Intrasplicing - Analysis of Long Intron Sequences

S. Ott, Y. Tamada, H. Bannai, K. Nakai, and S. Miyano

Pacific Symposium on Biocomputing 8:339-350(2003)

INTRASPLICING - ANALYSIS OF LONG INTRON SEQUENCES

S. OTT*†, Y. TAMADA*‡, H. BANNAI†, K. NAKAI†, S. MIYANO†

†*Human Genome Center, Institute of Medical Science, The University of Tokyo*
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
{*ott, bannai, knakai, miyano*}*@ims.u-tokyo.ac.jp*

‡*Department of Mathematical Sciences, Tokai University*
1117 Kitakaname, Hiratsuka-shi, Kanagawa, 259-1292, Japan
tamada@ims.u-tokyo.ac.jp

We propose a new model for the splicing of long introns, which we call *intrasplicing*. The basic idea of this model is that the splicing of long introns may be facilitated by the splicing of inner parts of the intron prior to the splicing of the long intron itself. Since long introns have up to about 100,000 bases, this model seems to be a likely explanation of their splicing. To investigate the possibility of this model, we develop a new computational method for the analysis of DNA sequences with respect to splicing. We analyze the genomic sequence of four species with our method and derive several results indicating that intrasplicing may be an appropriate model for the splicing of at least part of the long intron sequences.

1 Introduction

Nuclear splicing is known to play an essential role in the expression of genetic information of eukaryotes. The molecular components responsible for carrying out the splicing process, which removes the introns from the pre-mRNA sequence, are becoming more and more elucidated and a large amount of research is spent on alternative splicing, aberrant splicing and on splicing inhibitors as well as splicing enhancers[2,3,4]. However, while the splicing reaction can be easily imagined to happen on short introns of a few hundred bases, it is poorly understood how long introns can be correctly recognized by the splicing machinery. Such introns can contain up to about 100,000 bases (and even more for some species[10]) while the known splice signals are limited to an area of about 50 bases around the beginning and the end of the intron. Therefore, it seems rather unlikely that such introns are removed from the pre-mRNA in a single reaction, because this would require the beginning and the end of such introns to get spatially very close.

Recursive splicing has been proposed as one explanation for the splicing

*These authors contributed equally to this work. Please address correspondence to these authors.

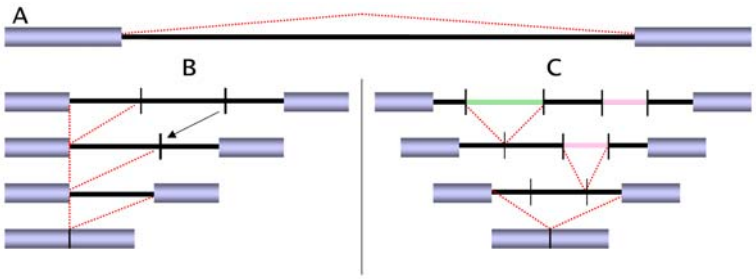


Figure 1. Three models of long intron splicing: A) splicing in one step, B) recursive splicing, C) intrasplicing.

of long introns[6]. In this model, splice sites of consecutive splicing reactions are generated by preceding splicing reactions. Recursive splicing has been shown to occur in one of the long introns of *Drosophila*[6], but it might be only one of several ways of splicing a long intron. Therefore, we propose a more general model for the splicing of long introns, which we call *intrasplicing*. In this model the long intron is shortened by several splicing reactions occurring within the long intron without the use of the long intron splice sites itself. As soon as the remainder of the long intron becomes sufficiently short, it is cut out of the pre-mRNA by a single splicing reaction. This model is less restrictive, because splicing reactions do not have to yield new splice sites and the process as a whole can take place in many different ways. We use the term *intraintron* to denote (putative) introns within introns.

The aim of this work is to investigate the possibility that intrasplicing is an accurate model for the splicing of long introns in a computational way (see Methods). We make use of an intron prediction program as well as the genomic sequence of human (*H. sapiens*), mouse (*M. musculus*), fly (*D. melanogaster*), and zebrafish (*D. rerio*).

We derive several results indicating that intrasplicing may be an explanation or a partial explanation for the splicing of long introns.

2 Methods

2.1 Outline of our Approach

In order to evaluate the possibility of intrasplicing, we have developed a computer program, which can be outlined as follows. First, our program learns an intron predictor for short introns (at most 400 bases) using two thirds of

the currently available mRNA data of the particular species. Then it uses the remaining third of the data to evaluate the probability that a prediction is true with respect to its score. Using the intron predictor for short introns and the probability function, our program recursively searches the highest scoring intraintron candidate within a long intron and cuts it out, until the remainder of the long intron becomes sufficiently short.

