# The role of pre-mRNA secondary structure in gene splicing in *Saccharomyces cerevisiae*

by

Sanja Rogic

B.Sc., The Faculty of Mathematics, The University of Belgrade, 1994
M.Sc., The University of British Columbia, 2000

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

The Faculty of Graduate Studies

(Computer Science)

The University Of British Columbia

October 12, 2006

# Abstract

The process of gene splicing, which involves the excision of introns from a pre-mRNA and joining of exons into mature mRNA is one of the essential steps in protein production. Although this process has been extensively studied, it is still not clear how the splice sites are accurately identified and correctly paired across the intron. It is currently believed that identification is accomplished through base-pairing interactions between the splice sites and the spliceosomal snRNAs. However, the relatively conserved sequences at the splice sites are often indistinguishable from similar sequences that are not involved in splicing. This suggests that not only sequence but other features of pre-mRNA may play a role in splicing. A number of authors have studied the effects of pre-mRNA secondary structure on splicing, but these studies are usually limited to one or a small number of genes, and therefore the conclusions are usually gene-specific.

This thesis aims to complement previous studies of the role of pre-mRNA secondary structure in splicing by performing a comprehensive computational study of structural characteristics of *Saccharomyces cerevisiae* introns and their possible role in pre-mRNA splicing. We identify long-range interactions in the secondary structures of all long introns that effectively shorten the distance between the donor site and the branchpoint sequence. The shortened distances are distributed similarly to the branchpoint distances in short yeast introns, which are presumed to be optimal for splicing, and very different from the corresponding distances in random and exonic sequences. We show that in the majority of cases, these stems are conserved among closely related yeast species.

Furthermore, we formulate a model of structural requirements for efficient splicing of yeast introns that explains previous splicing studies of the

RP51B intron. We also test our model by laboratory experiments, which verify our computational predictions.

Finally, we use different computational approaches to identify any structural context at the boundaries or within yeast introns. Our study reveals statistically significant biases, which we use to train machine learning classifiers to distinguish between real and pseudo splice sites.

# Contents

# List of Tables

# List of Figures

# Acknowledgements

Being co-supervised by three supervisors can be both an enriching and challenging experience. In my case, it was much more the former, since my mentors complemented each other so well. First and foremost, I want to acknowledge my primary thesis supervisor, Alan Mackworth, who provided me with unconditional research, personal and financial support for many years during my Masters and Doctoral studies. His constant encouragements and invaluable advice were essential for my progress. I am also extremely grateful to Holger Hoos, whose research guidance was indispensable throughout my Doctoral studies. I especially appreciate his extensive engagement in the final phases of my thesis writing when his thorough comments helped me to significantly improve my dissertation. And last, but not least, I extend sincere appreciation to Francis Ouellette, my third thesis co-supervisor, who taught me a lot about Biology and Bioinformatics and who played a vital part in creating a great Bioinformatics community in Vancouver, which I am proud to be a member of.

I would also like to thank Anne Condon, additional member of my thesis committee. Anne was always available to meet with me and offer me valuable professional and personal advice.

I am thankful to Ben Montpetit, my research collaborator, for his great laboratory work and his patience when answering my numerous questions. I am also grateful to Ben's supervisor Phil Hieter, who made our collaboration possible.

My PhD experience was made more enjoyable by my friends and colleagues from the Beta lab. I would like to thank all of them for their friendship, help with research problems and our fun discussions and lab outings.

I owe so much to my parents, for their unconditional love and support.

Finally, but most importantly, my deepest gratitude goes to my family, my husband Novak for his endless patience, tolerance and encouragements and for his help in producing many figures in this document and my beautiful and clever children Pavle, Tara and Luka, who inspire me to always go forward.

To Novak, Pavle, Tara, Luka and Sofia

# Chapter 1

# Introduction

The central dogma of molecular biology states that the flow of genetic information is from DNA to RNA to protein: genes, which are parts of DNA that store genetic information, are transcribed, i.e., copied, to messenger RNA (mRNA), which carries this information outside the cell nucleus into the cytoplasm where it gets translated into proteins (Figure 1.1). In eukaryotic organisms, protein-coding genes are often interrupted by intervening sequences, called introns, that must be removed from mRNA before it gets translated in order to produce functional proteins.

Splicing of precursor mRNA (pre-mRNA), which involves the excision of introns from a primary RNA transcript and ligation of exons into mature mRNA, is one of the essential cellular processes in eukaryotic organisms. Although this process has been extensively studied since the discovery of splicing almost three decades ago (Chow et al., 1977; Berget et al., 1977), resulting in a thorough understanding of the splicing pathway and identification of the numerous components of the splicing machinery, there are still unanswered questions. One of them is: how are the splice sites accurately identified and correctly paired across the intron? It is currently believed that identification is accomplished, at least partially, through basepairing interaction between the conserved sequences at the exon/intron boundaries and within introns and the small spliceosomal RNAs. However, these conserved sequences are short and not well defined, and are often hard to distinguish from the numerous unutilized sequences throughout the genome.

The relative conservation of the splicing signals is used for the computational identification of exon/intron boundaries, which is an essential part of gene-finding in eukaryotic genomes. Failure to distinguish between real splice sites and unused, 'pseudo' sites is one of the the major reasons for

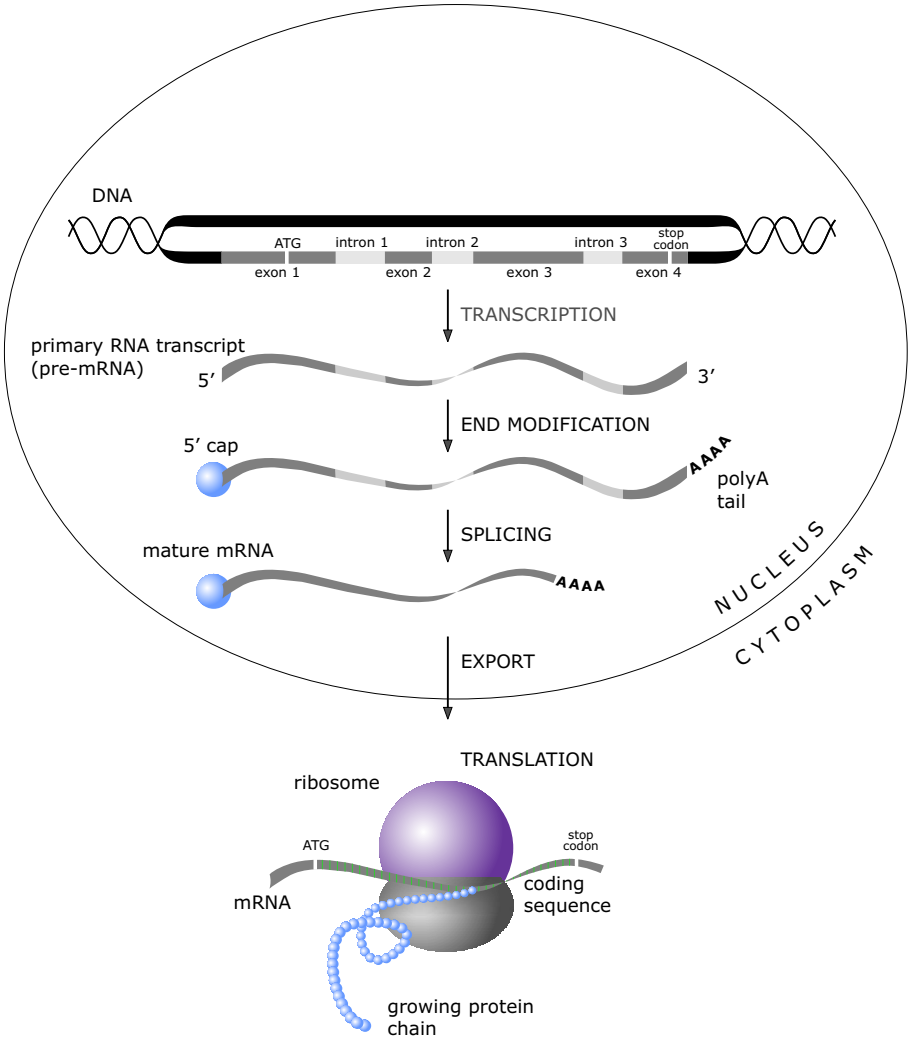



Figure 1.1: The molecular processes involved in the pathway leading from DNA to protein in eukaryotic cells.

the low accuracy of computational gene prediction in longer genomic DNA sequences (Pertea et al., 2001; Rogic et al., 2001).

The fact that the known primary sequence determinants are not able to unambiguously specify splice sites has prompted scientists to speculate about the role of pre-mRNA secondary structure in splicing. There is a large body of biological literature, a sample of which is given in Section 2.3, that describes the different ways in which pre-mRNA secondary structure can affect splicing. In many cases, the induced structural changes have an inhibitory effect on splicing, suggesting that certain structural arrangements or structural elements are important for splicing. The secondary structure of pre-mRNA is also indicated to have a regulatory effect on alternative splicing. However, most of these experimental studies have been focused on the analysis of a single gene or small set of genes and the universal role of pre-mRNA secondary structure has not been determined.

Elucidating the complex details of the gene-splicing process is of significant importance for biology and medicine: it has been estimated that $\sim 15\%$ of human genetic diseases are caused by errors in splicing (Krawczak et al., 1992). This number is likely to be larger since the study focused only on point mutations in vicinity of splice sites ignoring mutations in other *cis* and *trans* splicing factors (Faustino and Cooper, 2003). Consequently, improved understanding of the splicing process and splice site recognition would lead to better computational models and higher prediction accuracy of gene-finding programs. This, again, is an important goal of genomics, where one of the major tasks is accurate annotation of large volumes of genomic sequences generated by numerous genome sequencing projects, which is initially accomplished by computational methods.

This highlights the importance of the problem we study in this thesis, namely, correlation between pre-mRNA secondary structure and splicing. Even though the splicing mechanism is, in general, universal for all eukaryotes, there are some minor differences between the organisms that need to be considered. Therefore, it is beneficial to focus on a single organism when establishing and testing our initial hypothesis about the above-mentioned relationship, which can later be extended and/or modified to apply to other

eukaryotic species.

For the research work in this thesis, we chose *Saccharomyces cerevisiae*, the simplest intron-containing eukaryote, which is also a well-established model organism. We selected *S. cerevisiae* based on its thorough, experimentally supported annotation, large number of splicing studies and limited intron sizes. The *S. cerevisiae* dataset that we use for our study is carefully assembled using three different intron databases and relies on comparative genomic studies to confirm annotated splice sites.

We begin our study exploring the hypothesis suggested by Parker and Patterson (1987) that yeast introns with large distances between the 5' splice site and the branchpoint sequence can fold into secondary structures that would shorten this distance to one that is optimal for spliceosome assembly. This hypothesis was confirmed for a limited number of yeast introns by comprehensive biological experiments that demonstrated that the existence of such secondary structure elements is essential for splicing efficiency. Structural elements that exhibit a similar effect on splicing were also found in some introns of *Drosophila melanogaster* and its related species. Furthermore, shortening of long distances between the essential splicing signals was observed for higher eukaryotes, more specifically mammals, where folding of long intron sequences is facilitated by protein binding and interactions (see Section 2.3).

These studies indicate that pre-mRNA secondary structure within introns might be essential for efficient splicing of long introns in all eukaryotic species, but it is hard to claim universality of the phenomenon based on a very small sample size. Motivated by this limitation, we perform a more extensive computational study of all long introns in *S. cerevisiae*, with the rationale that combining computational evidence for a large dataset of introns with convincing biological evidence for a few introns will strengthen the hypothesis. We commence our study by searching long yeast introns for secondary structure elements that could bring the 5' splice site and the branchpoint sequence into closer proximity and then analyze the resulting shortened distances with respect to distances between splicing signals in short yeast introns, which are assumed to be optimal for splicing.

After our initial analysis, we further refine our method to take into account additional biological evidence, and relax our initial criteria for the targeted structural elements within the long yeast introns. The computational analysis in this part is based on previously mentioned biological studies that examined the relationship between secondary structure formation within yeast introns and experimentally determined splicing efficiency levels. We also reconsider the computation of the secondary structure of introns and shortened distances between the splicing signals to make them less error sensitive. Encouraged by promising results of the computational study, we validate our hypothesis using biological laboratory experiments.

In the second part of this thesis, we consider a different role of pre-mRNA secondary structure in gene splicing: can secondary structure elements serve as additional identifiers of exon/intron boundaries and introns? We employ different approaches to identify thermodynamically stable or conserved structural motifs or specific structural contexts in the vicinity of splice signals or within introns. We also apply a thermodynamical approach to detect splicing signals based on knowledge of their interactions with small nuclear RNAs.

Finally, we use a machine learning approach to investigate the potential of discovered structural characteristics of yeast introns to improve the accuracy of computational splice site and intron recognition in *S. cerevisiae*.

In summary, we perform a comprehensive study of secondary structure characteristics of yeast introns and their relationship to pre-mRNA splicing, using a combination of computational, statistical, phylogenetic and experimental approaches. We consider various aspects and functions of RNA secondary structure. The obtained results support our initial hypothesis that pre-RNA secondary structure is capable of modifying distances between important splicing signals in long introns, bringing them into closer proximity, which is thought to be optimal configuration for splicing. The identified structural elements are also conserved among *Saccharomyces sensu stricto* species, indicating their functional significance. Our model that describes RNA structural requirements for splicing is able to explain the confusing results of several previously reported experimental studies. We further val-

idate it by carefully designed experimental testing.

Our computational and statistical exploration of structural characteristics within introns and around the splice sites identifies a number of structural features that are specific to yeast introns and not observable in a random RNA sequence with the same sequence features. Namely, the 5' splice site sequences are found to have a tendency to be relatively free of secondary structure and to bind more favorably to snRNAs than pseudo sites; branchpoint sequences tend to have more free bases when folded globally, as observed in some previous studies (Hall et al., 1988; Stephan and Kirby, 1993; Mougin et al., 1996; Chen and Stephan, 2003) and certain structural motifs are found to be over-represented in the neighbourhoods of 5' and 3' splice sites.

Our machine learning experiments based on these findings give further support to the validity of our splicing model and demonstrate the ability of these methods to reduce the number of false positive splice site and intron predictions.

## Organization of the thesis

In Chapter 2, we provide the necessary background for the topics covered in this thesis, give an overview of the biological literature that motivated our research and discuss existing work that incorporates secondary structure information in computational splice site prediction methods. Chapter 3 describes important characteristics of *S. cerevisiae* introns and splicing, discusses general features of intron architecture and describes the assembly and properties of the STRIN dataset of yeast introns and the dataset designed for phylogenetic analysis. Chapter 4 introduces the notion of a zipper stem, an RNA structural element that shortens the distance between the 5' splice site and the branchpoint sequence, and describes different computational approaches for its identification. The second part of the chapter investigates the conservation of zipper stems among closely related *Saccharomyces sensu stricto* species. In Chapter 5, the initial model that describes RNA structural requirements for splicing is further refined to take into account

additional biological evidence and to allow relaxing of the initial criteria for the zipper stems within long yeast introns. The chapter also discusses the validation of our model using laboratory experiments done in collaboration with Ben Montpetit and Phil Hieter from the Department of Medical Genetics at UBC. Further structural analysis of yeast introns, as well as the structural context at the splice signals, are discussed in Chapter 6. The machine learning experiments described in Chapter 7 test whether the efficiency of pre-mRNA splicing can be predicted based on secondary structure characteristics of introns. In the second part of the chapter, the weak structural signals discussed in Chapter 6 are used to improve the accuracy of computational splice site and intron prediction by filtering out false positive predictions based on classification methods from machine learning. Chapter 8 summarizes and discusses the results obtained and outlines directions for further research.

Supplementary information and data are provided in a number of appendices: the sequences in the STRIN dataset are listed in Appendix A; the procedure for experimental testing of our splicing model is described in Appendix B; printouts from our StructureAnalyze procedure which calculates structural characteristics important for splicing, are provided in Appendix C; and the sequences of the RP51B intron mutants that we designed for the purposes of experimental verification are specified in Appendix D.

# Chapter 2

# Background and related work

In this chapter, we first introduce the reader to the general process of pre-mRNA splicing and identification of splice sites by the spliceosome machinery. Next, we describe computational methods for predicting splice sites that are based on various sequence-based approaches. The secondary structure of RNA and its computational prediction, which we heavily base our thesis on, are also discussed. A sample of the large body of literature on the effects of pre-mRNA secondary structure on splicing is presented, providing the motivation for our thesis research. Finally, we describe existing attempts to integrate pre-mRNA structural information with sequence information to predict splice sites.

