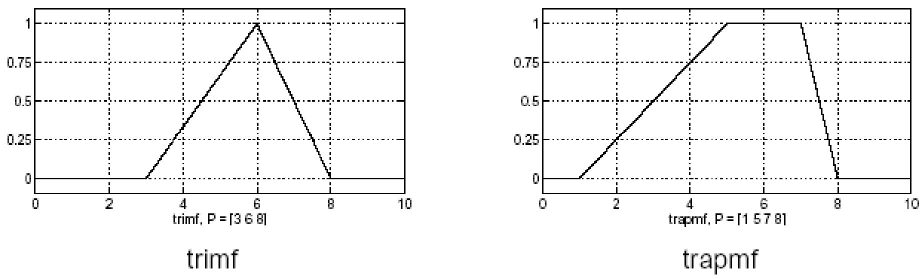


Figure 4.1: Linear membership functions. The figure illustrates the simplest form of membership functions (linear) that map a given input space to a membership value between 0 and 1. The left part of the figure shows a *triangular* membership function while the right part depicts a *trapezoidal* membership function.

rules. The fuzzy logic based methods use heuristic rules to generate flexible as well as robust models. This concept is especially useful for noisy data or for data with possible missing input information. In addition, it is also a time-efficient way of obtaining a robust approximation to an optimal mathematical model which might be too complicated for a fast design. The five steps necessary to implement a fuzzy logic system are described below.

4.1.1 Fuzzification of inputs

The linguistic terms attributing input parameters are translated into membership functions. A membership function is a curve that defines how each point in the input space is mapped to a linguistic term with a certain degree of belief (between 0 and 1). Once the input space is encoded into such a membership function, all subsequent analysis can be performed using linguistic terms which facilitates a heuristic framework for rules. The simplest membership functions are formed using straight lines (Figure 4.1).

4.1.2 Application of fuzzy operators

Fuzzy operators are an extension of standard boolean operators and can be called a superset of standard Boolean operators. Extreme fuzzy values of 1 (absolutely good) and 0 (absolutely bad) will implicate standard boolean logical operations. Fuzzy logic extends these operations for continuous input values, in other words allowing the operations on all real numbers between 0 and 1. Figure 4.2 graphically contrasts the multi-valued logic of fuzzy systems with the two-valued boolean logic.

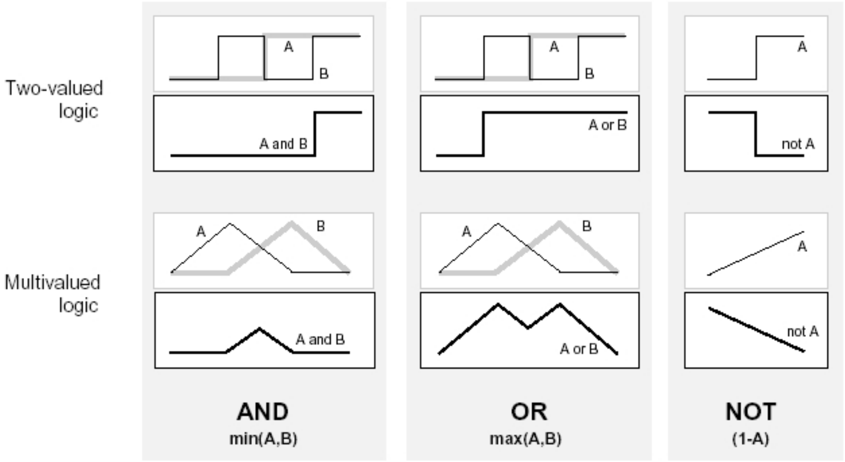


Figure 4.2: Fuzzy Logic Vs Boolean logic. The upper part of the figure displays plots corresponding to the basic logical operations (AND/OR/NOT) on boolean variables. The lower part of the figure shows the application of such these logical operators on continuous variables as used in fuzzy systems. The dark plots correspond to the output after performing the corresponding operations. While the boolean operations give only an *exact yes/no* (0/1) answer, fuzzy definition of variables allows the possibility of a *partial yes.* Image source: `http://www.mathworks.com/access/helpdesk/help/toolbox/fuzzy`

4.1.3 Application of implication method

A core aspect of the fuzzy systems is the *if-then* rule statement (Figure 4.3). While fuzzy variables and logical operators are the verbs of fuzzy logic, if-then rule statements implicate conditionality that comprise fuzzy logic. A single fuzzy if-then rule is of the form
if x is A then y is B
where A and B are linguistic values defined by fuzzy sets. However, it is important to note that every rule has a corresponding weight. The weight of a rule defines its significance in a way that a change in input parameters of a higher weighted rule will affect the output more than if the same change was observed for another input parameter corresponding to a lowly weighted rule. Therefore, this allows to encode the relative significance of different rules with respect to each other.

4.1.4 Aggregation of all outputs

An amalgamation of outputs derived from different rules into a single fuzzy set is achieved by a process called aggregation. The common aggregation methods are *maximum*, *probabilistic* and *sum*.

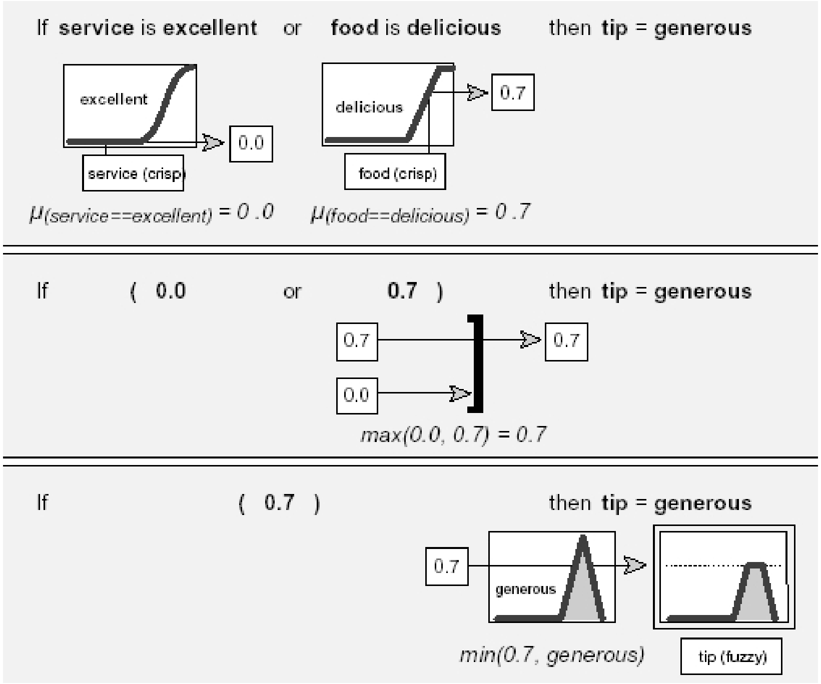

If service is excellent or food is delicious then tip = generous
excellent
0.0
service (crisp)
delicious
0.7
food (crisp)
μ(service==excellent) = 0 .0
μ(food==delicious) = 0 .7
If (0.0 or 0.7) then tip = generous
0.7
0.0
0.7
max(0.0, 0.7) = 0.7
If (0.7) then tip = generous
0.7
generous
min(0.7, generous)
tip (fuzzy)

Figure 4.3: Application of implication method. The diagram illustrates the interpretation of an if-then rule for deciding the amount of tip in a restaurant based on two fuzzified inputs viz., food quality and service quality. For the set of input values (service=3, food=8), the two fuzzy statements are resolved to get the membership values of 0.0 and 0.7 for the service and food statements respectively. Subsequently, the fuzzy operator (OR: *max()*) is used to resolve the multiple membership values (0.0 and 0.7) into a single value. Finally, the membership value is applied to the output membership function to derive the fuzzy output (tip). Image source: `http://www.mathworks.com/access/helpdesk/help/toolbox/fuzzy`

4.1.5 Defuzzification of fuzzy terms

In the final step, the aggregated but still fuzzy output needs to be translated into a discrete numerical value. This process of defuzzification of the output can be performed using different methods like the centroid computation, the average of the maximum value etc.