Step 1: Learn a short intron predictor from mRNA data.
Step 2: Evaluate the intron predictor.
Step 3: Use the intron predictor to compute a likely way for intrasplicing for all long introns.

By doing so for all long introns longer than 10,000 bases, the program computes several statistics concerning the intraintron candidates used. This cutting procedure is then conducted for several different sequences such as sequences derived from Markov Chains, and statistics are computed for these sequences in the same way. If intrasplicing is an appropriate model for long intron splicing, we expect to find many high scoring intraintron candidates in long intron sequences. It is also expected that intrasplicing statistics between intronic sequences and sequences which are close to exons, will show significant differences. We explain the three main steps in more detail.

2.2 Intron Predictor

We have implemented an intron prediction program, which is similar to the one used in Lim and Burge[8]. This predictor includes inhomogeneous first-order Markov Models for the splice sites, a weight matrix model for the branch point, a length score, and an intron composition score. All introns longer than a certain threshold are removed from the data during training of this predictor. We only describe differences to Lim and Burge's method here.

The intron predictor of Lim and Burge is designed for predicting very short introns of length at most 134. For our purpose, we also need to predict introns longer than this, because introns of length 200 or 300 can easily be thought to be spliced in a single reaction. In order to get higher prediction accuracies for introns of length up to 400, we changed the

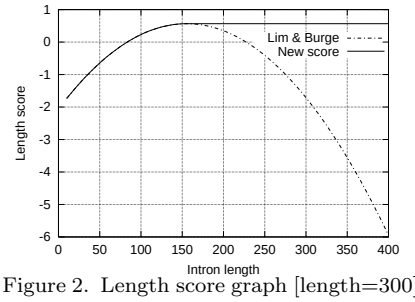


Figure 2. Length score graph [length=300].

length score as shown in Figure 2.

The intron composition score used in Lim and Burge[8] is defined as the sum of the pentamer scores contained in the intron candidate. Scoring in this way yields a length dependent score, since longer candidate regions get a higher score if they are composed of high-scoring pentamers. To remove this length-dependency, we used the average of the pentamer scores instead and evaluated the optimal weight for this score.

Furthermore, in order to improve the prediction accuracy, we conduct a clustering of splice sites in the training procedure of our intron predictor and learn one inhomogeneous first-order Markov Model for each cluster. We use this set of Markov Models by applying all models and choosing the highest score. We use five clusters for the 5' site and two clusters for the 3' site, because this configuration performed best in preliminary computations. The branch point score was not calculated for species other than human, because it has been reported to contribute very weakly to intron prediction accuracy[8].

Figure 3 shows short intron prediction accuracies with respect to the maximal length of predicted introns[a].

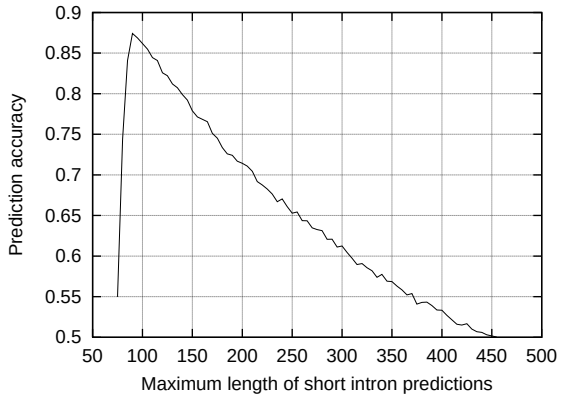


Figure 3. Short intron prediction accuracy. The accuracy is measured as sensitivity plus specificity divided by two. Therefore, 0.5 can be achieved in a trivial way and corresponds to a powerless predictor.[b]

[a] We used human mRNA sequences listed in RefSeq[9] for this computation.

[b] With TP denoting the number of true positives and FP, TN, FN correspondingly, the sensitivity is defined as $\frac{TP}{TP+FN}$, and the specificity is defined as $\frac{TP}{TP+FP}$. Therefore, a predictor selecting all candidates as positives has score 0.5.