## 2.1 Pre-mRNA splicing and splice site recognition

The process of gene splicing, which involves excision of introns from a primary mRNA transcript and ligation of exons into mature mRNA, is one of the essential steps in protein production. In cells, this process is usually catalyzed by a large ribonucleo-protein complex, called spliceosome, which is composed of five small nuclear RNAs (U1, U2, U4, U5 and U6)*, assembled into small ribonucleo-protein particles (snRNPs) and numerous non-snRNP splicing factors. The exceptions are group I and II introns, which are capable of self-splicing, tRNA introns, where splicing is catalyzed by protein enzymes, and U12-type introns, which are spliced by a compositionally distinct spliceosome (Abelson et al., 1998; Staley and Guthrie, 1998; Lopez and

*The U3 snRNA is not involved in splicing, but participates in the processing of pre-ribosomal RNA.

Séraphin, 2000).

Splicing consists of two consecutive trans-esterification reactions. In the first reaction, the donor splice site at the 5' exon/intron junction is cleaved and the intron 5' end is ligated to the branchpoint, located typically 20 to 40 nt upstream of the 3' splice site. In the second reaction, cleavage of the acceptor (3') splice site releases the intron as a lariat structure and 5' and 3' exons are joined together (Figure 2.1).

Splice site recognition and spliceosome assembly occur simultaneously: the 5' splice site is initially recognized through complementary base-pairing interactions with the U1 snRNA (Mount et al., 1983; Zhuang and Weiner, 1986). In higher eukaryotes, this base-pairing stretches over approximately nine nucleotides (nt), encompassing the last two or three exonic nucleotides and the first five or six nucleotides of the intron. Subsequently, the branchpoint sequence base-pairs with the U2 snRNA (Black et al., 1985; Parker et al., 1987; Zhuang and Weiner, 1989; Wu and Manley, 1989). This interaction involves the U2 snRNA sequence GGUG and the branchpoint consensus signal YYRAY, with the unpaired branchpoint adenosine (A) bulged out of the RNA duplex (Query et al., 1994). The U4/U5/U6 tri-snRNP is then added to this pre-mRNA-snRNA complex, the U6 base-pairing with the 5' splice site intronic sequences (Kandels-Lewis and Séraphin, 1993; Sontheimer and Steitz, 1993) and U5 forming non-canonical base-pairing interactions with the 5' and 3' terminal exonic nucleotides (Newman and Norman, 1992). The complex then undergoes a series of structural rearrangements transforming into a mature spliceosome that is capable of catalyzing splicing reactions (Staley and Guthrie, 1998). This summary is a simplified version of this complex event, since it neglects important functions of the associated splicing factors.

Splicing of introns has to be performed with single-nucleotide precision in order to produce functional proteins. This requires that the actual splice sites be accurately recognized and correctly paired across the intron. The recognition of splice sites is, at least partially, achieved by formation of Watson-Crick base-pairs between some spliceosomal snRNAs and short consensus sequences located at the 5' splice site and the branchpoint (an ex-

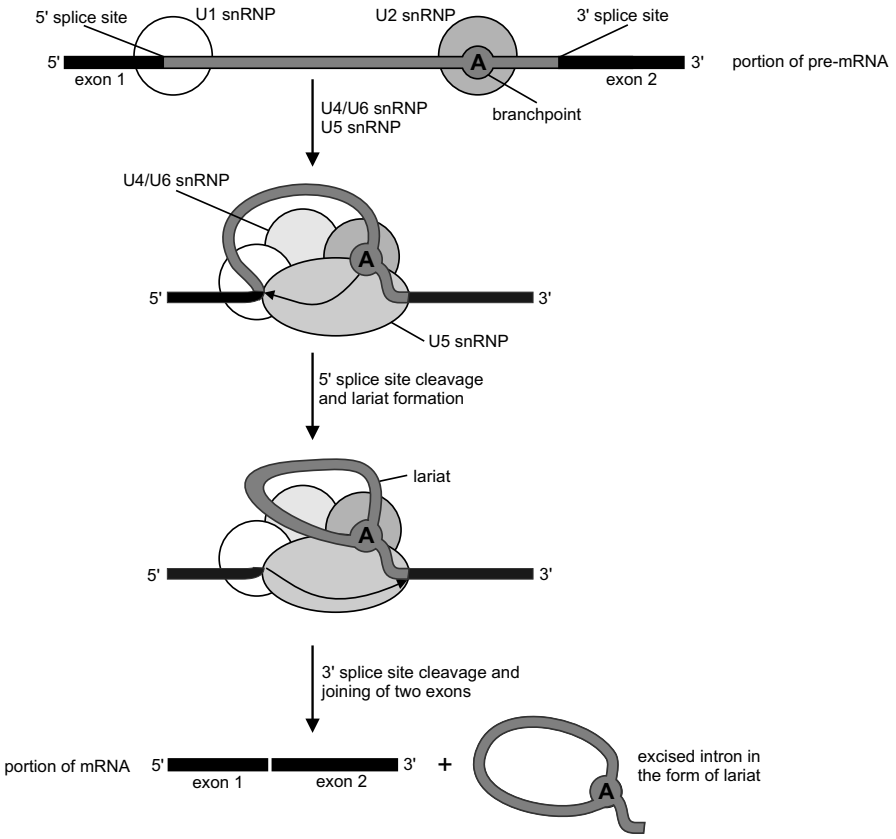



Figure 2.1: Splicing of pre-mRNA.

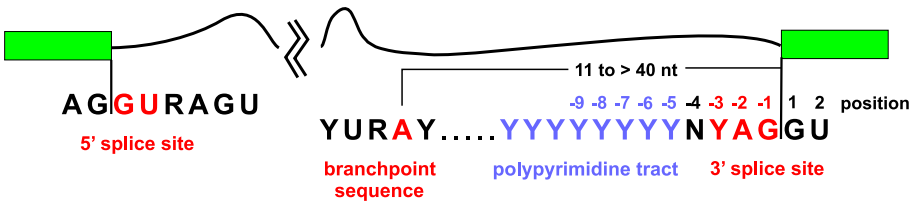



Figure 2.2: Consensus sequences involved in intron splicing in humans. R stands for purines (adenine (A) and guanine (G)) and Y stands for pyrimidines (cytosine (C), thymine (T) and uracil (U)). N stands for any nucleotide. The figure was modified from Moore (2000).

ample for human introns is given in Figure 2.2). Conserved sequences are also found at the 3' splice site and in the form of a polypyrimidine tract (located immediately upstream from the 3' splice site), which mediate splicing through their interactions with splicing factors and non-base-pairing interactions with snRNAs and other intronic sequences (Madhani and Guthrie, 1994). However, these consensus sequences are not uniquely associated with functional splice sites; there are numerous occurrences of these signals throughout the genome not utilized by the splicing machinery. This is illustrated in a study by Sun and Chasin (2000), where positional weight matrices (described in the next section) were trained on 2400 instances of real human donor and acceptor sites to search for splice sites in the 42-kb human *hprt* gene, which contains 8 introns. This approach identified 8 real donor sites along with over 100 pseudo donor sites that have scores higher than the lowest scoring real donor site. The results were even more discouraging for acceptor sites, since 683 pseudo sites were predicted. It is still not fully understood how the precise specificity required to distinguish correct splice sites from similar 'pseudo-sites' is achieved or how the correct donor/acceptor pairs are brought together.

### Computational splice site recognition

Identification of splice sites is an essential component of computational gene-finding in eukaryotic genomes. Relying on biological knowledge and results,

researchers in computational biology approach this problem by modeling consensus sequences around splice sites and within introns. Various methods are used to model splicing signals, such as the following: the simple consensus sequence model, which looks for either a specific sequence motif or allows some alternative nucleotides at certain positions in the motif; position weight matrices, which represent the frequency of appearance of the A, C, G, and T nucleotides at each position of the consensus sequence; and weight arrays, which exploit statistical dependences between adjacent nucleotides (Fickett, 1996; Burge, 1997; Salzberg, 1997). Weight matrices and weight arrays are used to score candidate sequence motives. Neural networks and decision trees are also used for identification of splicing signals (Hebsgaard et al., 1996; Kulp et al., 1996; Burge, 1997).

These sequence sensors are usually not used in isolation, but are integrated with content sensors that use coding statistics to distinguish between coding and non-coding regions. The integrated approaches can either be stand-alone splice site predictors or gene-finders that attempt to identify entire gene structures (splice sites in intron-containing genes and in the boundaries of coding regions). These methods yield better accuracy for splice site recognition because they eliminate false positive splice sites that do not have an expected shift in coding potential (Brunak et al., 1991). There are a number of methods used to combine signal detection with coding statistics for stand-alone splice site prediction, including neural networks (Hebsgaard et al., 1996); Bayesian networks (Arita et al., 2002; Churbanov et al., 2006); rule-based expert systems (Vignal et al., 1999); and discriminant analysis (Solovyev et al., 1994).

An example of an integrated approach for predicting splice sites is linear discriminant analysis (Solovyev et al., 1994), which we use to initially predict donor and acceptor splice sites as described in Section 7.2.2. Linear discriminant analysis is a procedure that finds a linear combination of sequence measures that provides maximum discrimination between real and pseudo-sites. The linear discriminant function is of the following form:

$$z = \sum_{i=1}^{p} \alpha_i x_i \tag{2.1}$$

where $x_1, \ldots, x_p$ are the values of $p$ features of sequence $x$, and $\alpha$=($\alpha_1, \ldots, \alpha_p$) is a vector of coefficients derived from the training set by maximizing the ratio of between-class variation to within-class variation of $z$ . The input sequence $x$ is classified as a real site if $z \geq c$ and as a pseudo site if $z < c$, where $c$ is a threshold constant derived from the training set in the same way as the coefficient vector $\alpha$. Sequence features used to detect donor sites include: agreement with consensus region, average triplet preference in the potential coding region, coding statistics for the coding and intron regions (octanucleotide preference), and the G-richness of the region. The features used for acceptor sites include: agreement with consensus region, average triplet preference in the branchpoint region, agreement with typical polypyrimidine region, as well as coding statistics for the coding and intron regions.

The splice site prediction accuracy of current gene-finding programs is 70-80% when tested on short genomic sequences containing exactly one gene with a relatively simple exon/intron structure (Rogic et al., 2001). The accuracy level drops significantly for more realistic, longer genomic sequences containing multiple multi-intronic genes. The reason for this drop in accuracy is that consensus sequences used to identify splice sites are short and not well defined, and the number of false positive signals that are accepted by signal sensors grows with the length of a DNA input sequence. For the linear discriminant analysis described above, the reported numbers of false positive predictions are on average 1.5 false donor sites per true donor site, and 6 false acceptor sites per true acceptor site (Solovyev et al., 1994). However, our analysis in Section 7.2.2 shows that on the yeast dataset used in this thesis the results are much worse: on average there are more than five false donor sites per true donor site and eight false acceptor sites per true acceptor site.

## 2.2 RNA secondary structure and prediction

Ribonucleic acid (RNA) is a nucleic acid polymer with a backbone of ribose sugar rings linked by phosphate groups. Each sugar has one of the four bases adenine (A), guanine (G), cytosine (C), and uracil (U) linked to it as a side group. The phosphate groups link the 5' carbon of one ribose to the 3' carbon of the next, which imposes a directionality of the backbone from 5' to 3'. RNA usually occurs as a single strand and in many cases forms inter-molecular base-pairing interactions, which constitute the secondary structure of the molecule. Basepairing is accomplished through hydrogen bonding between complementary bases: C can pair with G and U can pair with either A or G. These basepairs are called Watson-Crick or canonical basepairs (G-U is called a wobble pair).

The secondary structure of RNA is composed of a set of elementary structures such as stems, hairpin loops, internal loops, bulges and multi-loops. An example of RNA secondary structure with annotated structural elements is given in Figure 2.3. Pseudoknots are another type of RNA structural elements that occur frequently in nature and in some cases have important functions (Pleij and Bosch, 1989). However, they are often considered to be a part of the RNA tertiary structure.

For many functional RNAs, tertiary structure is a key determinant of their biological function. However, our understanding of tertiary structure formation and interactions is limited, as is the available experimental data. Secondary structure is generally believed to play a crucial role in tertiary structure formation, since most tertiary interactions are thought to arise after the formation of a stable secondary structure, when the molecule is able to bend around the flexible, single-stranded regions (Brion and Westhof, 1997; Tinoco and Bustamante, 1999). The tertiary structure interactions that arise in the later stages of folding are usually too weak to disrupt secondary structure that has already formed.

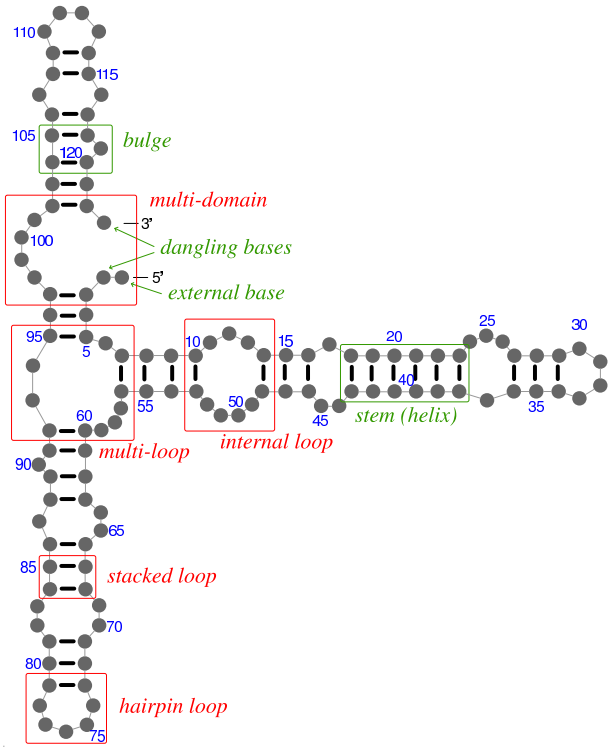



Figure 2.3: An example of RNA secondary structure where the bullets represent the nucleotides in the RNA sequence and the black lines between them represent basepairing interactions. The basic structural elements are annotated. The figure was taken from Andronescu (2003).

### RNA secondary structure prediction

The most common approaches for RNA secondary structure prediction are based on the assumption that, at equilibrium, any RNA molecule folds into its lowest free energy state. Therefore, the aim of structure prediction is to determine a minimum free energy (MFE) structure. Most MFE secondary structure algorithms use dynamic programming to perform a complete evaluation of all feasible structures given an RNA sequence, and determine one with minimum free energy.

The calculation of free energies is based on a nearest neighbour thermodynamic model, which considers the energy contributions of stacking interactions between neighbouring basepairs, assuming these energies are independent and additive. The energy of an RNA molecule is calculated by adding energy contributions of stacking interactions and various types of loops. The energy contributions of stacked basepairs and some types of loops have been experimentally determined, while for the other structural elements they have been estimated (Xia et al., 1998; Mathews et al., 1999). These energy contributions are encoded as the parameters of the energy model. The most widely used energy model is Turner's energy model (Freier et al., 1986; Turner et al., 1987; Turner and Sugimoto, 1988; Mathews et al., 1999).

The most commonly used MFE programs for RNA secondary structure prediction are mfold (Zuker and Jacobson, 1998) and RNAfold (Hofacker et al., 1994). Both programs use dynamic programming for the identification of the MFE structures and base their energy calculations on Turner's energy model. The running time for these two algorithms is $O(n^3)$.

The MFE folding approach has a number of shortcomings:

- Prediction of MFE secondary structures has limited accuracy, that is, the predicted structures are not guaranteed to accurately reflect the physical ground state of respective RNAs. This is partially due to the imperfect underlying energy model, which contains a number of approximative and extrapolated parameters. The experimentally derived parameters are based on very short RNA strands ($\sim$ 20 nt)

that were used to determine the free energy of different secondary structure elements. Consequently, applying these thermodynamic parameters to longer RNA sequences will sometimes result in inaccurate free energy calculations. In general, the prediction accuracy decreases with increasing RNA sequence length, and computational prediction is considered unreliable for sequences longer than several hundred nucleotides (Morgan and Higgs, 1996; Mathews et al., 1999).

Another approximation is the nearest neighbour thermodynamic model itself, since the independence and additivity of structural elements constituting RNA structure are assumptions that may not be entirely accurate.

- Another simplification that most of the prediction algorithms make is that they can predict only pseudoknot-free secondary structures. This limitation is a consequence of the dynamic programming approach, which cannot handle pseudoknots in their most general form. There are ways to include some classes of pseudoknots in predictions but this always results in substantially increased computational time. In fact, if all pseudoknots are included, the secondary structure prediction problems becomes *NP*-hard (Lyngsø and Pedersen, 2000).

  Some examples of RNA secondary structure prediction programs that include pseudoknot predictions are *pknots* by Rivas and Eddy (1999) ($O(n^6)$ time and $O(n^4)$ space), *pknotsRG* by Reeder and Giegerich (2004) ($O(n^4)$ time and $O(n^2)$ space), both considering only a restricted class of pseudoknots, and *HotKnots* by Ren et al. (2005) and a genetic algorithm by Gultyaev et al. (1995), the latter two of which are based on heuristic approaches.