4.2 Fuzzy logic based prediction of (alternative) splice events

In the absence of a defined model to describe splicing, fuzzy logic provides a good approximation for estimation of the reliability of EST-based predictions of splice events. This section describes a fuzzy-logic approach to compute quality values of splice junctions as well as entire exons. Subsequently, confidence in the predicted alternative splice events is computed based on the quality values of the relevant exon-intron junctions.

4.2.1 Definition of membership functions for splicing evidences

In order to build the fuzzy logic model, different evidences of splice events from the EST data are extracted and encoded into fuzzified parameters. The membership functions of these parameters are described below.

1. *Common Boundaries (GeneNest):* Start/end positions shared by multiple clones in the EST assembly may result from sequencing of incompletely spliced clones. Therefore, these common positions provide an evidence for splice junctions. This parameter is encoded into a triangular membership function with five categories varying from *very low* (VL) to *very high* (VH) (Figure 4.4).

2. *Common Boundaries (SpliceNest):* The exons shared among multiple transcripts are observed as common boundaries in the mapping of EST consensus sequences to the genome. These common boundaries are therefore considered as a positive evidence for the existence of the splice-junction. Figure 4.5 shows the corresponding membership function. The plateau towards the right of membership function implies that after a certain number of common boundaries (n=5), an additional boundary will not contribute to more confidence in the splice junction.

3. *Tolerance for detection of the common boundary (SpliceNest):* A tolerance is allowed while delineating the common boundaries in alignments, such that in the absence of common boundary results the nearest boundary is considered as a common boundary. However in these cases, the tolerance used is penalized with

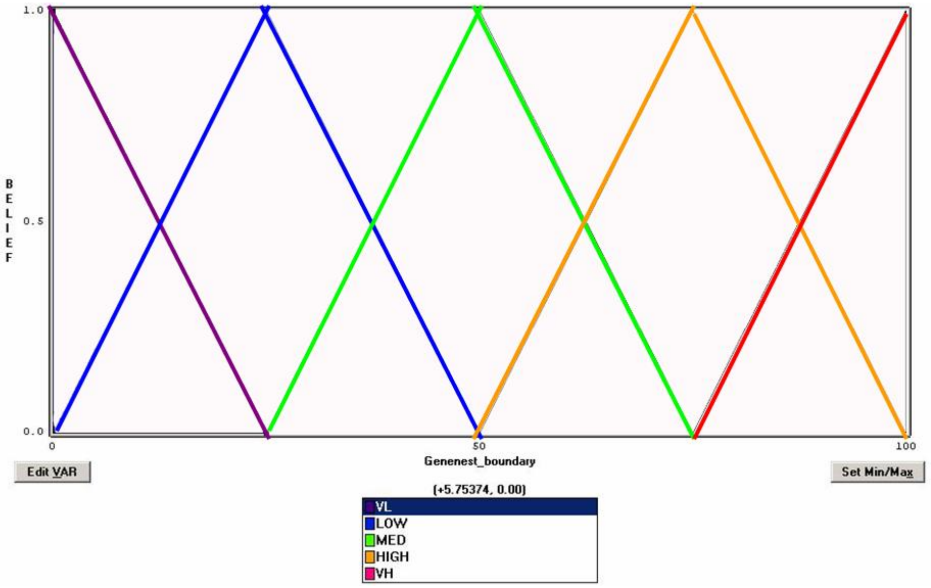


Figure 4.4: Membership function of common boundaries (GeneNest). The input range is categorized into five categories varying from *very low* (VL) to *very high* (VH).

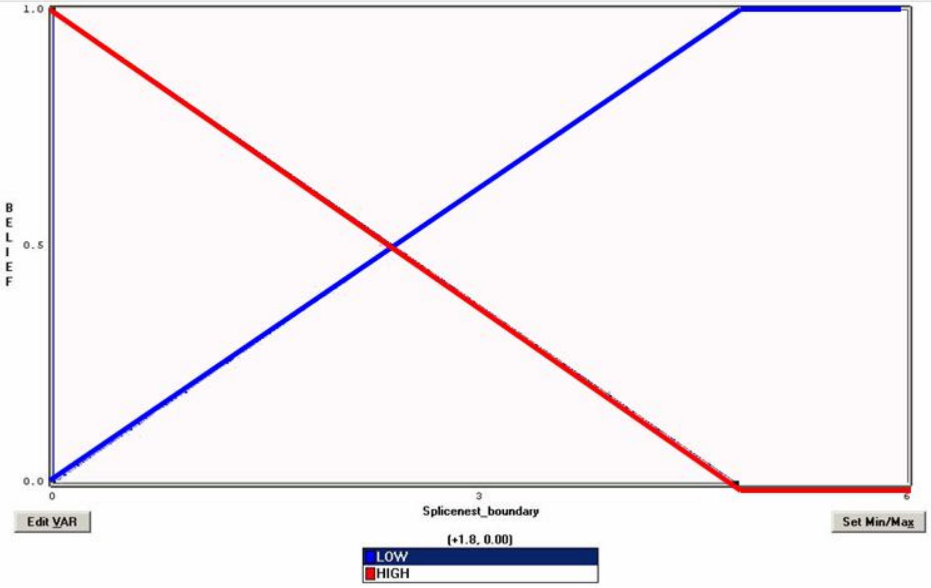


Figure 4.5: Membership function of common boundaries (SpliceNest). The input range is categorized into *low* and *high*. The plateau towards the right of membership function implies that after a certain number of common boundaries (n=5), an additional boundary will not contribute to more confidence in the splice junction.

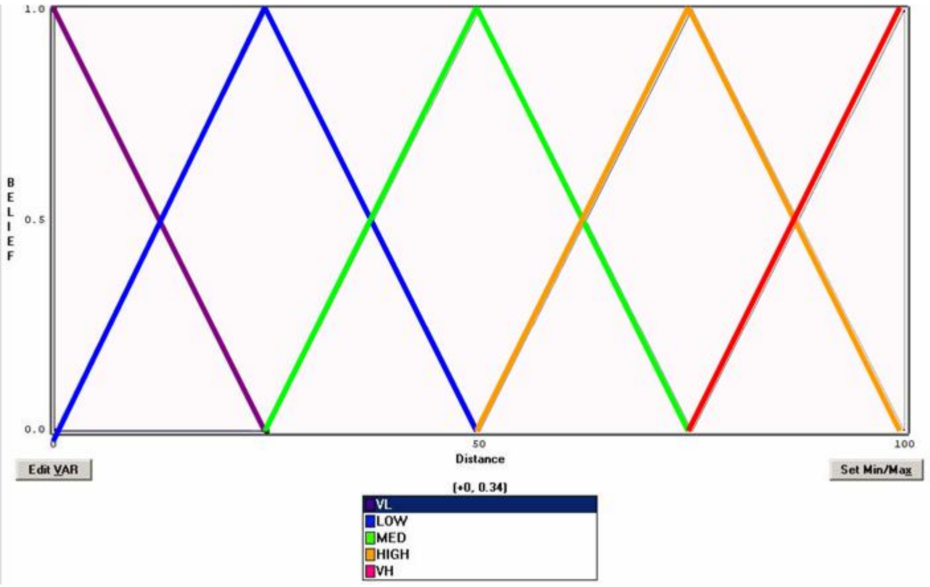


Figure 4.6: Membership function of tolerance used for detection of the common boundary (SpliceNest). The input range is categorized into five categories (*very low* (VL) to *very high* (VH)).

the increasing distance between the two respective boundaries. This parameter is encoded into a triangular membership function (Figure 4.6).