2.3 Evaluating the Intron Predictor

After training the intron predictor on two thirds of the mRNA data set, we use the remainder of the data set to evaluate the predictor. We compute the specificity of predictions with respect to the score as shown in Figure 4. That means for each score s, we compute the ratio of the number of real introns scoring above s and the total number of predictions scoring above s. This ratio can be viewed as the probability that a predicted intron candidate of score s is in fact an intron. For very high scores, the function shows a minor instability, because such scores are very rare.

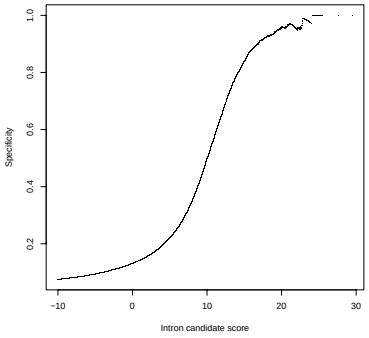


Figure 4. The probability function for short intron predictions.

2.4 Simulated Intrasplicing

The intrasplicing procedure recursively searches for the highest scoring intraintron candidate within the sequence of a long intron and cuts out the region of this candidate until the remainder of the long intron becomes short enough. The reason for selecting only the highest scoring candidate at a time is that non-overlapping candidates are not affected by the cutting and can still be selected in a later step. Since overlapping candidates cannot be selected later, it would be an interesting approach to try different possibilities, but a computation of all possibilities is not feasible. Therefore, choosing the candidate with the highest likelihood is a straightforward way around this problem.

The main parameters of this procedure are the degree to which the long intron is shortened, and the maximum length of short intron predictions. While predicting longer introns promises an easier intrasplicing, increasing the maximum length would limit the quality of intrasplicing predictions, since the accuracy of short intron predictions would decrease, as seen in Figure 3.

We also generated several different sequences to do a comparison of the results. For each long intron, we generated sequences of the same length to exclude the influence of sequence lengths. The sequences were generated from Markov Chain models of exon sequences, intron sequences, gene sequences, and intergenic sequences. Markov Chains are widely used to model DNA sequences[5]. In most computations, we used fifth-order Markov Chains, but we also evaluated the influence of the order of the Markov Chains (see Results).

Furthermore, we concatenated the sequences of all exons of a species and chose random substrings of this sequence of length equal to the particular long intron. Finally, we made use of a 0th-order Markov model of the whole available genomic data, i.e. using the base content of the genome as the only information for sequence generation.

In order to analyze the interior sequence of long introns only, influence of the 5' and 3' splice sites of the long intron had to be excluded. Therefore, the first 8 characters and the last 20 characters of each sequence were not used by our intrasplicing procedure. These values correspond to the length of the intron part of the splice sites as used in Lim and Burge[8].

We used data of the Ensembl database[7] for all of the computations[b]. The Markov Chain models of intergenic DNA were calculated only for human using GenBank[1] data (December 2001). Most of the computations were also performed using mRNA sequences from GenBank, which are listed in the RefSeq[9] database, but the results did not show significant differences to the results derived from Ensembl data. Only the Ensembl results are presented here.

The minimum length for long introns was set to 10,000 bases for all computations concerning a single intron. We excluded long introns with more than 0.5% of non-`acgt`-characters or more than 20 non-`acgt`-characters within a part of 100 characters from the analysis. Furthermore, to avoid the influence of characters other than `a,c,g,t` in the sequence data, we replaced each such character in a random way by one of the nucleotide characters it represents.

The number of long introns meeting the above criteria was 6468 for human, 330 for mouse, 71 for fly and 31 for zebrafish. For fly and zebrafish, no long intron was excluded by our criteria, but for mouse the number of ruled out long introns was very high. Therefore, 3% of non-`acgt`-characters were allowed and the 100-bases-criteria was skipped for mouse in some computations (see Results) yielding a set of 1052 long introns.

The computation was conducted using a PC cluster with 64 Pentium4 CPUs with 2 GHz for about two weeks.

3 Results

3.1 Analysis of Human Sequences

We applied the intrasplicing procedure to all known human long introns meeting the criteria described in Methods. Figure 5 shows the distribution of prob-

[b] *H. sapiens* release 4.28, *M. musculus* release 4.1, *D. melanogaster* release 4.3, *D. rerio* release 4.06.