- RNA molecules do not fold in isolation, and contact with proteins or other RNA molecules may play an important role in structure formation. Considering our current biological knowledge and ability to computationally model the folding and structure of RNA sequences, we will not be able to take into account the cellular environment and

RNA-protein molecular interactions in the near future. However, recently some attempts have been made to model the secondary structure interactions between two or more RNA molecules (Rehmsmeier et al., 2004; Andronescu et al., 2005; Mückstein et al., 2006).

One way to deal with inaccuracies of RNA secondary structure prediction is to also compute near-optimal structures. The prediction of suboptimal structures was first proposed by Wuchty et al. (1999) and later implemented in the mfold and RNAfold programs. The computation of suboptimal structures in addition to the optimal structure is important not only because the native structure can be buried in the near-optimal space due to the inaccuracy of the underlying energy model, but also because there is evidence that some RNA structures can oscillate between different structures or exist in a population of structures (Christoffersen and Mcswiggen, 1994; Betts and Spremulli, 1994; Freyhult et al., 2005). It has been shown that, on average, the accuracy of secondary structure prediction algorithms increases by more than 20% when 750 suboptimal structures are generated, as opposed to generating the MFE structure only (Mathews et al., 1999).

Another important advance in MFE structures prediction was made by McCaskill (1990), who proposed a dynamic programming algorithm for calculating the partition function. The partition function is a quantity from statistical mechanics that encodes the statistical properties of a system in thermodynamic equilibrium. The partition function for the ensemble of all possible secondary structures for a given RNA sequence can be calculated using the following formula:

$$Q = \sum_{S \in \mathcal{S}} e^{-\Delta G(S)/RT} \tag{2.2}$$

where $\mathcal{S}$ is the set of all structures for the given RNA sequence, $\Delta G(S)$ is the free energy of the structure $S$, $R$ is the physical gas constant with the value $R = 1.987\ cal \cdot K^{-1} \cdot mol^{-1}$, and $T$ is the temperature.

The partition function algorithm allows calculation of basepairing probabilities within the thermodynamic ensemble of structures. It has been

shown that these basepairing probabilities provide measures of confidence for MFE structure prediction (Mathews, 2004) and that they are less affected by uncertainties in energy parameters than is MFE structures (Layton and Bundschuh, 2005).

There are other approaches to RNA secondary structure prediction that are not based on MFE computation. One example is *Sfold* (Ding and Lawrence, 2003; Ding et al., 2005, 2006), which samples structures according to their probabilities derived from Boltzmann statistics. The derived structures are further clustered according to structural similarity, and a small number of centroids is returned that can be taken as a representative ensemble of potentially relevant structures.

The preceding discussion about the computational methods for RNA secondary structure prediction considered the availability of a single RNA sequence. When several related RNA sequences from different organisms are available, it is possible to predict conserved secondary structure using comparative structure analysis. This approach is discussed further in Section 4.4.

## 2.3 Secondary structure of pre-mRNAs and its effects on gene splicing

The fact that splice sites are not specified unambiguously by primary sequence prompted scientists to speculate about the effects of higher-order RNA structure on gene splicing. One of the early experiments conducted by Solnick (1985) showed that when an exon is sequestered (i.e., isolated from the rest of the pre-mRNA) within a loop of potential RNA structure, it is omitted from the mature mRNA. Similar experiments indicated that sequestering of a 5' splice site can have a negative effect on splicing (Eperon et al., 1988). Goguel et al. (1993) went slightly further in their experiments, determining the size of the stem containing the 5' splice site that is needed for splice site inhibition to be observed ($>$ 9 nt). They observed a similar splicing effect if the branchpoint sequence was enclosed in a 15-nt long

stem. Experiments *in vivo* showed a weaker effect on splicing. A study in a plant, *Nicotiana plumbaginifolia*, revealed that 18-24 nt RNA hairpin loops strongly inhibit splicing when they sequester the 5' splice site or are placed within a short intron (Liu et al., 1995). 3' splice sites that are sequestered within such structures are still utilized, and for some introns efficiency of splicing is improved if hairpins are present. A somewhat contradictory observation emerged from a study by Lin and Rossi (1996), where artificially introduced stem/loop structures inserted 5 nt downstream of the 3' splice site in the ACT fusion gene abolished splicing *in vivo*. A potential explanation for this effect is that single-stranded sequences in this region are required for effective involvement with U5 snRNA (Newman and Norman, 1992).

It was also shown that pre-mRNA has a positive effect on splicing by shortening the effective distance between the 5' splice site and the branchpoint to the optimal distance. Experiments on several *Saccharomyces cerevisiae* genes identified complementary sequences in the vicinity of 5' splice sites and branchpoints whose base-pairing interactions are essential for splicing efficiency (Newman, 1987; Goguel and Rosbash, 1993; Libri et al., 1995; Charpentier and Rosbash, 1996; Howe and Ares, 1997). Existence of a similar intronic intra-molecular structure in the yeast U3A snoRNA precursor was confirmed by comprehensive structural probing of the molecule (Mougin et al., 1996). This stem/loop structure is also conserved in the second U3 snoRNA gene in *S.cerevisiae*, even though the intronic sequences between the two genes are significantly different (Brulé et al., 1995). Similar secondary structure interactions were also identified in *Drosophila melanogaster's Adh* intron 1: mutational analysis showed that either the disruption of the identified stem or its stabilization resulted in reduced splicing efficiency (Chen and Stephan, 2003). This structure was found to be conserved in the *Drosophila* subgenus, and it has been proposed that its role is to force the branchpoint sequence downstream into an unpaired conformation.

Functional stem regions were also found between the branchpoint sequence and the 3' splice site: these structures have been shown to have an important shortening effect on unusually long distances between these two

splicing elements (Gattoni et al., 1988; Chebli et al., 1989) and also to enable utilization of a distant 3' splice site by sequestering the closer, alternative one (Deshler et al., 1989).

A recent study by Martinez-Contreras et al. (2006) indicates that intronic secondary structure interactions may be important for efficient splicing of long mammalian introns. The authors observed that the insertion of two hnRNP A1 protein binding sites at the end of an artificially enlarged mammalian intron increased splicing four-fold. It was suggested that two bound hnRNP A1 proteins interact, thus causing the intron to 'loop-out'. Replacing these binding sites with 20-nt inverted repeats had a similar effect on splicing, indicating that looping-out of introns is important for efficient splicing. The observed phenomenon is in agreement with the role of secondary structure interactions observed in yeast, and the author suggests that yeast secondary structure interactions are substituted by hnRNP protein interactions in mammals. This is supported by over-representation of hnRNP-binding-site-like motifs at the ends of mammalian introns.

Changes in pre-mRNA secondary structure can also be the cause of human genetic diseases: for example, exclusion of exon 7 in human SMN1 and SMN2 genes, caused by disruption of 24-nt stem/loop structure in intron 7, is the cause of spinal muscular atrophy (Miyaso et al., 2003). The secondary structure serves as a splicing enhancer that binds some uncharacterized protein factors. There is also a competing theory which argues that primary structure changes, more specifically, loss of a sequence based splicing enhancer or gain of a splicing silencer, are responsible for exclusion of exon 7 in SMN2 (Cartegni and Krainer, 2002; Kashima and Manley, 2003).

The secondary structure of pre-mRNA has been shown to play a role in autoregulation of expression of some proteins: if production of a protein is in excess, the protein binds to its own pre-mRNA or mRNA and prevents splicing or translation. One example is the *S. cerevisiae* protein RPL30 (formerly known as L32). This protein regulates the splicing of its own gene by binding to a stem/loop structure that is formed between complementary sequences at the 5' end of the pre-mRNA transcript and at the 5' splice site of its only intron (Eng and Warner, 1991). It was proposed that RPL30

binding stabilizes the stem and prevents binding of U1 snRNA to the 5' splice site. An example of a different splicing autoregulatory function was observed in Yra1p, a small yeast RNA binding protein (Preker and Guthrie, 2006). If in excess, Yra1p is toxic to the organism; steady-state levels of the protein are essential for viability. This is accomplished through positive and negative splicing regulation. The YRA1 gene has one large intron (776 nt) and a non-canonical branchpoint sequence (GACUAAC), both of which are negative regulators of its splicing (shortening the intron or mutating the branchpoint sequence to the canonical one improves the splicing but has a negative effect on cell growth). The YRA1 intron contains a stem/loop structure that brings the 5' and 3' splice sites closer together and acts as a positive regulator of intron splicing. This stem is evolutionarily conserved in all budding yeast species. The disruption of the stem leads to reduced splicing levels and improved cell growth.

Pre-mRNA secondary structure is also found to have a regulatory effect on alternative splicing. For several cases of alternative splicing events it has been proven that secondary structure elements suppress expression of some exons in certain tissues, while they are normally expressed in others (Libri et al., 1991; Clouet d'Orval et al., 1991; Blanchette and Chabot, 1997; Coleman and Roesser, 1998; Hutton et al., 1998). An interesting example is the inclusion stem (iStem), a long-range RNA structure element, found in *Drosophila melanogaster's Dscam* gene (Kreahling and Graveley, 2005). The *Dscam* gene theoretically encodes 36016 different proteins due to alternative splicing of 95 of its 115 exons. The alternatively spliced exons are organized in four clusters and the iStem, which is found in the intron preceding cluster 4, is required for efficient inclusion of all the exons in this cluster. The iStem is a large stem/loop structure, with a 27-nt long stem and 275 nt in the loop, which is also conserved in other species of the *Drosophila* subgenus. The function of the iStem is not precisely known, but it is suspected that it serves as a binding site for some protein factors.

## 2.4 Using secondary structure information for splice site prediction

The numerous examples of RNA secondary structure affecting splicing prompted scientists to consider secondary structure elements as additional identifiers of spliceosomal signals. This hypothesis was tested for acceptor splice site prediction by Patterson et al. (2002) using a machine learning approach. They used Martin Reese's benchmark dataset for the evaluation of gene-finding algorithms (Reese et al., 1999) to extract 100-nucleotide-long subsequences centred around the acceptor splice sites (positive examples) and 100-nucleotide-long subsequences centred around AG dinucleotides not annotated as splice sites (negative examples). The resulting dataset contained 3960 subsequences, with the same number of positive and negative examples, and was used for training and testing in 10-fold cross-validation experiments. For each subsequence in the dataset, a comprehensive set of foldings was obtained using the mfold RNA secondary structure prediction algorithm (Zuker and Jacobson, 1998). The foldings, annotated by their free energy, were used to calculate three structural metrics: optimal folding energy, max helix, which is the probability of helix formation in a close neighbourhood, and neighbour pairing correlation model (NPCM), which was used to form an aggregate model of the structure by training two Markov models for positive and negative examples and using them to score sequences in the test dataset. The metrics were aggregated in feature vectors and used to train support vector machines (SVMs) and decision trees. SVM and decision tree classification was used to enhance the accuracy of traditional sequence-based splice site prediction methods. They achieved a 5-10% reduction in error rate, compared with strictly sequence-based approaches.

Another, more recent approach used nucleotide base-pairing information to predict splice sites in *Saccharomyces cerevisiae* introns (Marashi et al., 2006b) using a neural network approach. The authors selected 154 intron-containing yeast genes and predicted their MFE structures using the mfold algorithm. They considered 20-nt windows around donor and acceptor splice sites as positive examples and 20-nt windows around non-

splice-site GU and AG nucleotides as negative examples. Instead of using the traditional 4-letter RNA alphabet ({A,C,G,U}) they used an extended 8-letter alphabet, which considers if a nucleotide is basepaired or not ($\{A_S, C_S, G_S, U_S, A_L, C_L, G_L, U_L\}$, L = loop, S = stem). These 20-nt feature vectors were used for training and testing of separate three-layer-based perceptron neural networks for donor and acceptor sites. Half of the splice site instances were used for training and the other half for testing and cross-validation. Using this simple structure information improved the prediction accuracy of splice site prediction: the correlation coefficient for donor site prediction was 0.98 when structural information was used and 0.89 when only sequence information was used. For acceptor sites, the respective correlation coefficient values were 0.70 and 0.57.

The results from these studies show that using secondary structure information in addition to sequence-based measures leads to improved accuracy of splice site prediction. It also provides indirect evidence for the role of pre-mRNA secondary structure in gene splicing.

# Chapter 3

# Intron structure and splicing in *Saccharomyces cerevisiae*

Good biological datasets are essential for most types of bioinformatics analysis. If sequence datasets are used, it is very important that the number of sequencing errors be minimal and that sequence annotation, and the locations of gene structure elements, upstream and downstream gene regions and any additional, functionally important sequence motifs, are accurate. Any errors in data can result in faulty results and conclusions. The availability of data is also a concern. Even though the influx of biological data is enormous, sometimes it is hard to find a complete and good quality dataset for the type of analysis to be performed. Due to these concerns, we decided to conduct our research on *Saccharomyces cerevisiae*, the simplest intron-containing organism.

*Saccharomyces cerevisiae*, also known as brewer's or baker's yeast, was the first eukaryote to have its genome fully sequenced (Goffeau et al., 1996). The genome of this unicellular organism contains 12 million bases, divided among 16 chromosomes. Initial sequence annotation found 5885 potential protein-encoding genes, but this number is constantly being updated (Velculescu et al., 1997; Kowalczuk et al., 1999; Blandin et al., 2000; Wood et al., 2001; Kellis et al., 2003). The current number of open reading frames (ORFs) is 6604, of which 4412 are 'verified' ORFs, meaning that there exists experimental evidence that a gene product is produced in *S. cerevisiae* (SGD, July 2006).

Despite the differences in exon/intron structure between yeast and higher eukaryotes, and considering the universality of splicing among all intron

containing organisms, *S. cerevisiae* is often used as a model organism for splicing studies. This is due to its intrinsic advantages as an experimental system, given the possibility of a controlled growth environment and the simplicity of genetic manipulation (Goffeau et al., 1996). We chose to base our research on yeast for its thorough, experimentally supported annotation, large number of splicing studies and limited intron sizes.

## 3.1 *S. cerevisiae* introns and splicing

Unlike most eukaryotic genomes, the yeast genome has few introns. In August 2003, when we collected our data, the number of genes that were annotated to contain spliceosomal introns in the Ares lab Yeast Intron Database (`http://www.cse.ucsc.edu/research/compbio/yeast_introns.html`) (Grate and Ares, 2002) was 239. Although few in number, intron-containing genes produce the lion's share of yeast mRNA: more than 10000 of the nearly 38000 mRNA molecules made each hour are derived from genes that have introns (Ares et al., 1999). *S. cerevisiae* introns are usually limited to at most one per gene: 229 intron-containing genes in the Ares Database have only one intron, 8 contain 2 introns and 2 have several alternatively spliced intron variants. Yeast introns are shorter on average than introns in higher eukaryotes, are primarily located near the 5' end of the gene, and have highly conserved splice sites and branchpoint sequence (Spingola et al., 1999).

### 3.1.1 Conservation of splicing signals

The consensus sequence for the donor site is GUAUGU, which is by far the most commonly used 5' splice site. There are a few other variants that are used to a much lesser extent, but the first and fifth position (<u>G</u>UAU<u>G</u>U) are invariant among all annotated introns (Spingola et al., 1999). The branchpoint sequence, UACUAAC, is highly conserved in yeast, and only small deviations from the canonical sequence are tolerated (Langford et al., 1984). This is in contrast with mammalian branchpoint sequences, which are very poorly conserved (Padgett et al., 1986). Usually, only the first nucleotide

in the consensus is variable, with few exceptions. The last four positions are invariant for all annotated introns (Spingola et al., 1999). The vast majority of 3' splice sites have the consensus sequence YAG (Y = C or U). In yeast, the polypyrimidine tract, which is typically highly conserved in higher eukaryotes (see Figure 2.2), is conserved in $\sim 65\%$ of introns (Kupfer et al., 2004) and restricted mostly to U residues (Umen and Guthrie, 1995). We also discuss *S. cerevisiae* splice site signals and branchpoint sequence in Section 6.3, focusing on their basepairing interactions with snRNAs.

### 3.1.2 Pre-mRNA splicing

Splicing of pre-mRNA in *Saccharomyces cerevisiae* is generally the same as in other eukaryotic organisms. The splicing process starts with U1 snRNP recognition of the 5' splice site and subsequent binding to it, forming what is called a 'commitment complex'. Next, U2 snRNA binds to the branchpoint sequence, followed by the assembly of the U4/U5/U6 tri-snRNP with the pre-spliceosome. The U1 snRNA gets displaced from the 5' splice site, and the U6 snRNA establishes basepairing interactions with the donor site. The basepairing interactions between the U6 and U4 snRNAs are disrupted, and U6 basepairs with the U2 snRNA to form the mature (active) spliceosome. As in other eukaryotes, the splicing event consists of two consecutive trans-esterification reactions: the first one cleaves the pre-mRNA at the 5' exon/intron junction and the intron's 5' end is ligated to the branchpoint; the second one cleaves the pre-mRNA at the 3' splice site, releasing the intron as a lariat structure, and joins the 5' and 3' exons together.