4. *EST count:* The number of ESTs covering a splice junction provide a reliability measure for that particular event. Such a measure is encoded by the membership function illustrated in the figure 4.7. In this case as well the input space consists of five categories. However, an increased categorization at the lower end of the input space implicates a higher sensitivity when the number of ESTs covering the splice event is lower.

5. *Splice signal:* The presence of a splice signal is defined either by the presence of the consensus (GU-AG) splice signal or by the presence of a non-consensus splice signal with perfect alignment in vicinity (+/-10 bases) of the splice junction. Non-consensus splicing events are a known feature of splicing (Clark and Thanaraj (2002)) and therefore should not be left out. However, one needs additional positive hints for such cases, since several splice sites detected with non-consensus splice signal may be artifacts of the alignment procedures. This is due the fact that the alignment program (sim4) introduces gaps/mismatches to extend the alignment as long as the overall alignment quality does not fall below a defined threshold. By keeping a check on local alignment quality such data artifacts are removed. This information is used to estimate the reliability splice signal and is described by a membership function with four discrete classes (Figure 4.8).

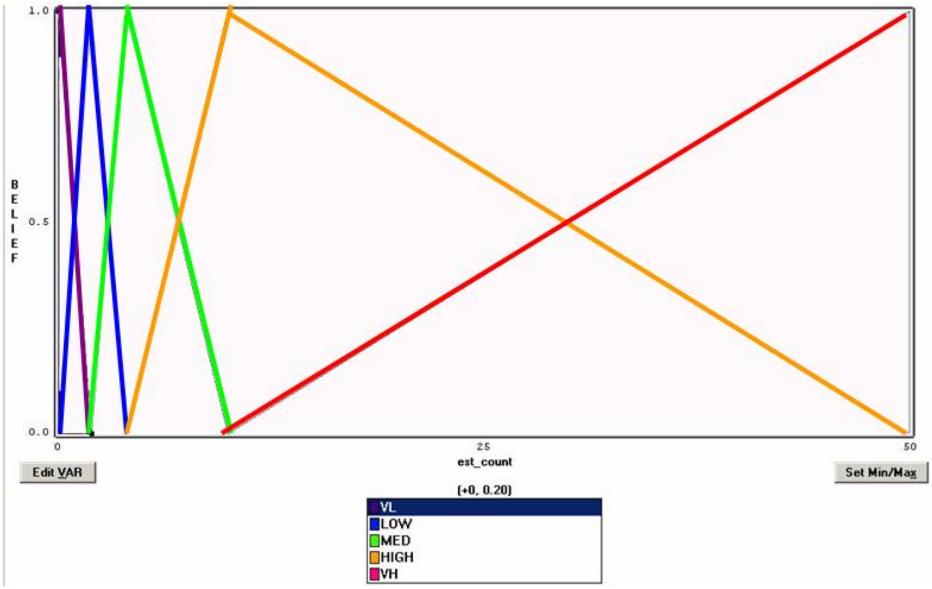


Figure 4.7: Membership function of EST count. The input range is categorized into five categories varying from *very low* (VL) to *very high* (VH). However, an increased categorization at the lower end of of the input space implicates a higher sensitivity when the number of ESTs covering the splice event is lower.

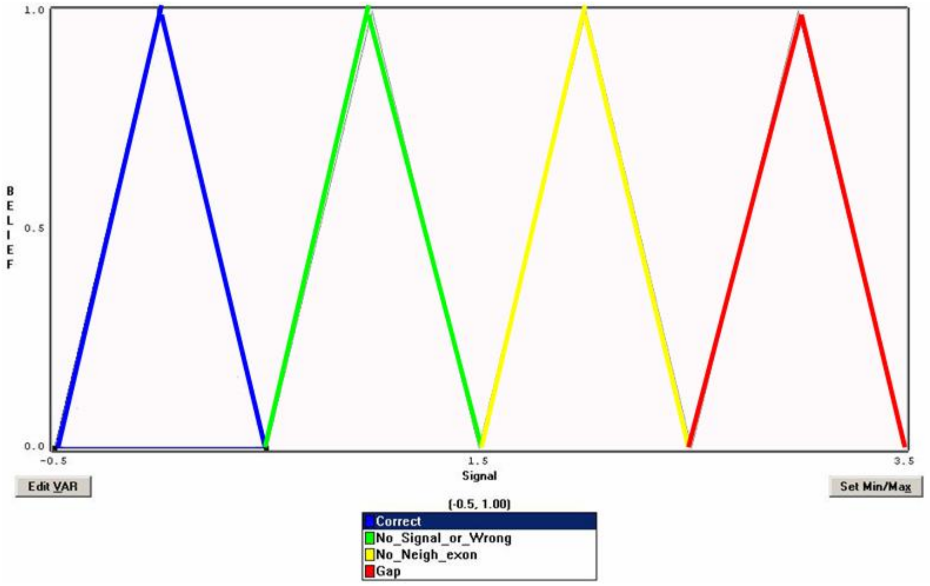


Figure 4.8: Membership function of splice signal. Although by definition it is triangular membership function, the input values will always be integer corresponding to one of the four categories shown. Apart from the wrong signal, there are two more categories defined for flexibly describing different splice events. These are called `no_neigh_exon` and `gap`, representing terminal exon junctions and gaps in the alignment respectively.

6. *Neighborhood signal quality:* An overall splicing direction of the transcript is computed and an inconsistency in direction of one of the splice signals is interpreted as an absence of splice signal for the respective junction. The membership function of this parameter is similar to that of the parameter *splice signal.*

4.2.2 Computation of quality values for exon-intron boundaries

All the parameters described in Section 4.2.1 are computed and a fuzzy logic system (TILShell Pro, Togai Infralogic Ltd) is designed to compute quality values for splice boundaries based on these parameters. Figure 4.9 illustrates the combination of these parameters into a final quality measure. This computation is splitted into two steps. First, the common boundary information together with the EST coverage of the splice junction is combined using a set of rules (Table 4.1: Rulebase 1) to compute an intermediate quality value. This quality measure is then refined (Table 4.1: Rulebase 2) using splice-signal information to compute a final quality measure for splice junctions. This two step procedure facilitates optimization of the extra weight required for rules related to splice signals information which are known markers of splice-sites (Thanaraj and Clark (2001); Weir and Rice (2004)).
Based on these weighted rules, the fuzzy logic software outputs a *C* function, which maps these input parameters to an output quality value for the splice boundary.

Afterwards, the quality of relevant exon-intron boundaries together with the sequence alignment score (Figure 4.10) are used to compute confidence values for all observed constitutive as well as alternative splicing events. Alignment score is also represented by a triangular membership function with five classes ranging from *very low* (VL) to *very high* (VH). The rules used for the computation are listed in Table 4.2. For incorporation into the SpliceNest database, a threshold (55%) for the confidence value was determined below which all observed splice events are considered as potential data artifacts. Therefore, only the splice events for which the computed confidence value is above this threshold are visualized in the SpliceNest database. The complete pipeline for the procedure is graphically illustrated in figure 4.11. For illustration purpose the common boundaries derived from GeneNest as well as SpliceNest are grouped into a single parameter. Similarly, the EST coverage as well as alignment score is grouped into a single parameter, viz. alignment quality.