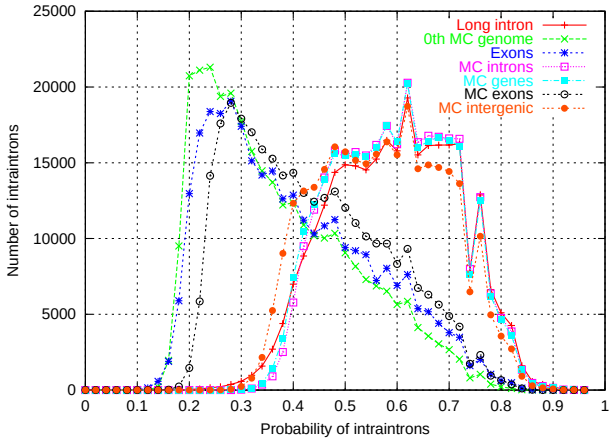


Figure 5. Comparison of probability distributions of various sequences, at least 60% of long introns are removed by simulated intrasplicing.

abilities of putative intraintrons for human. Each line represents one type of sequence. "MC" abbreviates Markov Chain. The sequences clearly fall in two distinct groups. One group consists of long intron sequences, the Markov Chain models for introns, genes, and intergenic sequences. The other group contains sequences from both real and Markov modeled exons, and the 0th-order Markov Chain of the genome. Therefore, long introns as well as the Markov Models for introns and genes show a clearly different behavior than the sequences in the second group. Since the major part of genes is constituted by introns, it is not surprising, that the Markov Models for introns and genes fall in the same group. Interestingly, the Markov Model for intergenic sequences also falls in the same group, though it shows differences to the other three members of the first group (orange line). The reason may be, that the splicing signal is weak enough to allow many sufficiently strong splice sites to occur in non-coding regions. This would imply that the splicing process can be stable against the insertion of intergenic sequences into introns.

Since the 0th-order Markov Model contains the nucleotide frequencies as the only information, it can be considered as a baseline for comparison with the other sequences. Surprisingly, exons as well as their Markov Model achieve only a very slight distinction from this baseline.

Figure 6 shows the average probability of intraintrons as a function of the parameter for stopping the intrasplicing procedure. This parameter specifies the percentage of the length of the original intron which is to be left. Therefore 0.7 indicates, that 30% of the intron have to be removed, while in the strictest

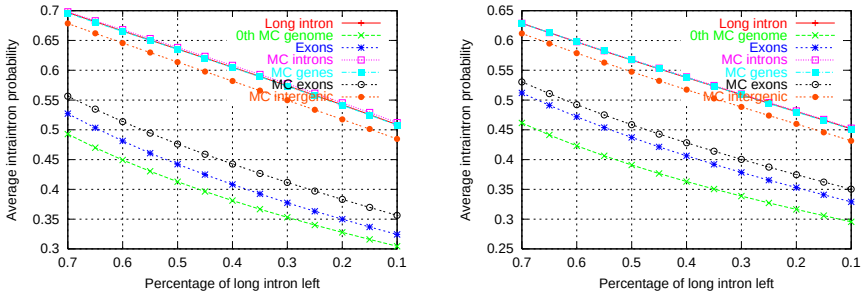


Figure 6. Average probability of intraintrons, (a) with intron composition score, (b) without intron composition score, the maximum length of intraintrons is set to 250.

case 90% are to be removed. The left graph shows the same clustering as we observed in Figure 5. Since the intron prediction program we used includes an intron composition score, we asked whether this might be the explanation for our observation. But as shown in the right graph, the general trends do not change, even if the intron composition score is not used.

Interestingly, while the Markov Model for exons achieves significantly higher intraintron probabilities than the sequences consisting of real exons, the Markov Model for introns does not separate from the intron sequences. This indicates that a fifth-order Markov Chain is a very good model for introns.

We also observe that the Markov Model for intergenic DNA shows a small difference to the intron sequences, if compared to the gap between the two groups. This could be explained, if introns have evolved from intergenic DNA and have further developed since then.

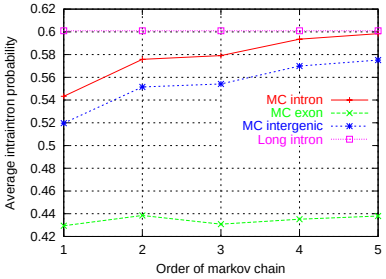


Figure 7. Influence of the order of the Markov Chain.

The influence of the order of the Markov Chain is shown in Figure 7, the value for long introns is shown for comparison. The maximum length of intraintrons is set to 250, and the long introns are removed to at least 60%. The 0th-order models all have low intraintron probabilities, but the 0th-order model for exonic sequences separates from intergenic and intronic sequences. The values for higher order Markov Chains of introns and intergenic DNA are converging towards the value for long introns,

but the values for Markov Chains of exonic sequences stay on the same level.