The yeast spliceosome contains more than 75 splicing protein factors, some of which are directly associated with snRNAs in small ribonucleoprotein particles (snRNP proteins) and others that are not parts of snRNPs (non-snRNP proteins). These proteins have essential roles in the splicing process: examples are the BBP protein that recognizes and binds to the branchpoint sequence prior to U2 snRNA binding, and the MUD2 protein that binds to polypyrimidine tract and 3' splice site (Brow, 2002).

While the five small nuclear RNAs are well conserved in length, sequence,

and especially structure among mammals, they are quite different in yeast (Kretzner et al., 1990). The sequence similarity between mammalian and yeast counterparts is very low (except for U6 snRNA), and three of the yeast snRNAs are significantly longer than in mammals (U2 snRNA is 6 times longer). However, careful structural analysis of these RNAs has revealed that most structural elements found in higher eukaryotes are also present in yeast snRNAs (Kretzner et al., 1990). It was shown that yeast-specific structural domains of snRNA can be deleted with little or no effect on splicing (Igel and Ares, 1988; Rymond and Rosbash, 1992).

## 3.2 *S. cerevisiae* intron dataset

In order to obtain a high quality yeast intron dataset we consulted three databases: the Ares lab Yeast Intron Database, the Yeast Intron DataBase, and the Comprehensive Yeast Genome Database. For additional information, we used the *Saccharomyces* Genome Database (SGD), which is an ultimate collection of genetic and molecular biological information about *S. cerevisiae*.

### The Ares lab Yeast Intron Database

The Ares lab Yeast Intron Database (AYID) is a searchable database that contains 239 *Saccharomyces cerevisiae* spliceosomal introns (Grate and Ares, 2002). For each intron, the following information is given: SGD feature, SGD synonyms, SGD locus and description. All these entries are linked to the SGD and MIPS (Munich Information Center for Protein Sequences) databases. FASTA files with intronic sequences with and without 50-nt flanking regions are also provided. The format of the AYID database is illustrated in Table 3.1.

Each gene in yeast has a systematic name and gene name. Systematic names are of the form YCXDDDS, where 'Y' stands for 'Yeast'; 'C' indicates chromosome number (A=1, B=2, etc.); 'X' can be either 'L' or 'R', indicating the position of the gene relative to the centromere (Left or Right);

| **SGD feature** | **SGD synonyms** | **SGD locus** | **Description** |
|---|---|---|---|
| SNR17A | | U3A snoRNA | [ares] Not in a protein ORF, but in U3A snoRNA |
| SNR17B | | U3B snoRNA | [ares] Not in a protein ORF, but in U3B snoRNA |
| YAL001C | FUN24 tsv115 YAL001C | TFC3 | RNA polymerase transcription initiation factor TFIIIC (tau), 138 kDa subunit |
| YAL003W | TEF5 YAL003W | EFB1 | Translation elongation factor EF-1beta, GDP/GTP exchange factor for Tef1p/Tef2p |
| YAL030W | YAL030W | SNC1 | Synaptobrevin (v-SNARE) homolog present on post-Golgi vesicles |

Table 3.1: An excerpt from the Ares lab Yeast Intron Database (Grate and Ares, 2002).

'DDD' is a three-digit number assigned to the gene bases on its relative position from the centromere (ORFs are numbered from the centromere to the telomere); and 'S' stands for 'W' or 'C', to indicate which strand the ORF is in (Watson – forward or Crick – reverse). The yeast gene nomenclature requires that the name consist of three letters (the gene symbol) followed by an integer (e.g., ACT1).

### The Yeast Intron Database

The Yeast Intron DataBase (YIDB) contains information about all introns encoded in the nuclear and mitochondrial genomes of *S. cerevisiae* (Lopez and Séraphin, 2000). It can be accessed at `http://www.embl-heidelberg.de/ExternalInfo/seraphin/yidb.html` (last accessed in August 2003). Introns are divided into tables according to the mechanism of excision: pre-mRNA introns, tRNA introns, the HAC1 intron, and group I and II introns. For 255 pre-mRNA introns, the following information is provided in a tabular format: ORF name, EMBL (European Molecular Biology Laboratory) database accession number, transcription frequency, partial sequence of the exon 1 (5' exon), sequences of 5' and 3' splice sites and the branchpoint as well as intron size. The entries are linked to the MIPS and EMBL databases.

| **ORF name** | **EMBL acc num** | **Trans freq** | **Exon 1** | **5' splice site** | **Branchpoint** | **3' splice site** | **Intron size** |
|---|---|---|---|---|---|---|---|
| YAL001C | L22015 | 1.1 | ggaa | gtatgtt | tttactaacga | taacgacacattgaag | 90 |
| YAL003W | L22015 | 52.3 | aagg | gtatgtt | attactaacaa | tctccttttaaaatag | 366 |
| YAL016W | L05146 | 4.5 | ctgc | gtatgtc | aatactaacgt | ataattgagtggtcag | 883 |
| YAL030W | U12980 | 2.3 | agct | gtaagta | tatactaactt | tcgtgtttatttttag | 113 |
| YAL042W | U12980 | 10.4 | gcgg | gtatgaa | agtactaacgg | aacttttcacttttag | 425 |

Table 3.2: An excerpt from YIDB (Lopez and Séraphin, 2000).

The format of the YIDB database is illustrated in Table 3.2.

#### The Comprehensive Yeast Genome Database

The MIPS Comprehensive Yeast Genome Database (CYGB) is a searchable database that contains a variety of data related to the genome of *S. cerevisiae* (Mewes et al., 2002). Its section about Hemiascomycetous yeast spliceosomal introns can be accessed at `http://mips.gsf.de/proj/yeast/reviews/intron/`, with further links to the *S.cerevisiae* intron table and a FASTA file with intron sequences (last accessed in February 2006). For 271 *S.cerevisiae* spliceosomal introns in the database, the following information is provided in a tabular format: ORF/gene name, intron name, intron length, partial sequences of exon 1 (5' exon) and 2 (3' exon), sequences of 5' and 3' splice sites, branchpoint sequence, distances from the 5' splice site to the branchpoint and from the branchpoint to the 3' splice site, reference, evidence (experimental or putative) and comments. The entries are linked to the MIPS and PubMed databases.

### 3.2.1 Dataset construction

We constructed our dataset by including introns that have consistent annotations between at least two of the three databases previously discussed. Since the vast majority of yeast intron-containing genes contain only one intron and only a few contain two introns per gene, we decided to include only the former, leaving the latter for later consideration. This was done solely to make the dataset more uniform. The number of introns found to have a consistent annotation between at least two databases was 227. Eleven of these

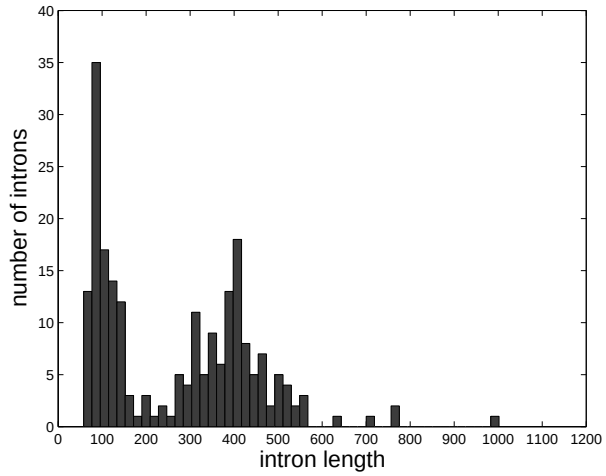
were excluded because they were not supported by the latest comparative genomic study (Kellis et al., 2003) and were marked as possible misannotations. An additional two introns, belonging to the genes YLR202C and YOR318C, have been excluded from the dataset because they were labeled as 'dubious' in SGD.

Consequently, our final yeast intron dataset contains 214 pre-mRNA introns. The consistency of annotation allows us to combine intron information from two sources in a classic database join operation, e.g., the intron sequence from the AYID database, which is not available in YIDB, and the branchpoint sequence from the YIDB database, which is not available in AYID. All of these introns are part of protein-coding genes, 95 of which code for ribosomal proteins, 84 have other, known cellular functions and 35 code for proteins of unknown function. The dataset contains 159 experimentally verified and 55 putative introns. The vast majority of introns are located in the translated portion of a gene, while 12 introns are located in the 5' untranslated region (UTR). This information was collected in January 2006 when the dataset was last updated. We will refer to this dataset as the STRuctural INtron (STRIN) dataset. Appendix A lists all the introns in the dataset (names of the genes which contain them) along with some of their characteristics.

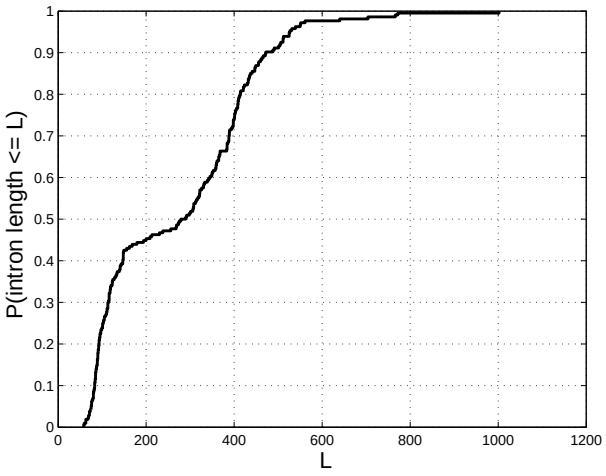
### 3.2.2 Length distribution and architecture of yeast introns

The length distribution of introns in the STRIN dataset is shown in Figure 3.1. We observe that this distribution is primarily bimodal: the first mode is located around 100 nt and the other at 400 nt. This is in agreement with previous observations of length distribution of yeast introns (Spingola et al., 1999). The shortest intron in the dataset has a length of 58 nt and the longest one has a length of 1002 nt.

The architecture of yeast introns was first examined by Parker and Patterson (1987), who tried to establish patterns of spatial arrangements between conserved sequence elements involved in splicing. The motivation for this work was the assumption that optimal spacing of conserved intron se-




(a)




(b)

Figure 3.1: (a) Distribution histogram and (b) cumulative distribution of intron lengths (L) in the STRIN dataset.

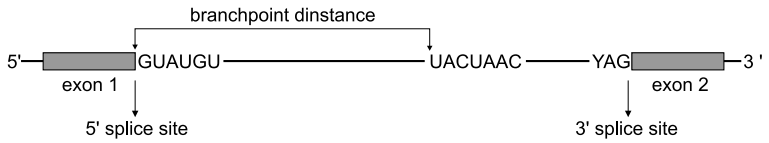



Figure 3.2: Architecture of a yeast intron: consensus 5' and 3' splice sites and branchpoint sequence are given (Y = C or U). The branchpoint distance is the distance between the 5' splice site and the branchpoint sequence. Some authors extend this distance up to the branchpoint adenine.

quences would promote spliceosome assembly by positioning snRNAs and other protein factors involved in splicing in proper geometry to interact with each other. Evidence of non-random spacing would support this assumption. For their study, Parker and Patterson used a dataset of 43 fungal introns, including 21 from *S. cerevisiae*. They distinguished two classes of introns based on the distance from the 5' splice site to the branchpoint sequence:

- 5' short (5'S) introns, for which the distance between two elements was about 40 nt, and
- 5' long (5'L) introns, with a significantly larger spacing of 200-450 nt.

Similarly, based on the distance between the branchpoint sequence and the 3' splice site, the authors proposed the following classification:

- 3' short (3'S) introns, for which the distance between two elements was between 5 and 15 nucleotides, and
- 3' long (3'L) introns, with a spacing of 22-137 nt.

A schematic of yeast intron architecture is given in Figure 3.2.

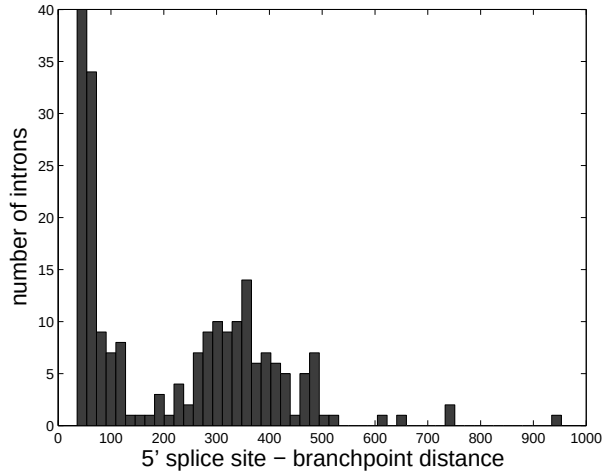
In order to see if these spatial classifications still hold for a larger set of yeast introns, we plotted distributions of distances between the 5' splice site and branchpoint sequence (which we call 'branchpoint distances') (Figure 3.3) and distances between the branchpoint and the 3' splice site (Figure 3.4). In AYID, these distances are referred to as 'lariat length' and 'tail

length', respectively. The branchpoint distance distribution appears to be bimodal, which is in agreement with the findings of Parker and Patterson (1987). Classifying STRIN introns into two groups based on their branchpoint distance yields 104 5'S introns and 110 5'L introns. 5' short introns range from 40 to 200 nt, with an average length of 115 nt and an average branchpoint distance of 71 nt, and 5' long introns range from 200 to 500 nt (with few longer exceptions), with an average length of 417 nt and an average branchpoint distance of 372 nt.

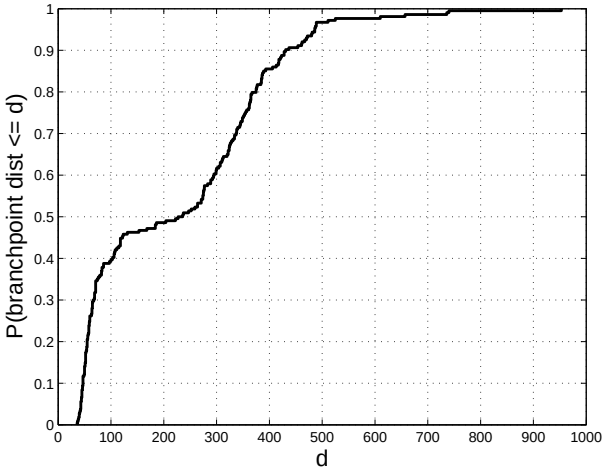
Visual comparison of the distributions in Figure 3.1 and Figure 3.3 suggests that the intron length and the branchpoint distance may be correlated in yeast introns. This is confirmed in the correlation plot given in Figure 3.5, which shows a tight linear correlation (with a Pearson correlation coefficient of $r = 0.99$) between the two intron characteristics. The same correlation coefficient value was reported in Kupfer et al. (2004), where it was found that high correlation between intron length and branchpoint distance is a common characteristic of five diverse fungi species.

The distribution of distances between the branchpoint and the 3' splice site in the STRIN dataset is unimodal (Figure 3.4), which is in contrast with the observations in Parker and Patterson (1987). This may be explained by the small size of the sample used by Parker and Patterson. The observed mode is near 35 nt, and the distances range from 10 to 80 nt, with a few longer exceptions. The correlation plot shown in Figure 3.6 does not show any evidence of a significant correlation between intron length and the distance between the branchpoint and the 3' splice site ($r = 0.13$). A similar correlation coefficient value was reported in Kupfer et al. (2004). The Pearson correlation coefficient between branchpoint distances and distances between the branchpoint sequence and the 3' splice site is $r = -0.005$.

Our analysis of spatial arrangements in the STRIN dataset supports in general the findings of Parker and Patterson (1987). The distances between conserved intron sequences are not random but grouped around one or two modes. The distance distribution for the branchpoint - 3' splice site suggests that a spacing between 10-80 nt is optimal for spliceosomal assembly, and thus is evolutionarily conserved. This is also supported by findings that the

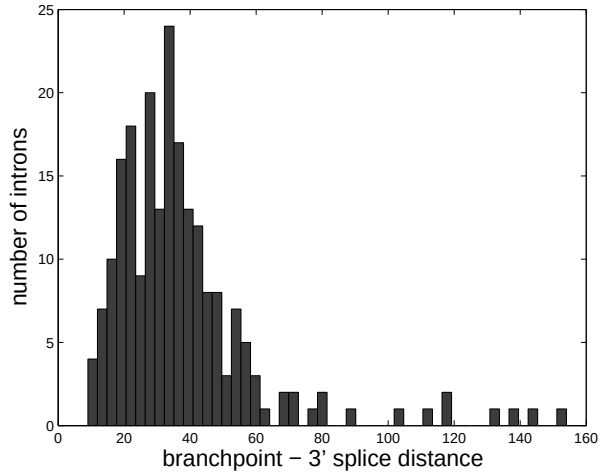



(a)

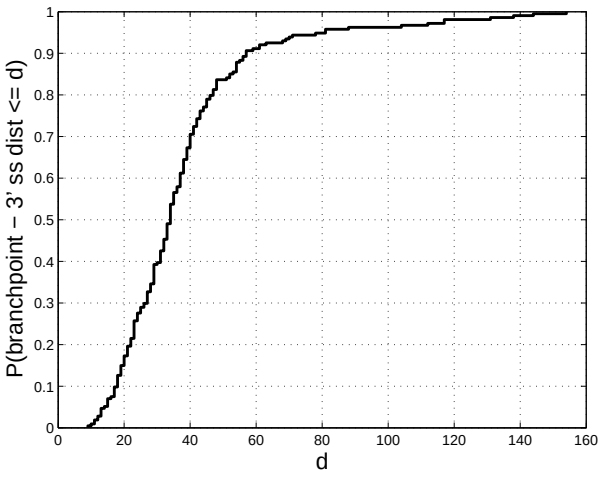



(b)

Figure 3.3: (a) Distribution histogram and (b) cumulative distribution of 5' splice site - branchpoint distances in the STRIN dataset.