4.2.3 Evaluation of the splice signal parameter

The effect of high weightage for rules related to splice sites was evaluated using systematic computer simulations. For every simulation all parameters, barring the

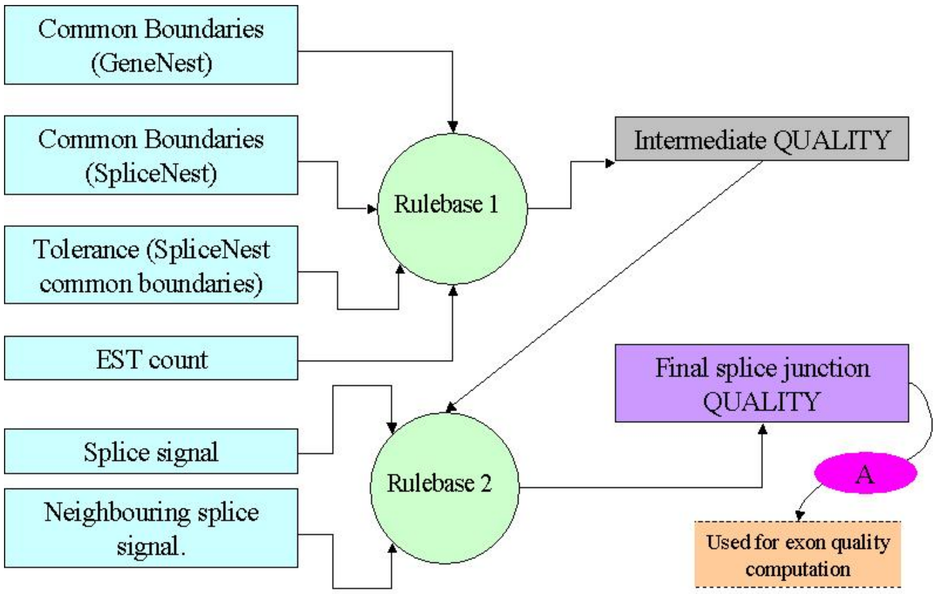

Common Boundaries (GeneNest)
Common Boundaries (SpliceNest)
Tolerance (SpliceNest common boundaries)
EST count
Rulebase 1
Intermediate QUALITY
Splice signal
Neighbouring splice signal.
Rulebase 2
Final splice junction QUALITY
A
Used for exon quality computation

Figure 4.9: Quality values for exon/intron boundaries. The computation of quality values for splice junction is splitted into two steps. First the common boundary information together with the EST coverage of the splice junction is combined using a set of rules (Rulebase 1: Table 4.1) to compute an intermediate quality value. This quality measure is then refined (Rulebase 2: Table 4.1) using splice-signal information to compute a final quality measure for splice junctions. This splice junction quality is used for subsequent computation of exon quality values (connector A).

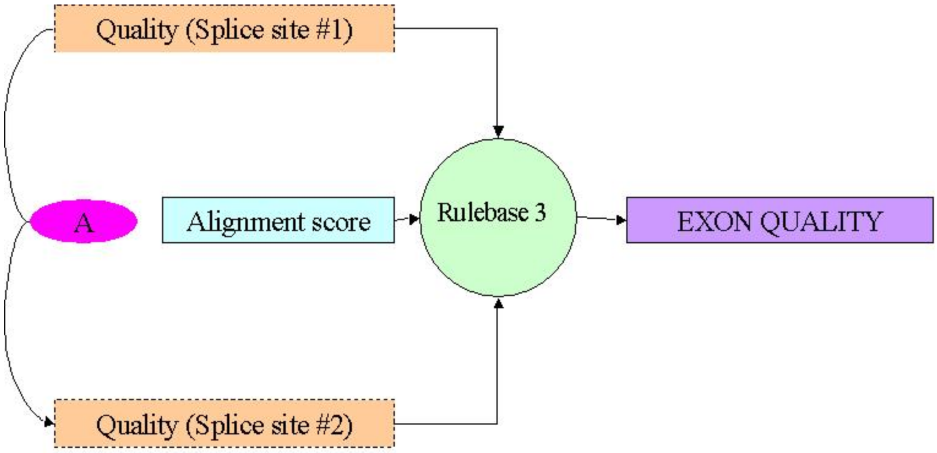


Figure 4.10: Computation of exon quality. The exon quality is computed by combining the sim4 alignment score as well as quality values of the relevant exon-intron boundaries (connector 'A' from Figure 4.9). The fuzzy logic rules governing the computation are listed in Table 4.2.

Rule#	Rule	Weight
Rulebase 1		
1	IF Genenest_boundary IS VL THEN quality (intermediate) IS VL	100
2	IF Genenest_boundary IS LOW THEN quality (intermediate) IS LOW	100
3	IF Genenest_boundary IS MED THEN quality (intermediate) IS MED	100
4	IF Genenest_boundary IS HIGH THEN quality (intermediate) IS HIGH	100
5	IF Genenest_boundary IS VH THEN quality (intermediate) IS VH	100
6	IF est_count IS VL THEN quality (intermediate) IS VL	100
7	IF est_count IS LOW THEN quality (intermediate) IS LOW	100
8	IF est_count IS MED THEN quality (intermediate) IS MED	100
9	IF est_count IS HIGH THEN quality (intermediate) IS HIGH	100
10	IF est_count IS VH THEN quality (intermediate) IS VH	100
11	IF Tolerance IS VL THEN quality (intermediate) IS VH	100
12	IF Tolerance IS LOW THEN quality (intermediate) IS HIGH	100
13	IF Tolerance IS MED THEN quality (intermediate) IS MED	100
14	IF Tolerance IS HIGH THEN quality (intermediate) IS LOW	100
15	IF Tolerance IS VH THEN quality (intermediate) IS VL	100
16	IF Splicenest_boundary IS LOW THEN quality (intermediate) IS LOW	100
17	IF Splicenest_boundary IS HIGH THEN quality (intermediate) IS HIGH	100
Rulebase 2		
18	IF quality (intermediate) IS VL THEN final_quality IS VL	10
19	IF quality (intermediate) IS LOW THEN final_quality IS LOW	10
20	IF quality (intermediate) IS MED THEN final_quality IS MED	10
21	IF quality (intermediate) IS HIGH THEN final_quality IS HIGH	10
22	IF quality (intermediate) VH THEN final_quality IS VH	10
23	IF Signal IS Correct THEN final_quality IS VH	30
24	IF Signal IS No_Signal_or_Wrong THEN final_quality IS LOW	30
25	IF Signal IS No_Neigh_exon THEN final_quality IS MED	30
26	IF Signal IS Gap THEN final_quality IS VL	30
27	IF Neighboring_Signal IS Correct THEN final_quality IS VH	10
28	IF Neighboring_Signal IS No_Signal_or_Wrong THEN final_quality IS LOW	10
29	IF Neighboring_Signal IS No_Neigh_exon THEN final_quality IS MED	10
30	IF Neighboring_Signal IS Gap THEN final_quality IS VL	10

Table 4.1: Rules for computing quality values for splice boundaries. The list corresponds to the two rulebases in figure 4.9. The first 17 rules correspond to rulebase 1 and the remaining correspond to rulebase 2.