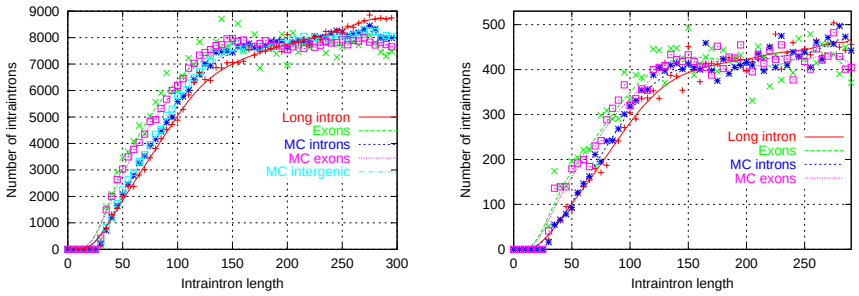


Figure 8. Length distributions of intraintrons, data points and smooth functions, (a) human, (b) mouse. At least 75% of long introns are removed by simulated intrasplicing.

In Figure 8, we show the distribution of lengths of intraintrons. We observe that except for long introns all sequences show a length distribution of the same shape as the length score we used for intraintron predictions (see Figure 2). But the long intron sequences show a different trend favoring longer intraintrons. That means that the long introns show a distinction to all other sequences used. This is an indication of a structure of intraintrons in long introns. It would be very interesting to see whether this trend continues for a longer range of the x-axis, but in lack of a powerful predictor of introns longer than 300 or 400 bases, our sight is limited to this range. Therefore, we suggest that this statistic shows a meaningful trend, which could be seen more clearly if such a predictor would be at hand.

For mouse, this observation cannot be made as clearly, but the amount of complete long intron sequences for mouse is still limited: using the criteria described in Methods, only 330 long introns were selected.

3.2 Comparison between Species

We also applied the intrasplicing procedure to the long intron data sets of mouse, fly and zebrafish described in Methods. Figure 9 shows the average intraintron probability of four species as a function of the upper bound for intraintron prediction. The curve for zebrafish shows a different degree of variation, probably due to the low amount of long intron

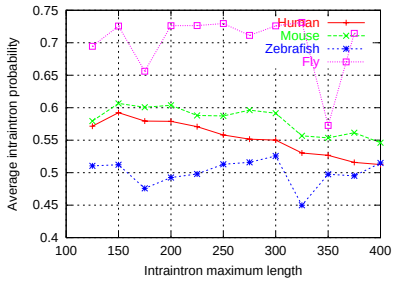
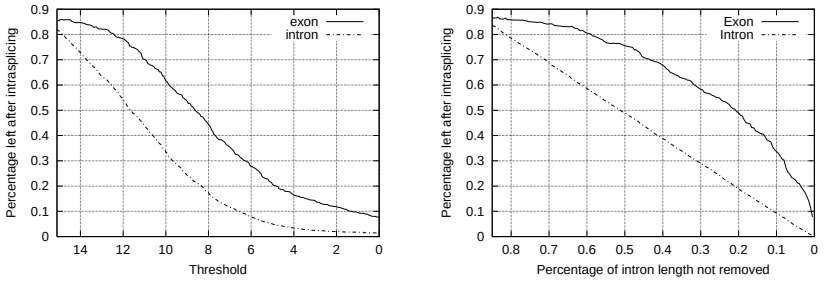


Figure 9. Species comparison of average intraintron probabilities, at least 75% of long introns are removed by simulated intrasplicing.

sequence data for zebrafish.[c] The overall tendency observed for human and mouse is an increase in intraintron probability, when the upper bound is increased from 125 to 150 bases, but further increase of the maximum length does not yield higher intraintron probabilities, which can be explained by the rapid drop of prediction accuracy for longer introns as shown in Figure 3.

It is interesting that fly shows higher intraintron probabilities than other species, because recursive splicing has been reported for fly[6].

3.3 Intrasplicing Simulation on Human Genes



(a) (b)

Figure 10. Comparison of percentage left after intrasplicing simulation.

To evaluate the plausibility of the intrasplicing model as a model for splicing in general, we applied it to whole genes, i.e., instead of trying to remove large parts of long introns, we used complete genes as an input to our intrasplicing procedure and examined to which degree exons and introns are removed for different threshold values. Figure 10 (a) shows the results when applying our intrasplicing procedure to each gene sequence until all putative intraintrons (maximum length 175) scoring above a threshold are removed from the sequence. Figure 10 (b) shows the percentage of intron/exon sequences remaining, as a function of the percentage of the remaining intron sequence.