(a)




(b)

Figure 3.4: (a) Distribution histogram and (b) cumulative distribution of branchpoint - 3' splice site distances in the STRIN dataset.

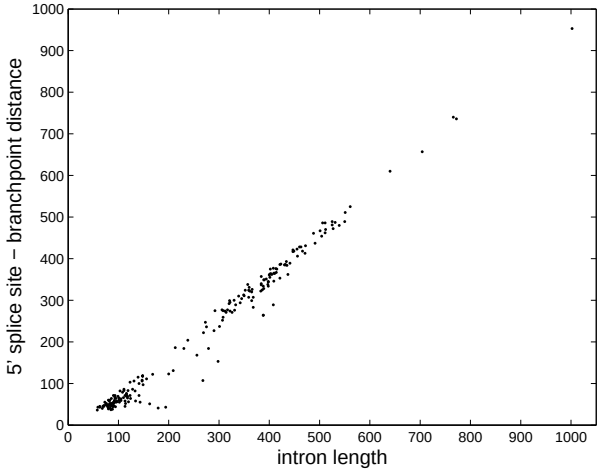



Figure 3.5: Correlation between intron length and branchpoint distance in the STRIN dataset ($r = 0.99$).

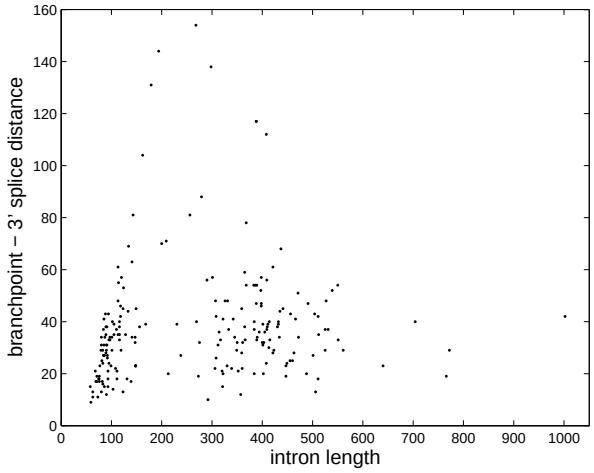



Figure 3.6: Correlation between intron length and branchpoint - 3' splice site distance in the STRIN dataset ($r = 0.13$).

distance from the branchpoint nucleotide is a critical parameter for 3' splice site activation (Luukkonen and Séraphin, 1997).

The observation of two modes for the branchpoint distribution leads us to believe that there are two optimal spacings between the 5' splice site and the branchpoint sequence. This is contradictory to our knowledge of spliceosome assembly, which is universal for any intron type. The shorter distance seems likely to facilitate direct interactions between the U1 and U2 snRNPs. The minimum distance needed for the assembly of the U1 and U2 snRNPs with pre-mRNA is around 40 nucleotides: the region bound to the U1 snRNP is approximately 15 nt (3 nt upstream of the splice junction and 12 nt downstream) (Mount et al., 1983), and the region bound to the U2 snRNP is about 35 nt (25 nt upstream of the branchpoint adenosine and 10 nt downstream) (Black et al., 1985). It was also shown that yeast pre-mRNA splicing requires a minimum branchpoint distance of 40 nt – shortening this distance leads to splicing inhibition (Thompson-Jäger and Domdey, 1987; Köhrer and Domdey, 1988). These findings are consistent with our data – the minimum branchpoint distance observed in the STRIN dataset is 42 nucleotides.

If shorter spacing is optimal for spliceosome assembly, how do the U1 and U2 snRNPs interact in 5'L introns? Parker and Patterson (1987) suggest the formation of pre-mRNA secondary structure that would bring the 5' splice site and the branchpoint into closer proximity. In their yeast dataset, they observed 12 introns with large branchpoint distances that have a potential to form helical structures between the 5' splice site and the branchpoint. A common feature of these structures is a relatively constant shortened branchpoint distance (obtained after subtracting nucleotides enclosed in the stem and loop) of around 45 nt that resembles the spacing in 5'S introns.

### 3.2.3 Secondary structure in yeast introns

Parker and Patterson's proposal is not unique. Several other authors have studied secondary structure elements in yeast introns and their effect on splicing. Experimental analysis of splicing efficiency of the *S. cerevisiae*

CYH2 gene, which involved deletion and rearrangement of intron sequences, identified two small regions, one downstream of the 5' splice site, and the other upstream of the branchpoint sequence, that were found to be essential for splicing *in vitro* and *in vivo* (Newman, 1987). The elements were found to be complementary in sequence, suggesting possible basepairing interactions.

In 1993, Goguel and Rosbash observed some communication between non-conserved sequences downstream from the 5' splice site and upstream from the branchpoint region of the RP51B *S. cerevisiae* intron. These interactions were further confirmed by Libri et al. (1995) and by Charpentier and Rosbash (1996), where comprehensive mutational and structure-probing analysis determined the exact structure of the stem-loop formed in the wild type (wt) intron. These studies also demonstrated that this complementary pairing is essential for efficient splicing *in vitro* and *in vivo*. Libri et al. found that if wild type basepairing interaction was disrupted, splicing could be restored by alternative basepairing, indicating that the existence, not the location, of the stem is essential for efficient splicing of the RP51B intron. The work of Libri et al. and Charpentier and Rosbash is discussed in more detail in Section 5.1.

A study of exon skipping in the multiply interrupted YL8A *S. cerevisiae* gene revealed the important role of complementary intron sequences that promote exon inclusion in mature RNA (Howe and Ares, 1997). The sequences are located downstream of the 5' splice site and upstream of the branchpoint, in each of the YL8A two introns. Destroying the complementarity of the sequences reduced the amount of correctly spliced pre-mRNA and induced exon skipping. Based on this observation, Howe and Ares suggest that these basepairing interactions serve as intron identifiers, promoting correct pairing of appropriate splice sites in multi-intron pre-mRNAs.

## 3.3 Intron dataset for phylogenetic analysis

Biological sequences and structures that have important functions are usually conserved during evolution. This makes comparative genomics, which compares genomes of related species, a powerful tool for identifying func-

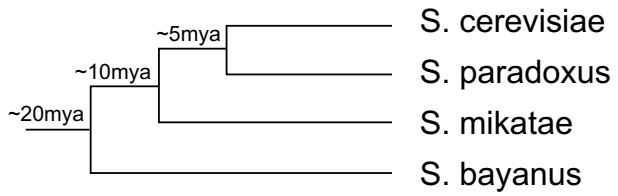



Figure 3.7: Phylogenetic tree for *Saccharomyces sensu stricto* species derived based on sequence divergence of ribosomal DNA sequences (Kellis, 2003). The time labels (in mya = million years ago) given at the nodes of the tree indicate when the species, represented by the branches coming out from a node, diverged from the most common ancestor.

tional elements without previous knowledge of function. In this thesis, we use a technique from comparative genomics, known as phylogenetic or comparative structure analysis, to look at the conservation of secondary structures or structural motifs, which we consider important for splicing.

For this purpose, we chose three species closely related to *Saccharomyces cerevisiae*: *S. paradoxus*, *S. mikatae*, and *S. bayanus*. These three species were sequenced and used for comparative studies, along with *S. cerevisiae*, by Kellis et al. (2003). The species were sequenced by seven-fold redundant coverage and assembled into 230-500 kb long scaffolds that cover most of the genome (∼95%). *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, and *S. bayanus* belong to the *Saccharomyces sensu stricto* group, and their phylogenetic tree is given in Figure 3.7.

*S. cerevisiae* has 90% (80%) average nucleotide identity with *S. paradoxus* in coding (intergenic) regions, 84% (70%) with *S. mikatae*, and 80% (62%) with *S. bayanus*. The species were found to have enough sequence similarity to allow reliable alignment of orthologous regions, but sufficient sequence divergence to allow recognition of functional elements (Kellis et al., 2003). The genomes of the four species were aligned by pair-wise comparison of the species' ORFs.

The comparative genetic analysis performed by Kellis et al. resulted in major changes in the *S. cerevisiae* gene catalogue, reducing the total number of genes by approximately 500 and redefining gene boundaries in more than

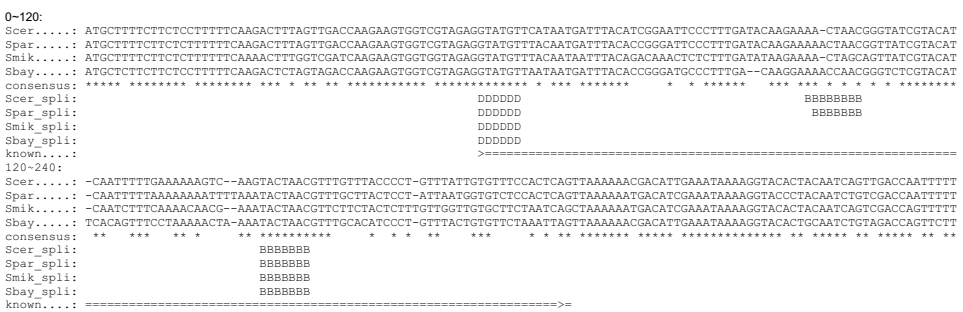



Figure 3.8: An example of a ClustalW alignment taken from the supplementary data by Kellis et al. (2003). Nucleotides conserved in all four species are marked with '*'. The intron is indicated by a double dashed line. Potential donor sites and branchpoint sequences are annotated by strings of Ds and Bs, respectively.

300 cases. They also identified 17 mis-annotated introns, which we excluded during construction of the STRIN dataset, and predicted 58 new introns.

We used multiple sequence alignments provided by Kellis et al. (2003) to extract the orthologous intron sequences. The alignments are provided as supplementary data and are available at `http://www-genome.wi.mit.edu/annotation/fungi/comp_yeasts/downloads.html` (last accessed in June 2006). They were produced by the ClustalW program (Thompson et al., 1994) and are in the format shown in Figure 3.8.

We searched for the multiple sequence alignments of introns from the STRIN dataset. There were 174 introns found in Kellis' data, which we further filtered to exclude alignments where only a *S. cerevisiae* sequence was present (YGR296W, YHL050C, YHR203C, YIL177C, YJL225C, YLR464W, YNL339C, YPL283C, YPR202W). We also examined the alignments for any inconsistencies with respect to donor, acceptor and branchpoint sequence alignments. This revealed additional problematic alignments where either the donor or branchpoint sequence was not properly aligned or the aligned sequences significantly diverged from the known consensus sequence. There were 9 intron alignments with these problems: YBR186W, YJR079W,

YKL002W, YKL157W, YLR093C, YLR211C, YNL012W, YOL047C, and YPL129W. These introns were excluded from our phylogenetic dataset. An additional four introns had problematic alignments between the *S. cerevisiae* and *S. bayanus* sequences: YGL137W, YJL177W, YLR078C, YOR182C. We excluded the *S. bayanus* sequences from these alignments. Finally, the intron YMR292W alignment between *S. cerevisiae* and *S. paradoxus* sequences was suspicious (the 7-nt sequence annotated as the branchpoint sequence for *S. paradoxus* was very different from the consensus sequence), so we excluded the *S. paradoxus* sequence from it.

These filtering steps reduced the number of alignments to 155. Among these there were 7 containing sequences from two species, 51 containing sequences from three species and 97 aligning all four species. An intron sequence from *S. cerevisiae* was present in each alignment, the *S. paradoxus* sequence was present in 146 alignments, the *S. mikatae* sequence in 120 alignments and the *S. bayanus* sequence in 134 alignments.

We calculated the average nucleotide identity in intronic regions based on these alignments: *S. cerevisiae* has 74% average nucleotide identity with *S. paradoxus*, 58% with *S. mikatae*, and 50% with *S. bayanus*.

We used these orthologous intron sequences from closely related species to investigate the conservation of intron architecture. We extracted the intron sequences from the alignments and computed intron length distributions for each species. The distribution histograms as well as correlation plots that compare intron length between *S. cerevisiae* and other species are given in Figure 3.9.

These intron length distribution histograms look very much like the distribution histogram for *S. cerevisiae* (Figure 3.1): the intron length range is almost the same and all distributions are bimodal, with the first mode located around 100 nt and the other around 400 nt. This observation as well as high correlation between intron lengths in *S. cerevisiae* and the other three species (see correlation plots in Figure 3.9; the corresponding Pearson correlation coefficients are $r = 0.97$, $r = 0.96$, and $r = 0.96$ for *S. paradoxus*, *S. mikatae*, and *S. bayanus*, respectively) indicate that intron lengths have been conserved among *sensu stricto* species.

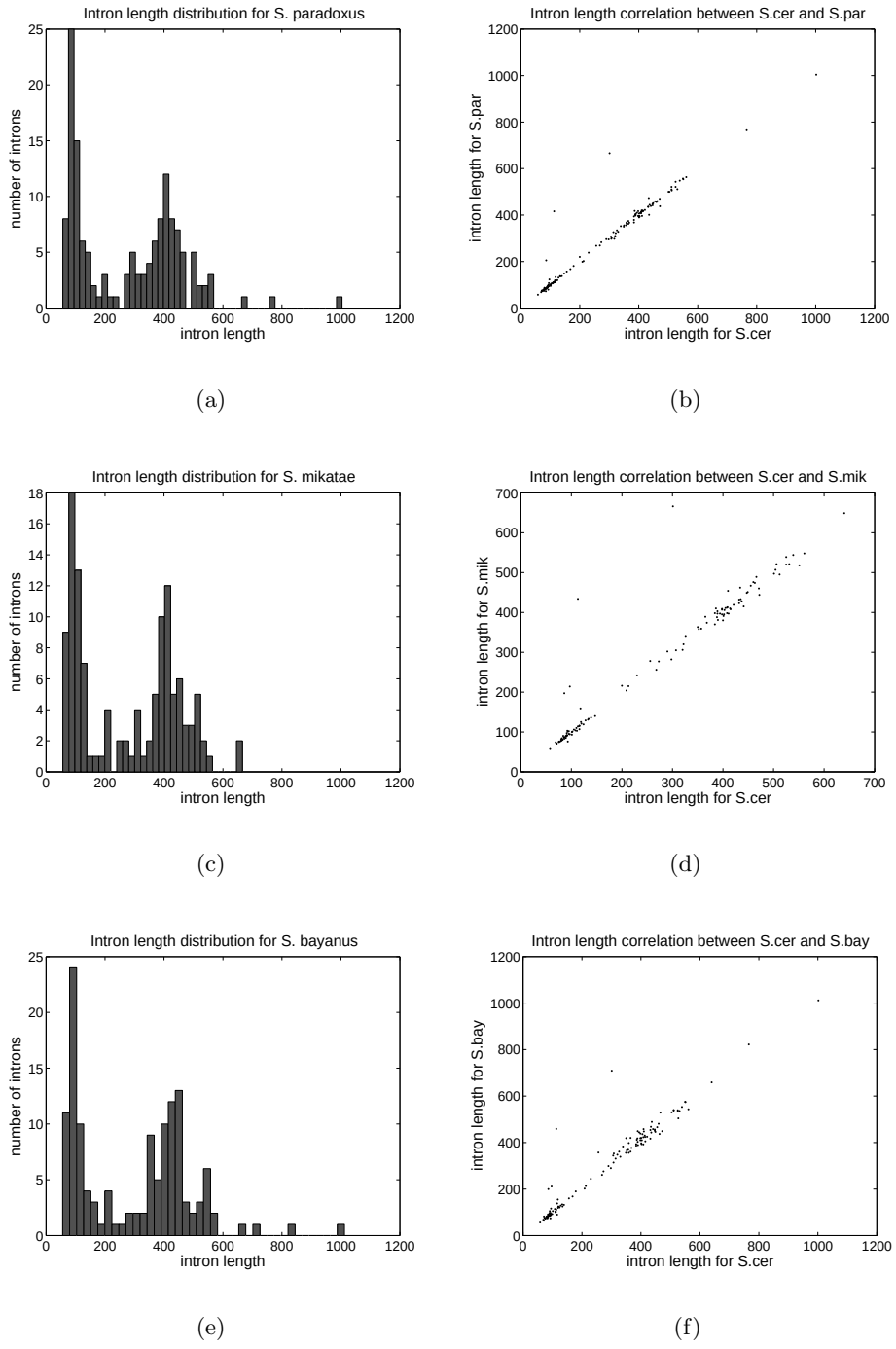



Figure 3.9: Intron length distribution histograms and intron length correlation plots between *S. cerevisiae* and *Saccharomyces sensu stricto* species: **(a), (b)** *S. paradoxus*, **(c), (d)** *S. mikatae*, **(e), (f)** *S. bayanus*.