Rule#	Rule	Weight
1	IF Quality_Boundary_1 IS VL THEN confidence (splice event) IS VL	100
2	IF Quality_Boundary_1 IS LOW THEN confidence (splice event) IS LOW	100
3	IF Quality_Boundary_1 IS MED THEN confidence (splice event) IS MED	100
4	IF Quality_Boundary_1 IS HIGH THEN confidence (splice event) IS HIGH	100
5	IF Quality_Boundary_1 IS VH THEN confidence (splice event) IS VH	100
6	IF Quality_Boundary_2 IS VL THEN confidence (splice event) IS VL	100
7	IF Quality_Boundary_2 IS LOW THEN confidence (splice event) IS LOW	100
8	IF Quality_Boundary_2 IS MED THEN confidence (splice event) IS MED	100
8	IF Quality_Boundary_2 IS HIGH THEN confidence (splice event) IS HIGH	100
10	IF Quality_Boundary_2 IS VH THEN confidence (splice event) IS VH	100
11	IF Alignment_Score IS VL THEN confidence (splice event) IS VL	25
12	IF Alignment_Score IS LOW THEN confidence (splice event) IS LOW	25
13	IF Alignment_Score IS MED THEN confidence (splice event) IS MED	25
14	IF Alignment_Score IS HIGH THEN confidence (splice event) IS HIGH	25
15	IF Alignment_Score IS VH THEN confidence (splice event) IS VH	25

Table 4.2: Rules for computing confidence values for splice events. The list corresponds to the rulebase in figure 4.10. To avoid high confidence values just as a result of frequently high (close to 95%) alignment scores, the rules related to the parameter `Alignment_Score` are weighted less.

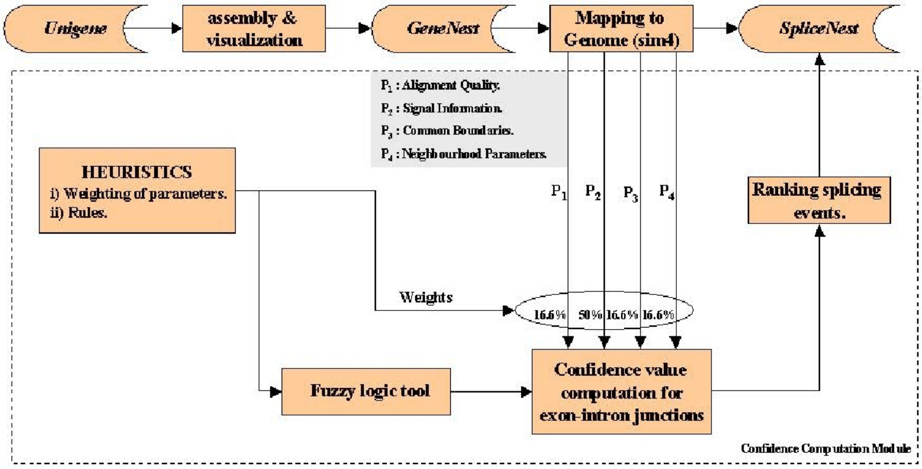


Figure 4.11: Flow-diagram of splicing confidence computation.

EST counts and *Common boundaries (SpliceNest)*, were fixed (*Common boundaries in GeneNest*=0; *Tolerance for detection of the common boundary (SpliceNest)*=0; *Neighboring signal information*=consistent). The quality value obtained over a range of *EST counts* and *SpliceNest common boundaries* is plotted. These plots were computed for the four different levels of signal information (Figure 4.8).

1. *Splice signal present:* In case the splice signal information is present, the quality values obtained are always above 86% (Figure 4.12). The only exception to this would be if the direction of neighboring splice signal is inconsistent, which will decrease the quality values of all boundaries in the transcript. As observed in the figure, the rate of increase in quality is more pronounced when the EST count is low. This reflects the membership function of *est_count*, in which the different classes are closer together for low EST counts than for higher EST counts (Figure 4.7). This implies a higher additional confidence in splice junction upon observing for example a third EST than observing an eleventh EST. The correlation of quality with the *common boundaries (SpliceNest)* is linear, thereby reflecting the symmetric membership function for the parameter.

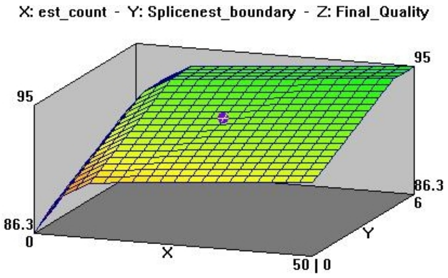


Figure 4.12: Boundary quality values when the splice signal is present.

2. *Start/end of the sequence:* The boundaries marking starts and ends of transcripts are not potential splice junctions. Therefore in the absence of splice signal information, the quality values are lowered (Figure 4.13). The overall landscape of the surface remains similar to the one obtained for boundaries representing splice junctions.

3. *Splice signal absent, or present in incorrect orientation:* As illustrated in Figure 4.14, the quality values are further reduced if there is no splice signal information for the non-terminal splice boundaries. Still, the shape of the plot is preserved since it is only dependent on the EST coverage and common boundaries.

4. *Gapped alignment:* The instances of gaps in the alignment of EST consensus sequences with the genome are unreliable and are therefore correctly attributed low quality values (Figure 4.15).

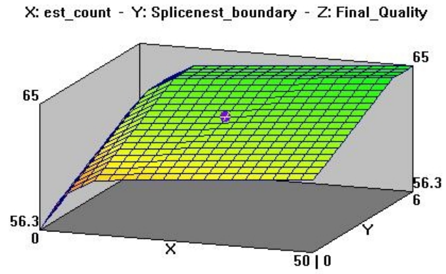


Figure 4.13: Boundary quality values for terminal exons.

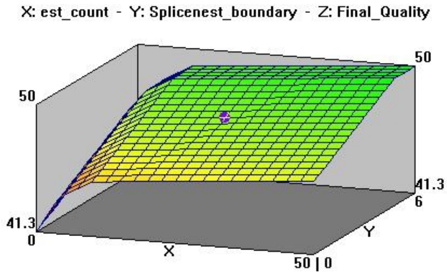


Figure 4.14: Boundary quality values when splice signal is absent.

The above simulations reveal that different levels of splice signal information lead to discrete sets of quality values. This is an effect of higher weights for the splice signal related rules. However for real data, variation in parameters that were fixed for the simulations fills the intermediate gaps.

4.2.4 Statistics for human data

The data used for our quality computation methodology was the June 2003 version of GeneNest consensus sequences (Human UniGene Build 161) and the subsequent mapping of these consensus sequences to the genome (SpliceNest, April 2003 freeze

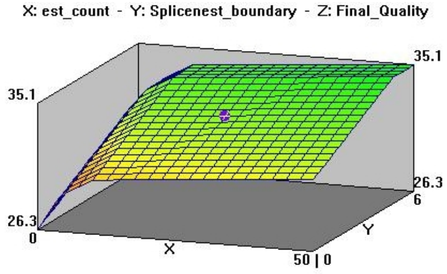
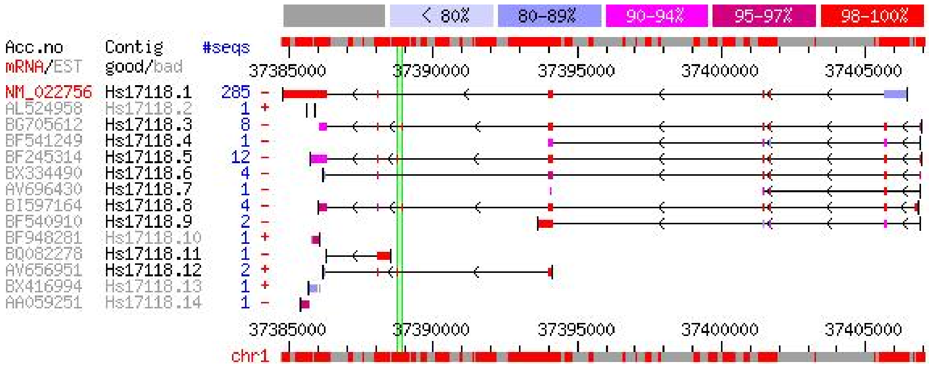