Our results show that exons are highly stable under intrasplicing. Thus, our model may suggest that simply everything (including introns and intergenic regions) except exons can be removed through intrasplicing.

[c]For this computation the dataset of 1052 mouse long introns was used. A computation on the set of 330 long introns with less non-`acgt`-characters did not yield a significantly different result.

4 Discussion

We have introduced a new model for the splicing of long introns and developed a new computational method in order to analyze whether this model might be appropriate or not. The results show a large difference in the average probability of putative intraintrons between two groups of sequences. Also the shape of the probability distributions is distinctive between both groups. This raises the possibility that splicing of intraintrons precedes and facilitates the splicing of the long intron itself, and that long introns contain a structure of (possibly nested) intraintrons.

Though intergenic DNA and long introns fall into the same group, their average intraintron probabilities and the length distribution of intraintrons show differences. This could be explained, if long introns have evolved from intergenic sequences. The observed intraintron length distribution for intraintrons shows a preference for longer intraintrons. A possible explanation is that longer intraintrons may facilitate a faster intrasplicing process.

The strength of our analysis method is limited by some different factors. First, the current intron prediction programs are powerful only for very short introns, and also these predictors are far from perfect. But in fact, if the intrasplicing model is the true splicing model for at least some introns, the model itself would explain the weakness of intron predictors: if intraintrons are (correctly) predicted as introns by an intron predictor, these predictions would be nevertheless classified as wrong predictions, because there is no data about intraintrons. This mistake would happen increasingly often for longer introns, since they are more likely to harbour intraintrons, which could explain the steep falling of the prediction accuracy (Figure 3).

Predictors making use of the context of introns within genes cannot be used for the analysis of intrasplicing. Therefore, this work shows that stronger intron prediction programs, which do not make use of the context information, would be very useful. This may become reality in the future, as the knowledge of splicing enhancers and splicing inhibitors grows.

We also note that if there are different ways for the splicing of long introns, the significance of our results may be limited by the usage of the whole set of long introns. Also the availability of long intron sequences is still limited for species other than human.

Another approach to improve the results of this work might be to try different ways of selecting intraintrons out of the long intron. For example, it can be thought that some intraintrons are already spliced before the long intron is completely transcribed. In this case, a procedure starting at the 5' region and progressing to the 3' region might yield stronger evidence.

In this work, we presented a computational approach to the known problem of long intron splicing. The intrasplicing model we proposed seems to be a likely explanation for long intron splicing and is worth to be examined. Since *in vitro* experiments on intermediate splicing products of long introns are presently difficult to conduct, there is the possibility that long introns contain a structure of intraintrons, though there are no experimental results supporting this hypothesis so far. The attractive model we proposed should therefore be pursued more extensively.

Acknowledgments

This research was supported in part by Grant-in-Aid for Encouragement of Young Scientists and Grant-in-Aid for Scientific Research on Priority Areas 'Genome Information Science' from the Ministry of Education, Culture, Sports, Science and Technology of Japan. We would like to thank the anonymous referees for valuable comments on this research.

References

1. D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, B.A. Rapp, D.L. Wheeler. GenBank. *Nucleic Acids Research*, 30(1):17–20, 2002.
2. B. J. Blencowe. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *TIBS*, 106–110, March 2000.
3. L. Cartegni, S.L. Chew, and A.R. Krainer. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature reviews*, 3:285–298, 2002.
4. J.F. Cáceres and A.R. Kornblihtt. Alternative splicing: multiple control mechanisms and involvement in human disease. *TRENDS in Genetics*, 18(4):186–193, 2002.
5. R. Durbin, S.R. Eddy, A. Krogh, G. Mitchison. Biological Sequence Analysis. Cambridge University Press, 1998.
6. A.R. Hatton, V. Subramaniam, A.J. Lopez. Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions. *Molecular Cell*, 2(6):787–797, 1998.
7. T. Hubbard, et al. The Ensembl genome database project. *Nucleic Acids Research*, 30(1):38–41, 2002.
8. L.P. Lim and C.B. Burge. A computational analysis of sequence features involved in recognition of short introns. *PNAS*, 98:11193–11198, 2001.
9. K.D. Pruitt, D.R. Maglott. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research*, 29(1):137–140, 2001.
10. D. Tollervey and J.F. Caceres. DNA processing marches on. *Cell*, 103(5):703–709, 2000.