We also looked at the conservation of branchpoint distances in the *sensu stricto* species. The branchpoint distances for the three related sequences were computed based on the multiple sequence alignments and known branchpoint distances for *S. cerevisiae* introns. The distribution histograms, as well as the correlation plots that compare branchpoint distances between *S. cerevisiae* and the other *sensu stricto* species, are shown in Figure 3.10. The correlation coefficients corresponding to the given correlation plots are $r = 0.999$, $r = 0.996$, and $r = 0.994$ for *S. paradoxus*, *S. mikatae*, and *S. bayanus*, respectively. This indicates that the branchpoint distances are conserved even better than the intron lengths and suggests functional importance of this intron characteristic. If pre-mRNA secondary structure plays a role in splicing, either to shorten the distance between the 5' splice site and the branchpoint sequence or in some other way, it seems reasonable to assume that these structural features of introns will be conserved among *sensu stricto* species. We investigate this hypothesis further in Sections 4.4 and 6.4.

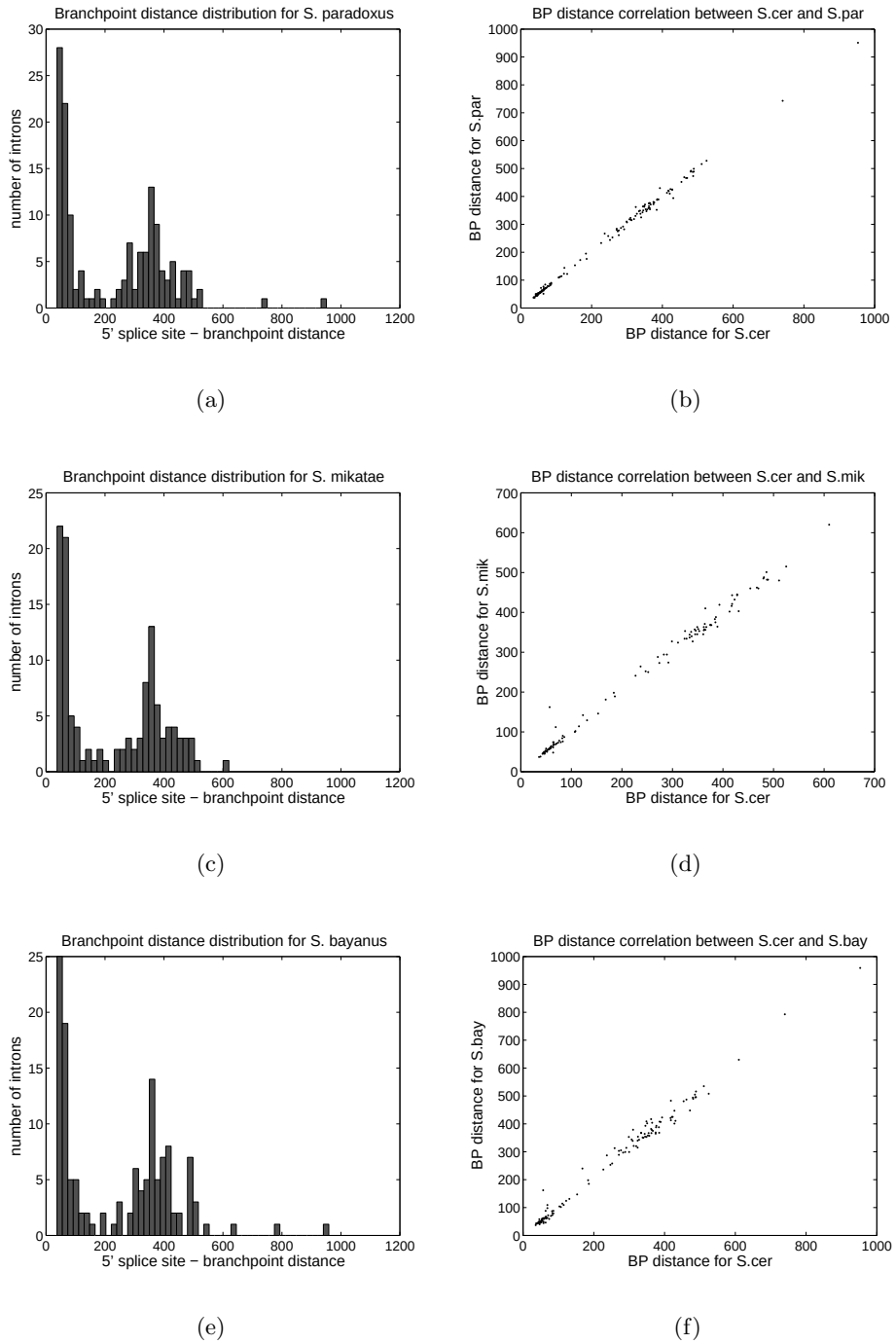



Figure 3.10: Branchpoint distance distribution histograms and branchpoint distance correlation plots between *S. cerevisiae* and *Saccharomyces sensu stricto* species: **(a), (b)** *S. paradoxus*, **(c), (d)** *S. mikatae*, **(e), (f)** *S. bayanus*.

# Chapter 4

# Zipper stems in long yeast introns

Motivated by the discoveries of complementary basepairing in several *Saccharomyces cerevisiae* introns, in this chapter we investigate if the ability to form a stem structure between the 5' splice site and the branchpoint is something common to all introns with long branchpoint distance. We search the secondary structures of 5'L STRIN introns for stems that bring the 5' splice site and the branchpoint sequence into closer proximity and analyze the effect of these stems on the resulting branchpoint distances. We experiment with different stem selection criteria and consider single and multiple stems. The results obtained generally support the hypothesis that stem structures in a long yeast intron can shorten the branchpoint distance to what is believed to be the optimal one.

The second part of the chapter investigates the conservation of zipper stems among closely related *Saccharomyces sensu stricto* species, using visual inspection of the multiple sequence alignments and secondary structures as well as automatic comparative structure approaches on a selected subset of introns.

## 4.1 Definition and initial identification of zipper stems

We used the Vienna RNA secondary structure package (Hofacker et al., 1994) to calculate secondary structures of 5'L introns in the STRIN dataset. Its RNA folding function, RNAfold, is based on a dynamic programming

algorithm, which calculates pseudoknot-free secondary structure with minimum free energy as well as the equilibrium partition function and basepairing probabilities. Version 1.4 of the package was downloaded from `http://www.tbi.univie.ac.at/~ivo/RNA/` (September 2003) and compiled. The input to the RNAfold function consists of one or more sequences in FASTA format. The output of the program is a secondary structure in dot-bracket notation, which is usually defined as follows:

**Definition 1 (RNA secondary structure in dot-bracket notation)** *Let $R = r_1 r_2 ... r_n$ be an RNA sequence ($r_k \in \{A, C, G, U\}$). The secondary structure of sequence $R$ is given by a string $S = s_1 s_2 ... s_n$, where $s_k$ ($1 \leq k \leq n$) is one of the symbols: '.', '(', and ')'. A basepair between bases $r_i$ and $r_j$, where $i < j$, is represented by $s_i =$ '(' and $s_j =$ ')'. Unpaired bases $r_k$ are represented by $r_k =$ '.'. For each position $i$ in $S$ the number of open brackets has to be greater than or equal to the number of closed brackets. The total numbers of opening and closing brackets in $S$ have to be equal.*[*]

The RNAfold function was run on a FASTA file with the 110 5'L introns from the STRIN dataset. For each intron sequence, the program calculated a structure with the minimum free energy. The structures were further computationally analyzed to find a stem that would shorten the branchpoint distance. A stem structure in RNA secondary structure prediction is formally defined in Definition 2, as follows:

**Definition 2 (Stem)** *A stem in an RNA secondary structure is defined by basepairing between two complementary subsequences of $R$: $r_i r_{i+1} ... r_{i+k}$ and $r_{j-k} r_{j-k+1} ... r_j$ and represented in secondary structure $S$ as a substring $s_i s_{i+1} ... s_{i+k}$ of opening brackets and a substring $s_{j-k} s_{j-k+1} ... s_j$ of closing*

[*]The pseudoknot-free RNA secondary structure in dot-bracket notation can also be defined by a simple context-free grammar:

$$S \rightarrow \varepsilon \mid . S \mid (S)\, S \tag{4.1}$$

*brackets, and where $s_i$ and $s_j$, $s_{i+1}$ and $s_{j-1}$, ..., $s_{i+k}$ and $s_{j-k}$ are matching brackets.*

Note that a stem as defined in Definition 2 is uninterrupted, i.e., does not allow for the occurrence of any unpaired bases that would form bulges or internal loops within the stem.

We designed an algorithm that identifies a stem that brings the 5' splice site and the branchpoint into closer proximity. The location of the branchpoint was obtained by taking the branchpoint sequence for a particular intron from the YIDB table (see Table 3.2) and finding its location in the intron sequence taken from the AYID database. The algorithm parses a string of brackets and dots given by the RNAfold program and looks at the stems that appear between the 5' splice site and the branchpoint sequence. The stem that has the largest distance between the complementary sequences forming it is selected and returned as the result. We call such a stem a 'zipper' stem, since it 'zips' the intron, bringing the 5' splice site and the branchpoint sequence closer together. The formal definition of a zipper stem is given in Definition 3. A zipper stem is always located between the 5' splice site and the branchpoint, even though these two sequences could be brought closer together by basepairing interactions between the first complementary sequence located between the 5' splice site and the branchpoint and the second complementary sequence located between the branchpoint and the 3' splice site. A generalization of the zipper stem definition that captures such effects will be considered later. An example of an intron structure, with annotated 5' splice site, branchpoint sequence and zipper stem, is given in Figure 4.1.

**Definition 3 (Zipper stem)** *A stem represented by two substrings $s_i...s_{i+k}$ and $s_{j-k}...s_j$ of matching opening and closing brackets, where $j < d$ ($d + 1$ is the first position of the branchpoint sequence), is called a zipper stem if it is the stem that brings the 5' splice site and the branchpoint closest together, i.e., if $j - i$ is maximal among all possible stems in $S$ for which $j < d$.*

Potential zipper stems were found for all 5'L introns in the dataset. Since it was previously speculated that the role of these stems would be to

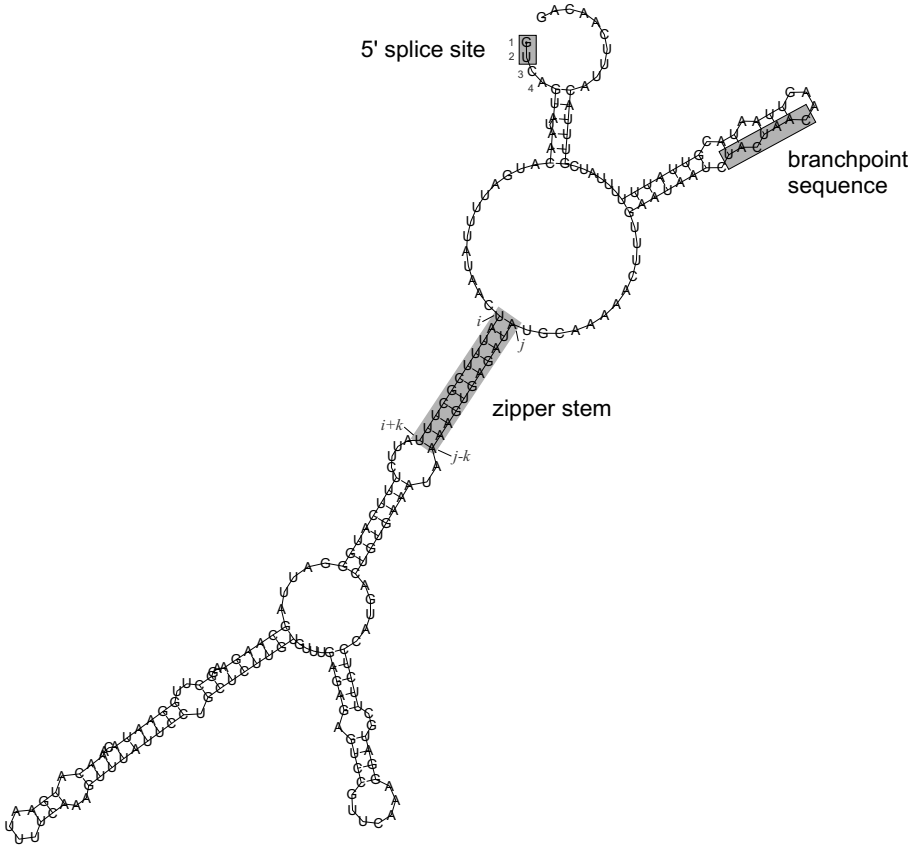



Figure 4.1: Secondary structure of the YGL030W intron. The dot-bracket notation for this structure with highlighted zipper stem is shown at the bottom of the figure.

shorten the distance between the 5' splice site and the branchpoint to the operational distance that is observed for 5'S introns (Parker and Patterson, 1987; Newman, 1987; Goguel and Rosbash, 1993; Libri et al., 1995) we tested if we could observe this effect in our dataset. Once the locations of the potential zipper stems were found, we calculated the shortened branchpoint distance for all the sequences.

**Definition 4 (Shortened branchpoint distance)** *If $s_i...s_{i+k}$ and $s_{j-k}...s_j$ are two substrings of matching opening and closing brackets that form the identified zipper stem, the shortened branchpoint distance has the following value:*

$$\bar{d} = i + (d - j) \tag{4.2}$$

*where d is the original branchpoint distance (the first position of the branchpoint sequence is d + 1); for an illustration see Figure 4.2.*

If the algorithm fails to identify a zipper stem for a certain sequence, its original, linear branchpoint distance will be included in the distribution of shortened distances.

The distribution of shortened branchpoint distances for 5'L STRIN introns, along with the distribution of branchpoint distance for 5'S introns, is given in Figure 4.3. The cumulative distribution plot is shown in Figure 4.4. The two distributions do not appear to support the hypothesized effect. The datasets have almost identical means (72 nt for 5'S introns and and 71 nt for zipped 5'L introns), but many zipped 5'L introns have branchpoint distances shorter than 40 nt or longer than 200 nt, which is not observed for the 5'S introns.

To statistically test these differences we performed the Kolmogorov-Smirnov test (KS test) (Hollander and Wolfe, 1999), which is used to determine whether the underlying probability distributions of two datasets differ. The KS test is non-parametric, i.e., it does not require any assumption about distribution of data (unlike Student's t-test) and is not dependant on data binning (unlike chi-square test), which makes it suitable for our analysis.

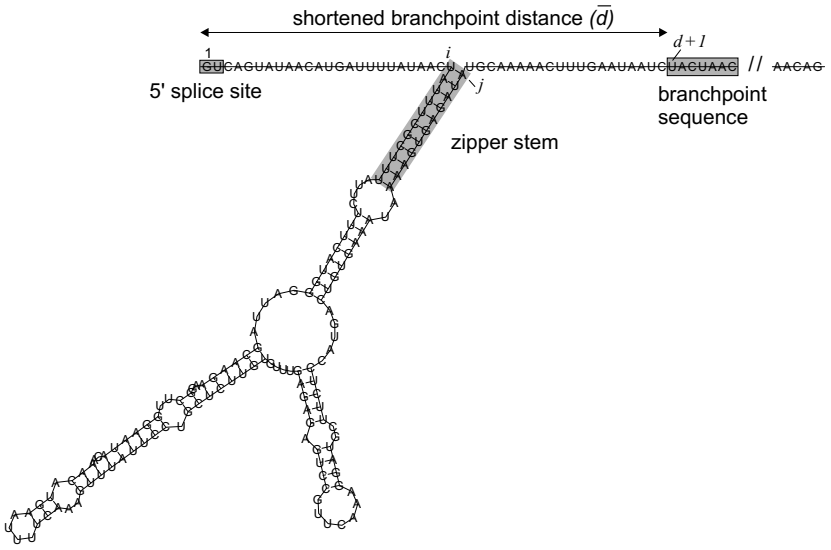


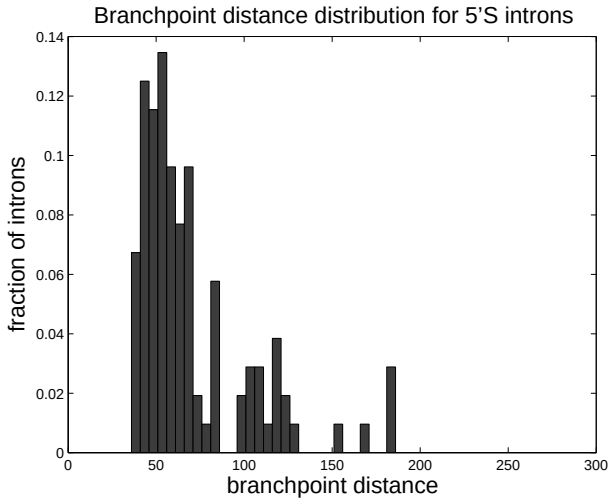

Figure 4.2: Illustration of shortened branchpoint distance.