Figure 4.15: Boundary quality values for gapped alignment.

of human genome). Figure 4.16 shows a schematic alignment of EST consensus sequences with the genome sequence (SpliceNest). An example scenario of highlighting the alternative regions (i.e. confidence > 55%) would be if some of the contributing exon boundaries are shared by other transcripts with at least 2 ESTs covering the alternative region with non-consensus splice signal. Alternatively, if the splice-signal is consensus (GT-AG), only a single EST is enough for the confidence value to cross the threshold. Potential alternative regions below the confidence cut-off are not interpreted as splice isoforms. Applying this strategy for analyzing human genes represented by GeneNest EST clusters (total 108094 clusters), constitutive splicing is observed in about one third of these genes (33270), out of which 45% are alternatively spliced. The remaining two thirds of the clusters represent singleton clusters, clusters that are not mapped to the genome, and clusters that do not contain any reliable splice site. Nevertheless, some of these might reflect single exon genes like the human melanocortin 4-receptor gene (Brocke et al. (2002)). To filter out such real single exon genes from data artifacts would require further experimental analysis or verification by an independent dataset.
Our computation of quality values forms the basis to define a confidence measure for the predicted splice variants and to differentiate them from the misaligned transcripts. However, the variants with confidence values below the threshold might include some of the atypically expressed variants (e.g. variants with non-consensus splice signals, lowly expressed variants), which lack the redundant sequence information needed to remedy the otherwise low sequence quality of ESTs. Still, such a confidence measure provides flexibility in selecting a set of (high confidence) splice variants, for experimental and/or theoretical analysis.

Types of predicted alternative splice events Based on the pattern of exon-intron boundaries described by the alternative transcripts, the splice variants are grouped into four types (Figure 2.7), viz. skipped exons, multiple skipped exons, alternative donors/acceptors and retained introns. This classification provides an opportunity to further refine the splicing predictions. The reason is that different splicing types are predicted using different number of boundaries. This in-turn translates into different reliabilities for various splicing types. Therefore, the most reliably recognized splice events are skipped exons (58% of all alternative splice events) since these events are nicely defined by splice signals on both sides of the alternative exon as well as on the respective intron. A subset of the skipped exon events, the multiple skipped exons (11% of all alternative splice events) are usually predicted due to long introns derived from very few (frequently tumor-related) ESTs. These splice events might represent splice variants in a different biological context (e.g. tumors, Wang et al. (2003)). Alternatively, this unusual splicing behavior might reflect leakage of splicing machinery using randomly chosen donor/acceptor sites. In contrast to the skipped exons, retained introns (5% of all alternative splice events) are represented by a single splice signal, thereby reducing the confidence in these boundaries. Additionally, frequent contami-



Acc.no mRNA/EST	Contig good/bad	#seqs	
NM_022756	Hs17118.1	285	-
AL524958	Hs17118.2	1	+
BG705612	Hs17118.3	8	-
BF541249	Hs17118.4	1	-
BF245314	Hs17118.5	12	-
BX334490	Hs17118.6	4	-
AV696430	Hs17118.7	1	-
BI597164	Hs17118.8	4	-
BF540910	Hs17118.9	2	-
BF948281	Hs17118.10	1	+
BQ082278	Hs17118.11	1	-
AV656951	Hs17118.12	2	+
BX416994	Hs17118.13	1	+
AA059251	Hs17118.14	1	-

Figure 4.16: Visualization of alternative splicing in SpliceNest (Hs.17118). The uppermost and lower-most horizontal bars represent the genomic sequence, which is greyed out in the regions representing repeats. The alignments in the middle represent different contigs/transcripts of a cluster/gene. These contigs split up into exons (thick bars) and introns (lines) with the presence or absence of both splice signals GT-AG (arrows on the intron line). A vertical line at the end of a bar represents the end of a consensus sequence. The color of the exon represents percentage alignment with the genomic sequence (as defined by the top color bar). The green vertical bars label significant differences as alternative splicing, while the potential data artifacts like genomic contamination towards the ends of consensus sequences (e.g. the leftmost exon of contig Hs.17118.9) are not labeled.

nation of mRNA samples by un-spliced RNAs lead to cDNAs and therefore ESTs that represent genomic sequence (Sorek and Safer (2003)). Such genomic contamination in turn leads to a splicing pattern resembling a retained intron. Consequently, the retained intron events are the least reliable type of alternative splicing. Nevertheless, optimally aligned retained introns that are covered by multiple ESTs are detected although with a lower confidence than skipped exon events with similar alignment quality and EST coverage.

4.2.5 Validation via known instances of alternative splicing

The database AEDB (Alternative Exon Database, Stamm et al. (2000)) is a collection of experimentally verified human alternative exons. The data is manually gathered from literature and is therefore of very high quality. This database was used to validate our method for prediction of splice events and the computation of confidence values. The sequence information in the database was mapped to the GeneNest consensus sequences using BLAST. 797 exons of the total 1022 human exons in the database could be successfully mapped to at least one of the GeneNest consensus sequence. In order to evaluate the computed quality values, the consensus positions of these *known* exons need to coincide with one of the exons detected in SpliceNest. This was the case for 583 of the exons. Most of the remaining exons (147) had only one confirmed boundary matching to a SpliceNest boundary. These matching boundaries/exons were used for the validation of the computed boundary qualities as well as the alternative splicing confidence values.

Quality of alternative exons The distribution of the quality values for known alternative exons (from AEDB) was contrasted with all alternative exons detected in SpliceNest. In this analysis, we observed that most of the known exons are attributed high quality values (Figure 4.17). Only three of the known exons have quality values below the defined threshold of 55%. These cases reveal instances of non-optimal removal of the linker sequences from EST data which in turn led to gaps while aligning the EST-based consensus sequences to the genome. The quality values were similarly high for the matching boundaries of the exons with only one verified SpliceNest boundary. Notably, the sensitivity evaluation provides an incomplete assessment of the method in the absence of specificity, which cannot be computed without large scale experimental analysis.

In principle, such a validation should be performed over all known exons. However, in the absence of a repository of experimentally verified constitutive exons, validation via *only* alternative exons provides an estimate of the base level of sensitivity. This is because constitutive exons surrounding the alternative exon are shared with at least one additional transcript. The higher copy number of constitutive exons compared to the corresponding alternative exons potentially leads to higher EST coverage and therefore higher quality values.

Confidence values for alternative splicing events The 583 known alternative exons that mapped completely to an exon in SpliceNest also formed the basis for evaluating the computed confidence values for alternative splice events. However, only 388 of these exons were detected as *alternatively skipped* in SpliceNest. The reason for a majority of undetected alternative exons is the absence of the corresponding intron in the SpliceNest data. In a few additional instances, the intron was tagged as a *bad* intron due to the lack of sufficient splice-site information and/ or non-optimal alignment. The confidence values of the 388 instances of identified alternatively skipped exons were contrasted to the confidence values of all predicted alternatively skipped exons. In contrast to the observed alternative exon splice events, *almost all* (384 out of 388) of the known alternative splicing events were attributed high confidence values (Figure 4.18).