The test calculates the maximum distance between two cumulative distributions ($D$ statistics), and then the p-value associated with this number can be obtained. For p-values smaller than the selected significance level, the null hypothesis that the two datasets stem from the same underlying distribution is rejected. We used a significance level of 0.05 for all statistical tests performed.

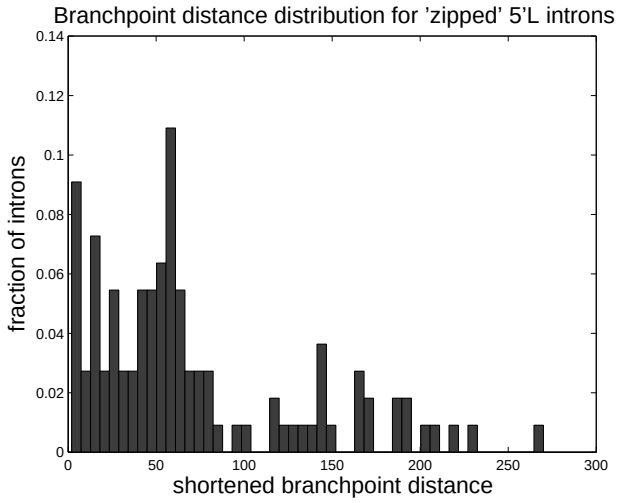
Applying the KS test to the datasets of branchpoint distances for 5'S introns and branchpoint distances for zipped 5'L introns, we obtained $D = 0.3$ with a corresponding p-value $< 0.0001$. Thus, the null hypothesis that the two datasets are of the same form is rejected.

## 4.2 Length-bounded zipper stems

Upon closer inspection of the intron structures with very short identified zipper stems, we usually found longer, more thermodynamically stable stems




(a)




(b)

Figure 4.3: Distribution histograms for (a) branchpoint distance for 5'S introns ($d$) and (b) branchpoint distance for zipped 5'L introns ($\bar{d}$).

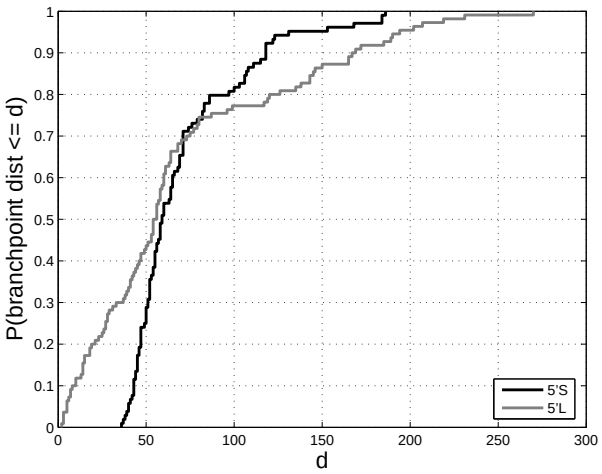



Figure 4.4: Cumulative distributions of the 5' splice site - branchpoint distance for 5'S introns and 5'L introns folded in secondary structure.

that could zip the intron between the 5' splice site and the branchpoint. One example is given in Figure 4.5. In this case the zipper stem found was only 2 nt long and is actually part of a larger stem formed by basepairing interactions between the 5' splice site and the branchpoint sequence. It is questionable if this basepairing interaction is present *in vivo* since the U1 and U2 snRNPs have to bind to these sequences in order to initiate splicing. The other possibility is that there are enzymes that open this structure previous to splicing (Wagner et al., 1998; Wang et al., 1998). In any case, even if the 5' splice site and the branchpoint were not enclosed in this stem, a 2-nucleotide-long stem does not seem thermodynamically stable enough to hold two ends of the molecule together, especially if the loop size is larger. Therefore, we modified our algorithm to find zipper stems whose length is greater than or equal to a minimum stem length defined in the algorithm. We ran the algorithm for arbitrary minimum stem lengths of 3, 5 and 7 basepairs (bp). For a stem length of at least 7 bp, the algorithm failed to find a stem in many introns from the dataset. The best results, in terms of $D$ statistics value, were obtained when the minimum length requirement for the zipper

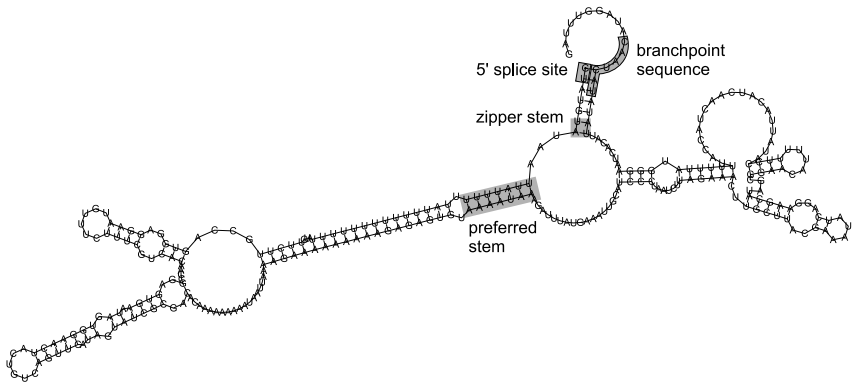


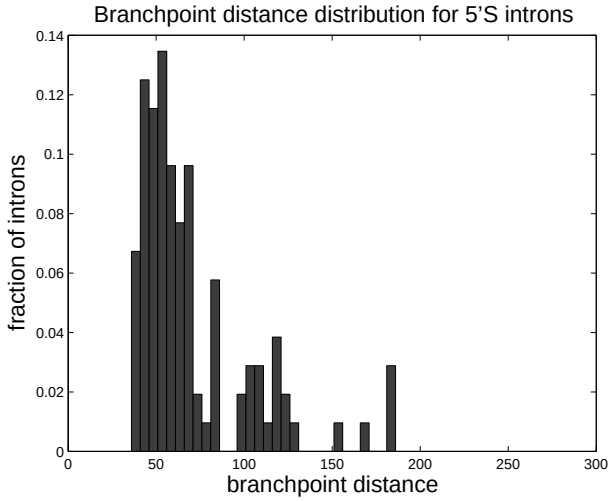

Figure 4.5: Secondary structure of the YDL079C intron.

stem was 5 bp. The distribution of shortened branchpoint distances is shown in Figure 4.6, along with the original branchpoint distance distribution for 5'S introns.

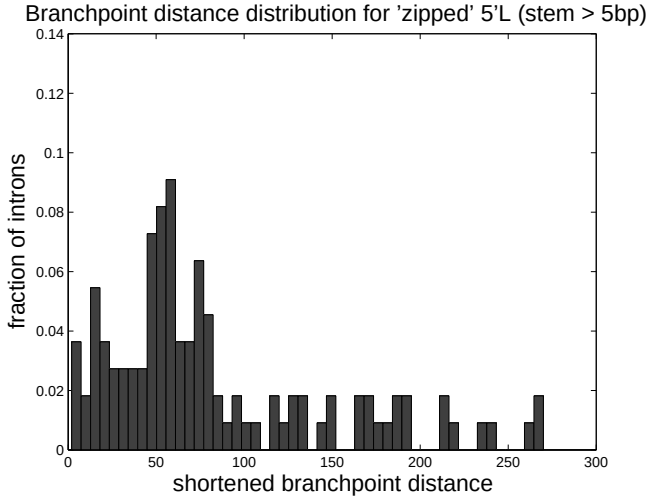
Imposing the minimum stem-length requirement did improve results, but the null hypothesis that the dataset of 5'S branchpoint distances and the dataset of shortened 5'L branchpoint distances stem from the same underlying distribution was still rejected (p-value = 0.024). Cumulative distribution plots for the two datasets are given in Figure 4.7. From Figure 4.6(b) and Figure 4.7 it is evident that the number of zipped introns that have very short branchpoint distances has been reduced; however, this also resulted in an increase of the zipped introns with longer ($>$ 200 nt) branchpoint distances.

### 4.2.1 Control datasets

Despite the differences between the two distributions, their modes are still around the same value of about 50 nt. To test the significance of this phenomenon we investigated if a similar effect can be observed for random and exon sequences. For this purpose, we generated two control datasets: a dataset of randomly generated RNA sequences and a dataset of exonic




(a)




(b)

Figure 4.6: Distribution histograms for (a) branchpoint distance for 5'S introns and (b) branchpoint distance for zipped 5'L introns with a minimum zipper stem length of 5 nt.

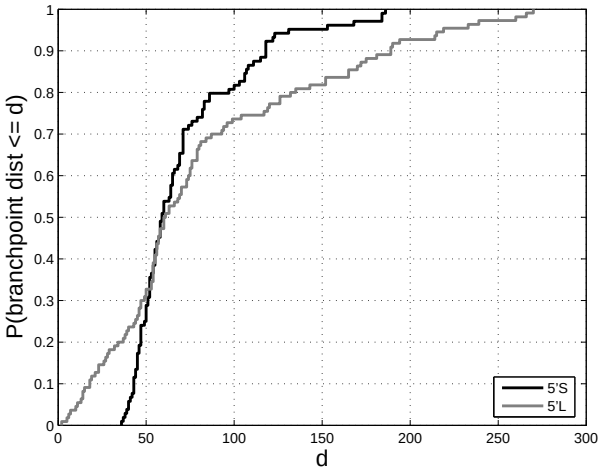



Figure 4.7: Cumulative distributions of the 5' splice site - branchpoint distance for 5'S introns and 5'L introns folded in secondary structure with a minimum zipper stem length of 5 nt.

subsequences. The first dataset contains 500 randomly generated DNA sequences that on average have the same GC content as the STRIN dataset (34%) and whose length distribution is identical to the length distribution of 5'L introns. The second dataset was generated by extracting the windows of exonic sequences from STRIN exons (exons from genes that have STRIN introns) by sliding a window of variable length, which is drawn from the 5'L intron length distribution, over the exon sequences. The resulting dataset has 449 exonic sequences that have the same length distribution as the 5'L STRIN introns.

The sequences in these control datasets were folded using RNAfold, and the obtained MFE secondary structures were processed to find potential zipper stems between the beginning of the sequence and the assumed branchpoint location. This location is the same as the branchpoint location of the intron whose length was used to model a particular random or exonic sequence. Identified zipper stems had to be at least 5 bp long. The distributions of resulting shortened branchpoint distances are plotted in Figures

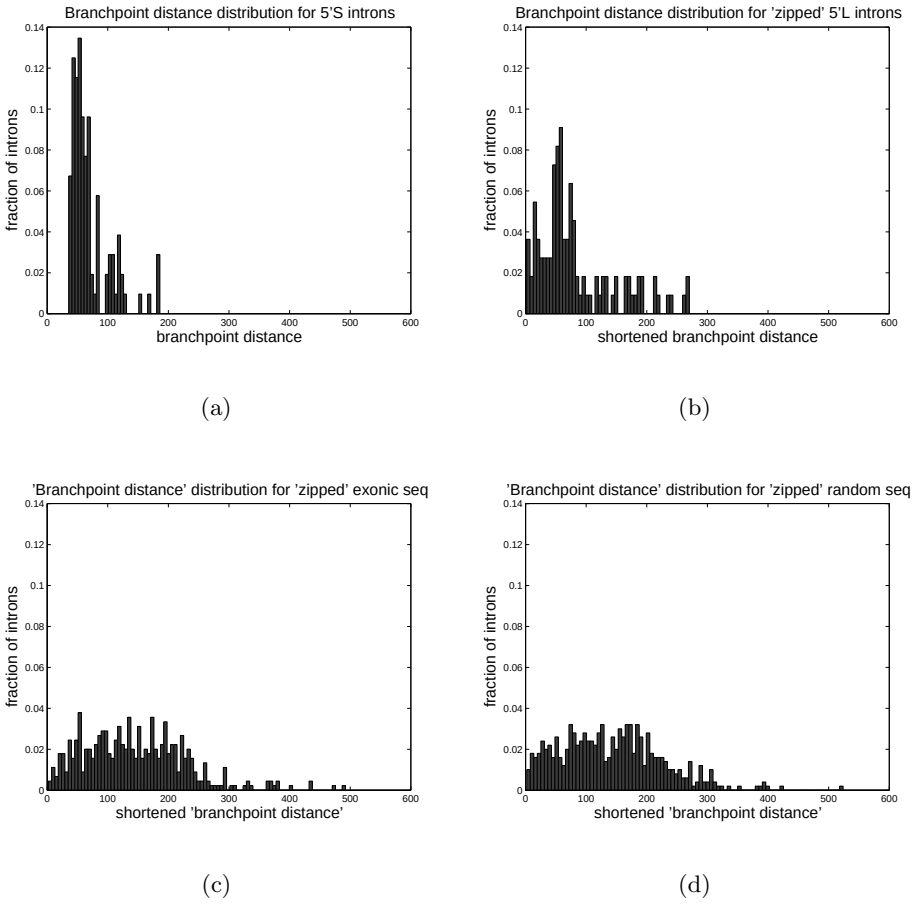



Figure 4.8: Distributions for (a) branchpoint distance for 5'S introns, (b) branchpoint distance for zipped 5'L introns , (c) branchpoint distance of zipped exonic sequences, and (d) branchpoint distance of zipped random sequences. All zipped structures have a minimum zipper stem length of 5 nt.

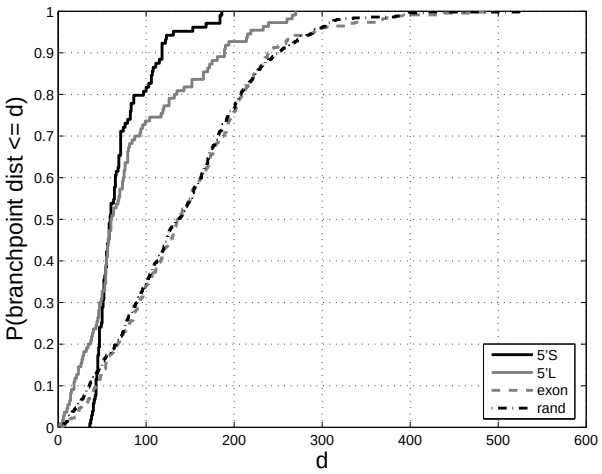



Figure 4.9: Cumulative distributions for the branchpoint distance in 5'S introns, shortened branchpoint distance in zipped 5'L introns, and shortened branchpoint distance of zipped exonic and random sequences. All zipped structures have a minimum zipper stem length of 5 nt.

4.8(c) and 4.8(d). The cumulative distributions for the branchpoint distance in 5'S introns, shortened branchpoint distance in zipped 5'L introns, and shortened branchpoint distance of zipped exonic and random sequences are given in Figure 4.9.

Figure 4.8 shows that the distributions for random and exonic sequences, which are very similar, are quite different from the other two distributions: the distribution mode at around 50 nt observed for both 5'S and 5'L introns does not exist for random and exonic sequences, whose distributions appear to be approximately uniform between 50 and 250 nt. The differences are also evident in cumulative distribution plots, shown in Figure 4.9, where the curves for random and exonic sequences overlap, while the curve for 5'L introns is closer to that for 5'S introns. Statistical analysis further emphasizes the differences between the datasets: the KS test comparing 5'S introns with exonic and random sequences produced $D = 0.5227$ (p-value $< 10^{-21}$) and $D = 0.5081$ (p-value $< 10^{-20}$), respectively. Table 4.1

| | **5'S introns** | **5'L introns** | **exon seq** |
|---|---|---|---|
| **5'L introns** | D = 0.20<br>p-value = 0.02 | | |
| **exon seq** | D = 0.52<br>p-value = $10^{-21}$ | D = 0.43<br>p-value = $10^{-15}$ | |
| **random seq** | D = 0.51<br>p-value = $10^{-20}$ | D = 0.41<br>p-value = $10^{-14}$ | D = 0.03<br>**p-value = 0.95** |

Table 4.1: Summary of KS test results for all pair-wise comparisons between datasets of STRIN 5'S introns, STRIN 5'L introns, exonic sequences and random sequences. The p-value highlighted in boldface is greater than 0.05, indicating that the hypothesis that two compared datasets stem from the same distribution cannot be rejected.

summarizes the results of the KS test for all pair-wise comparisons between the four datasets.