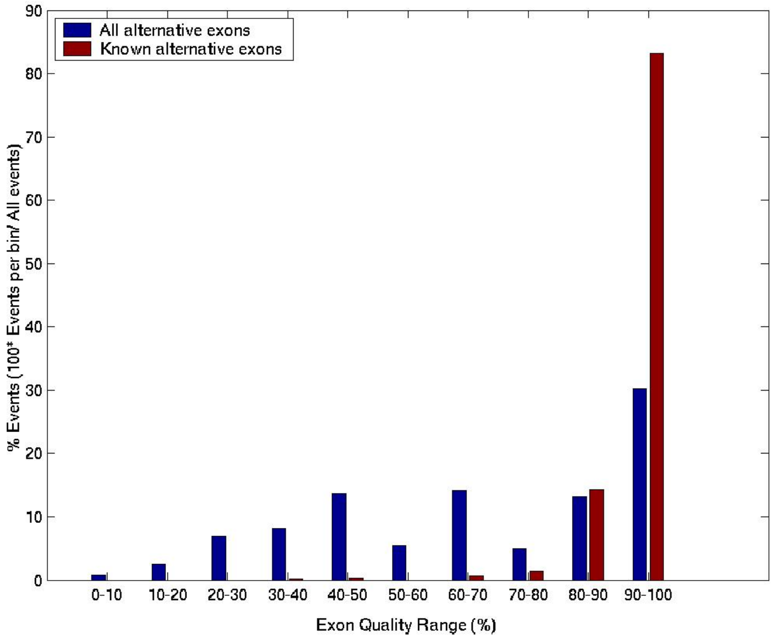


Figure 4.17: Quality values of known alternative exons vs predicted alternative exons. The distribution of quality values obtained for known alternative exons are contrasted with all the predicted alternative exons. The vertical axis is normalized for the total number of events which is several orders of magnitude different for the known and the predicted set. Almost all the known alternative exons are included among the high quality exons (80-100%).

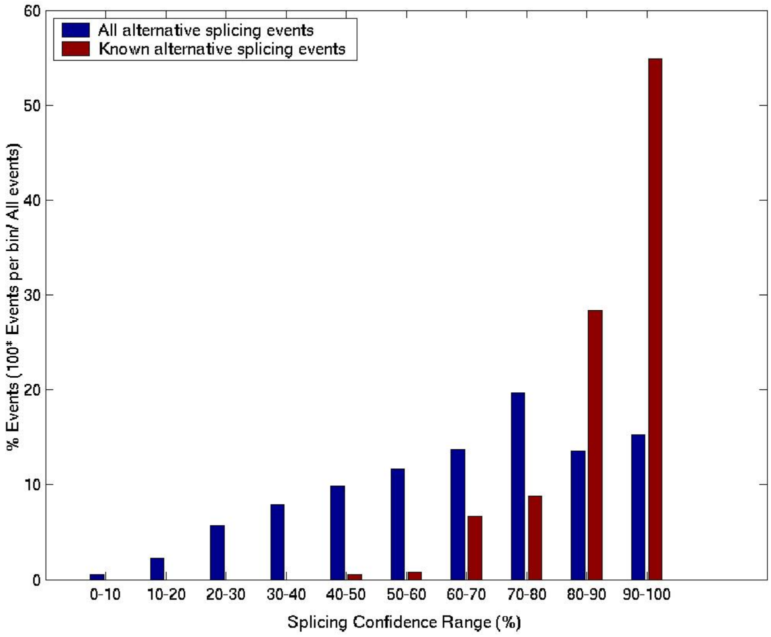


Figure 4.18: Confidence values of known alternative splicing events vs predicted alternative splicing events. The distribution of confidence values obtained for known alternative splicing events are contrasted with all the predicted alternative splicing events. The vertical axis is normalized for the total number of events which is several orders of magnitude different for the known and the predicted set. Since the confidence values include the quality of potentially less evident introns, the values are slightly shifted towards the left. Still, all of the known alternative splicing events are attributed significantly high ($> 60\%$) confidence values.

Serial	Perturbation Name	Description
1	EST mf dec	Peaks of EST count mem. func. (2,4,10)->(1,3,8)
2	EST mf inc	Peaks of EST count mem. func. (2,4,10)->(3,5,12)
3	EST weight dec	Weights of EST count rules halved
4	EST weight inc	Weights of EST count rules doubled
5	sp mf dec	Peaks of common SpliceNest bound mem. func. (5,6)->(4,5)
6	sp mf inc	Peaks of common SpliceNest bound mem. func. (5,6)->(6,7)
7	sp weight dec	Weights of common SpliceNest bound rules halved
8	sp weight inc	Weights of common SpliceNest bound rules doubled
9	ss weight dec	Weights of splice signal rules (30->20)
10	ss weight inc	Weights of splice signal rules (30->40)

Table 4.3: Perturbations applied to the fuzzy logic system. The EST count and splice boundary count are perturbed both using their weights and membership functions. The splice signals are discrete indicators of splicing therefore corresponding perturbations in membership functions was not possible. For these instances only the weights were varied in either direction.

4.2.6 Evaluation of robustness of the model

The Alternative Exon Database (AEDB) was further used to evaluate the robustness of the fuzzy logic model. Three parameters viz. *EST count*, *common SpliceNest boundaries* and *splice signals* were selected as a representative set for all parameters. Subsequently, the fuzzy logic system was systematically perturbed using these parameters. At first, the membership functions of each of the three parameters were changed symmetrically by altering the peaks of the respective membership functions (Table 4.3). Secondly, the weights imparted to the related rules were increased/ decreased. *Splice signals* being discrete indicators of splicing were perturbed only using the weights, since the corresponding perturbations in membership functions was not possible. In total, 10 different perturbations (Table 4.3) were applied to the model, one at a time. The resulting fuzzy models were used to compute quality values of the set of known alternative exons derived from the AEDB database. The quality values computed using the perturbed models and the initial model are plotted in Figure 4.19.

Classification into good/bad exons Across all perturbations, the quality values of three exons lie below the defined threshold of 55%. As discussed in Section 4.2.5, these low values result from the non-optimal removal of linker sequences from EST data. The linker sequences in turn led to gaps while aligning the EST sequences to the genome, thereby falsely implicating low quality values for these splice junctions. Such instances point to the need for improved detection of linker sequences.

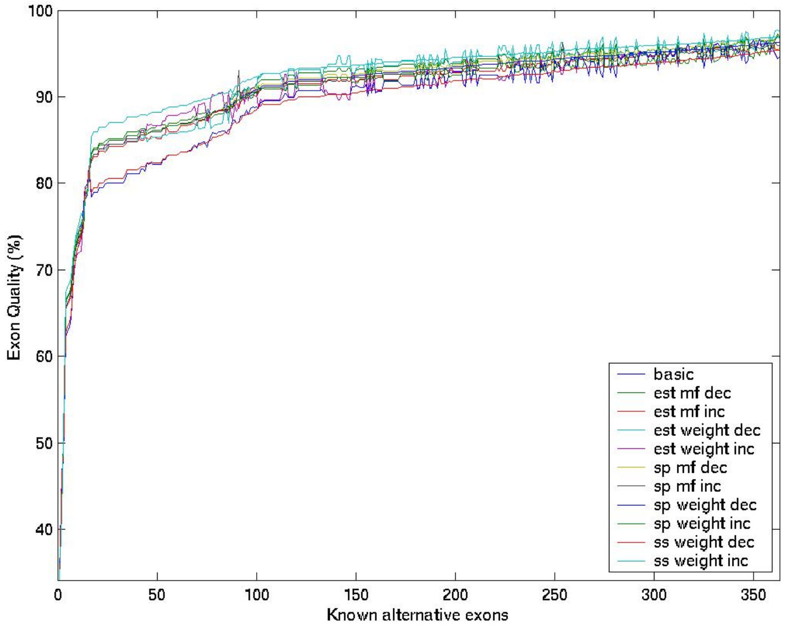


Figure 4.19: Quality values for the known alternative exons (AEDB) for different perturbations in the fuzzy logic model.

Variation in computed exon quality values For all the perturbations, the difference between the computed quality values using the initial model and the perturbed model is maximally 4.45 percentage points for any of the known exons. The mean difference in quality values was 1.71 while median was 1.52. Such small changes in quality values do not disturb the interpretation of the exon quality as being very low, low, medium, high or very high.

Ranked list of exons In the Figure 4.19, the exons are sorted by the quality values obtained by the initial model, thereby resulting in the corresponding curve (basic) being monotonically increasing. For any other curve, a break in monotonicity (peak/dip) implies changes in the quality based ranking of exons. The frequent peaks or dips in the computed exon qualities correspond to disturbance in the ranked list of exons. However, independent of the modeling, such fluctuations are an inevitable consequence of the high degree of freedom of the input data (large number of parameters).

4.2.7 Experimental validation

Experimental validation methods for transcriptome analysis are usually based on cDNA libraries. This is achieved either by the use of oligonucleotide arrays and/or by the use of RT-PCR based methods. While oligonucleotide arrays serve to provide large scale validation, automated RT-PCR experiments provide the same on a smaller scale. Still, RT-PCR experiments prove to be a better validation for initial understanding of the data. The reason is two-fold. On one hand oligonucleotide array based analysis includes problems related to signal-to-noise ratio, normalization methods and reproducibility of experiments. On the other hand, owing to the PCR amplification steps, the RT-PCR experiments are much more robust as well as reliable.
A set of 17 putative alternative splice events on chromosomes 21, 22 and X was arbitrarily selected using a previous version of SpliceNest data not employing the quality computation methodology. For these splice isoforms, PCR primers were generated on either side of alternative splice events using the primer design software GenomePRIDE (Haas et al. (2003)). The computed primers were then used for RT-PCR experiments on 40 different tissue samples (see Appendix B for a list of tissues). These primers were subsequently mapped to the updated version of SpliceNest (which includes the quality computation module), thereby formulating a basis for validation of the methodology. Out of these 17 alternative splice events analyzed by RT-PCR (Table 4.4), 11 were correctly attributed as true variants. Concomitantly, 4 variants were correctly assigned confidence values below the defined threshold. However, the remaining two predicted splice variants, with confidence values above the threshold, were not observed in the experiments. Nevertheless, the absence of both these cases in the RT-PCR experiments can be partially explained. The first false positive (Hs.25854) might represent a tissue and/or stage specific splice event since it is exclusively represented by ESTs derived from fetal brain. However, since our set of RT-PCR tissues include fetal brain tissue, this explanation may be valid only for other tissues. The second false positive case (Hs.49391) is better explained. The isoform is represented in EST data only via normalized libraries (discussed in Section 3.3.2). This implicates it to be a potentially rare transcript, therefore explaining the difficulty of detecting this transcript via standard PCR techniques. These features of the EST data frequently affect the tissue-specificity prediction tools and are discussed in detail in Chapter 5.

4.3 Alternative splicing in coding/non-coding regions

In order to understand the functional role of alternative splicing, the distribution of alternative splicing events in the coding and non-coding regions of the transcripts was investigated. To achieve that, a BLAST search against the SWISSPROT protein database was performed for all genes that were predicted to be alternatively spliced. The following criteria were then applied to create a dataset suitable for analyzing the

Cluster	Chromosome	Prediction	RT-PCR	Validation(Quality)
Hs.10267	22	Yes	Both variants verified	True Positive (88%)
Hs.75527	22	Yes	Both variants verified	True Positive (86%)
Hs.92260	22	Yes	Both variants verified	True Positive (89%)
Hs.351478	22	No	One variant verified	True Negative
Hs.129829	21	Yes	Both variants verified	True Positive (86%)
Hs.235887	21	Yes	Both variants verified	True Positive (91%)
Hs.181581	21	Yes	Both variants verified	True Positive (84%)
Hs.198308	21	No	One variant verified	True Negative
Hs.457939	21	No	One variant verified	True Negative
Hs.330208	21	Yes	One variant verified	False Positive (89%) - fetal brain.
Hs.49391	21	Yes	One variant verified	False Positive (89%) - Normalized ESTs
Hs.58668	21	No	One variant verified	True Negative
Hs.408790	21	Yes	Both variants verified	True Positive (86%)
Hs.75238	21	Yes	Both variants verified	True Positive (84%)
Hs.86958	21	Yes	Both variants verified	True Positive (87%)
Hs.821	X	Yes	Both variants verified	True Positive - Non-consensus splicing
Hs.1757	X	Yes	Both variants verified	True Positive (89%)

Table 4.4: Experimentally verified alternative splice events. The list shows the cluster ID from a previous version of SpliceNest database not employing the quality criterion with positive/negative results in the RT-PCR experiments. 15/17 clusters are now correctly annotated using the quality criterion, with the relevant quality values in parenthesis. The two mis-predictions might be a result of low expression and/or developmental stage specific splicing (fetal brain specific).

distribution of alternative splicing in known genes.
1. Consider only the transcripts of genes for which a gene symbol is annotated in the UniGene data.
2. Alternative splicing prediction quality > 70%, which enriches for transcripts with alternative regions covered by at least 3 ESTs and exon boundaries characterized by reliable splice signals.
3. Blast similarity score > 100.
4. Percent identity > 95%.
5. Consider only skipped exon events.

The alternative exons were then mapped to the 5' UTR, 3' UTR and coding regions of the genes.

In this analysis, 7974 spliced genes produced a good quality mapping to the protein sequences. Out of these, 2926 exons corresponding to 1730 genes are alternatively spliced. Among these exons, a large fraction (62%) map to the coding regions of the genes. Nevertheless, 14% of the exons map to the 5' UTR while 7% map to the 3' UTR. The remaining 17% of the exons overlap the translated and the un-translated parts of genes.

As expected, the coding regions contain the largest fraction (62%) of alternative variants, which are likely to result in functionally divergent proteins (Yan et al. (1996)). This large fraction of alternative splicing may partially reflect the current experimental focus on the analysis of protein sequences, therefore increasing the chance of detecting splice variants in coding regions whereas splice variant detection in the non-coding region is mainly based on ESTs. Furthermore, due to the experimental protocol of EST generation the 3' UTR is usually covered by a larger number of ESTs than regions further upstream, thus increasing the chance of detecting splice variants. Therefore, in 5' UTR the EST coverage is often much lower, thereby complicating the detection of splice variants in this region. Nevertheless, the fraction of alternative splice variants in the 5' UTR is two times higher than in the 3' UTR implying a higher functional importance of splicing in the 5' UTR as compared to 3' UTR. Assuming that alternative splice variants in the non-coding regions of the gene are not just related to nonsense mediated mRNA decay (NMD: Lewis et al. (2003)), these variants could for instance represent translational control mechanisms as summarized by Wilkie et al. (2003) & Kuersten and Goodwin (2003). In case of 5' UTR , alternative splicing may affect exons carrying a uORF that influence the expression of the downstream ORF (Jin et al. (2003)). Additionally, the 5' UTR variants along with a fraction of variants overlapping 5' UTR and coding region may represent alternative start-sites of transcription (alternative promoters: Itani et al. (2003); Delaloy et al. (2003); Brown et al. (1999)). Interestingly, these alternative promoters although revealed as splice variants may be related to a completely different regulation machinery. Prediction as well as analysis of these alternative promoters forms one of the outlook of this thesis.