It is necessary to point out that random and exonic sequences are different by nature, mostly because of the evolutionary pressure imposed on exon sequences, and that for some other characteristics they would probably exhibit profound differences (e.g., codon frequencies). The high similarity between the distributions for exonic and random sequences further emphasizes the non-randomness of the distribution mode for the shortened branchpoint distances observed in 5'L introns. One possible implication of the mode phenomenon is that the optimal branchpoint distance for splicing of yeast introns is about 50 nt, which in long introns, as our results suggest, is achieved by shortening of the original distance as a result of zipper stem formation.

### 4.2.2 Multiple zipper stems

Motivated by the considerable number of 5'L shortened branchpoint distances longer than 150 nt, which is not observed for 5'S branchpoint distances, we investigated the possibility that more than one zipper stem is involved in shortening the distance between the 5' splice site and the branchpoint. For this purpose, we modified our algorithm to look for an ensemble of zipper stems in the following way: once a zipper stem according to Defi-

nition 3 is found, the secondary structure $S$ is modified to exclude that stem (bases in the stem are marked as unpaired), and the search continues for the next zipper stem that brings the 5' splice site and the branchpoint sequence closest together. This process is repeated until no stems can be found that are larger than or equal to the minimum stem size defined in the algorithm. The shortened branchpoint distance of an intron zipped with several stems is calculated similarly as described before by counting the nucleotides between the 5' splice site and the branchpoint that are not enclosed in or between complementary sequences of the found zipper stems. An example is given in Figure 4.10, which shows four zipper stems identified by the algorithm. The black letters, which are not within the highlighted regions, are the nucleotides that are included in distance calculation.

The algorithm was run for the arbitrary minimum stem lengths of 3, 5, and 7 nucleotides and the best results, with respect to $D$ statistics values, were obtained for the minimum stem length of 5 bp. The shortened branchpoint distribution for 5'L introns, and the branchpoint distribution for 5'S introns, are shown in Figure 4.11.

Again, the modes of the two distributions are located at around 50 nt, but the distribution of the data for the lower values in each case is quite different, with many more sequences having very short branchpoint distances for the zipped 5'L introns. This observation is also supported by a KS test, which resulted in rejection of the null hypothesis that the two samples stem from the same underlying distribution (p-value $< 0.001$). The analysis was repeated with the two control datasets, this time resulting in distributions of shortened branchpoint distances for random and exonic sequences that were more similar to the original 5'S branchpoint distance distribution than the 5'L shortened branchpoint distance distribution ($D = 0.1971$ and $D = 0.1864$ for the exonic random sequences, respectively, p-value $< 0.001$ for both). However, similar to single-stem analysis, the distance distributions of control datasets appear more uniform, without a prominent mode, unlike the the distributions for 5'S and 5'L introns. This phenomenon can be observed in the cumulative distribution plot given in Figure 4.12.

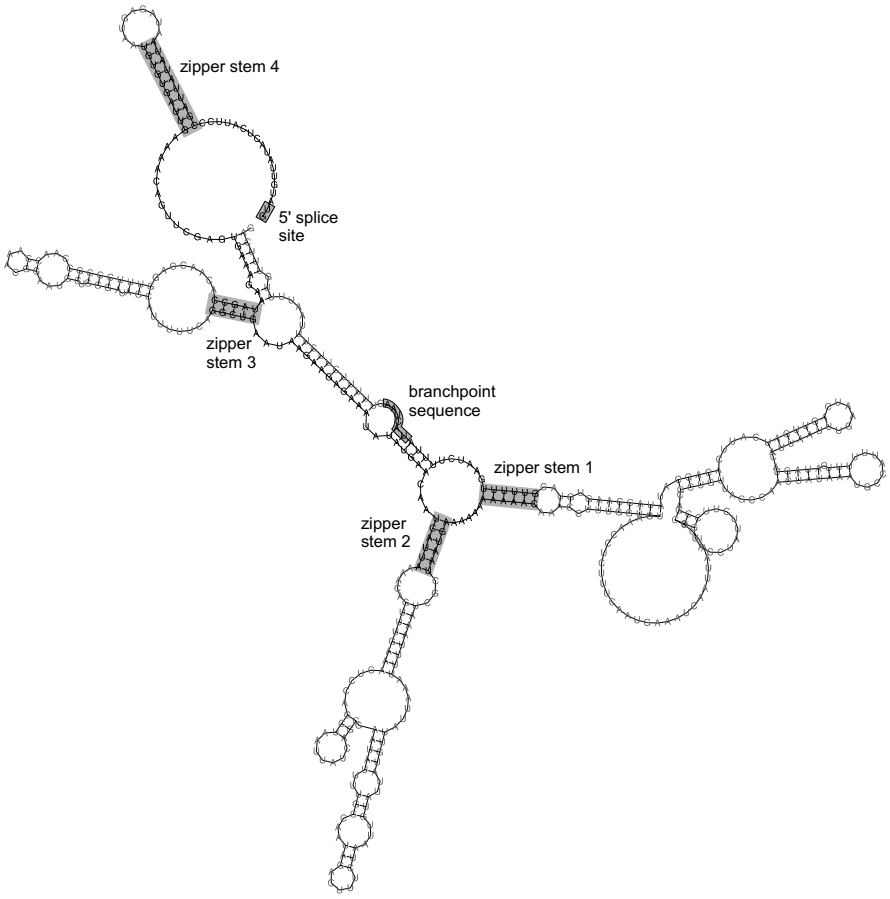
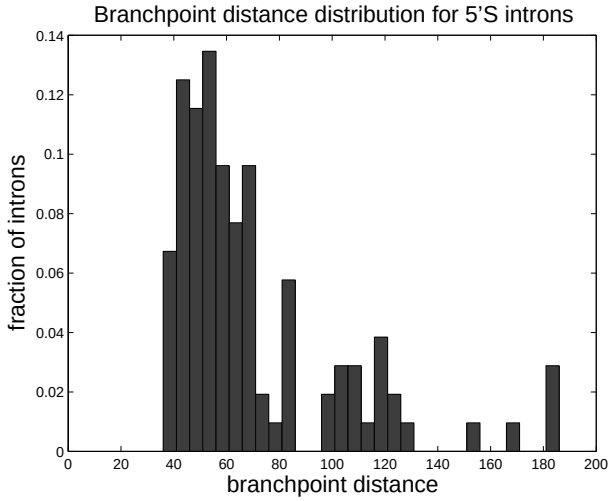
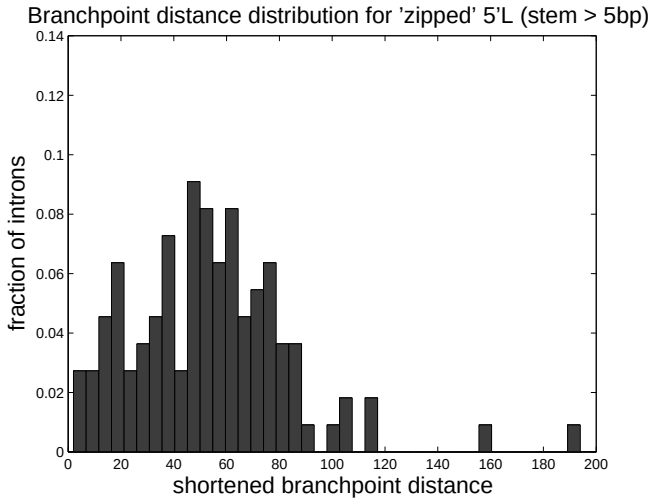



Figure 4.10: Secondary structure of the YGR214W intron. The four zipper stems are enumerated in the order by which they were identified by the algorithm. The shortened distance is calculated by counting nucleotides between the 5' splice site and the branchpoint sequence that are not enclosed in or between complementary sequences of the found zipper stems (black letters).




(a)




(b)

Figure 4.11: Distributions for (a) branchpoint distance for 5'S introns and (b) branchpoint distance for 5'L introns zipped with one or more stems with minimum length of 5 nt.

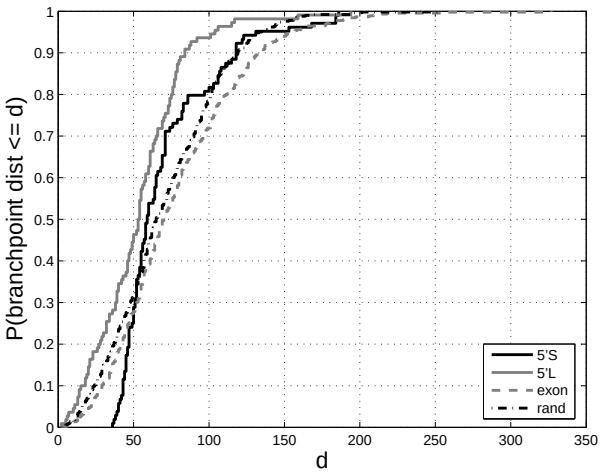



Figure 4.12: Cumulative distributions of the 5' splice site - branchpoint distance for 5'S introns and 5'L introns, random and exonic sequences zipped with one or more stems with minimum length of 5 nt.

In summary, in this section we analyzed the shortened branchpoint distances for STRIN 5'L introns that were zipped by one or more zipper stems, with or without the minimum stem-length requirement. We found that in the case of zipper stems with the minimum stem-length requirement, considering a single stem yields better results than when there are no constraints on zipper stem selection. The obtained shortened branchpoint distances are reasonably similar to the original branchpoint distances in 5'S introns, with the most common distance being $\sim$ 50 nt, but there are a number of distances that are outside of the optimal range. The corresponding distances for the two control datasets were found to be more uniformly distributed, which indicates that zipper stems in long introns have a specific effect on the branchpoint distances.

## 4.3 Thermodynamically stable zipper stems

As explained earlier, the zipper stems identified by our program have to be uninterrupted (Definition 3) and to satisfy a minimum length requirement. However, a stem length is just a crude approximation of its thermodynamic stability. It is possible to directly calculate a stem's thermodynamic stability, i.e., its free energy, using Turner's energy model (Freier et al., 1986; Turner et al., 1987; Turner and Sugimoto, 1988; Mathews et al., 1999). This model is described in detail in Section 2.2.

While Turner's energy model is usually used to calculate the free energy of an entire RNA molecule as well as for finding the minimum free energy structure for a given RNA sequence, it can also be used to calculate the free energy of some parts of an RNA structure. We used this approach to improve our method for zipper stem identification.

### 4.3.1 Identification of thermodynamically stable zipper stems

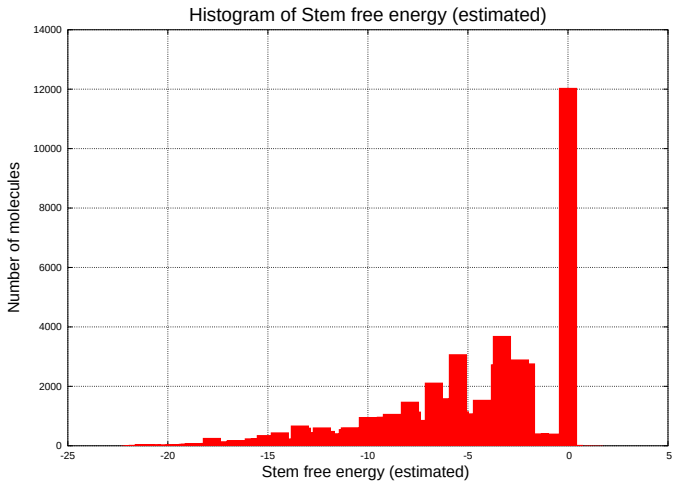
While our goal of finding the stem that maximally zips the intron (i.e., brings the donor site and the branchpoint sequences closest to each other) remains unchanged, we next consider different requirements for the zipper stems. As RNA stems in nature are often interrupted by internal loops and bulges, we modified our algorithm to also consider this type of stem. We accomplished this in the following manner: the algorithm first searches for the basepair $s_i s_j$ that has the maximal distance between the bases involved, i.e., $j - i$ is maximal among all possible basepairs found between the donor site and the branchpoint sequence. Once such a basepair is found, a subroutine is called to identify the stem containing that basepair (the basepair will be the first basepair in the stem).

The subroutine has one input parameter, the loop threshold ($t_l$), that controls the maximum allowable size of internal loops and bulges: an internal loop or bulge will be included as part of the stem only if its number of free bases is less or equal to $t_l$. In this way, stems with unrealistically large loops/bulges that would destabilize them are not allowed. The free

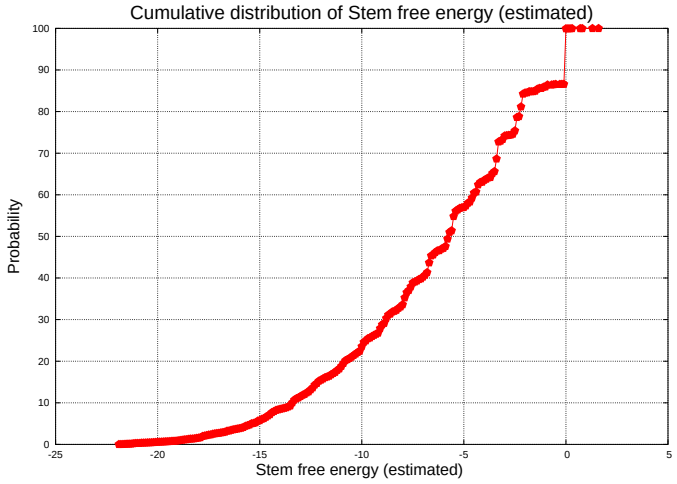
energy of a loop is roughly proportional to its size, with larger loops having higher free energy (Zuker and Jacobson, 1998). Thus, using a loop size threshold has essentially the same effect as using a loop energy threshold. The subroutine keeps extending the potential zipper stem, starting from the initial basepair, until it reaches a multi-loop or an internal loop or bulge whose number of free bases exceeds the loop threshold. The subroutine returns the last basepair that is considered part of the potential zipper stem along with the free energy of the stem. Once the free energy of the stem has been calculated, it is compared to the energy threshold ($t_e$), and if it is lower than this threshold, the stem is predicted as the zipper stem. If the stem is rejected, the RNA secondary structure is modified to exclude the found stem and the search is repeated. Once a thermodynamically favorable stem is found, the shortened branchpoint distance is calculated in the same manner as before.

Since the optimal free energy of a stem that is supposed to zip the intron is unknown, we performed a small empirical study to determine the range of possible values. To get some initial idea about the free energies of stems in naturally occurring RNA molecules, we used the RNA SSTRAND (RNA Secondary Structure and Statistical Analysis Database) database (Andronescu et al.) at `http://www.rnasoft.ca/sstrand/`. When last accessed (June 2006), this database contained 3356 RNA secondary structures on which users can perform various types of statistical analyses. Using the database interface, we obtained the distribution of free energies of stems over all RNA molecules in the database. Note that only continuous stems are considered for this analysis (Andronescu, 2006).

The distribution histogram in Figure 4.13 shows that the free energy of a typical stem without mismatches ranges between $-2$ and $-20$ kcal/mol. The large number of stems with a free energy of 0 kcal/mol corresponds to isolated basepairs (Andronescu, 2006). Guided by this distribution we have tried several values for the free energy threshold and analyzed the zipper stems found along with the corresponding shortened branchpoint distances. For $t_e > -5$, the zipper stems found are usually very short and appeared inadequate to hold the two ends of the intron together. There is usually a




(a)




(b)

Figure 4.13: Distribution histogram **(a)** and cumulative distribution plot **(b)** of free energies of stems in naturally occurring RNA molecules (obtained from SSTRAND database).

| $t_e \backslash t_l$ | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
|---|---|---|---|---|---|---|---|
| **−5** | 0.003 | 0.003 | 0.002 | 0.002 | 0.002 | 0.002 | 0.002 |
| **−7** | 0.019 | 0.029 | **0.055** | **0.055** | 0.038 | 0.017 | 0.017 |
| **−10** | $< 0.001$ | $< 0.001$ | 0.008 | 0.008 | 0.013 | 0.019 | 0.019 |
| **−12** | $< 0.001$ | $< 0.001$ | $< 0.001$ | $< 0.001$ | $< 0.001$ | 0.001 | 0.002 |

Table 4.2: The p-values from the KS test applied to the dataset of the 5'L branchpoint distances shortened by thermodynamically stable zipper stems and the dataset of the 5'S branchpoint distances. The numbers in the first row are the values for the loop threshold ($t_l$) and the numbers in the first column are the values for the energy threshold ($t_e$). The p-values highlighted in boldface are greater than 0.05; for these $t_e$ and $t_l$ values the hypothesis that two compared datasets stem from the same distribution cannot be rejected at the standard significance level of $\alpha = 0.05$.

more stable stem close by that would be identified if the energy threshold were lower. On the other hand, for $t_e < -12$ there is a large number of 5'L introns whose structures do not contain any stems that would satisfy the free energy criterion. Based on this analysis, we have chosen several values for $t_e$ that are within the acceptable range, namely: $-5$, $-7$, $-10$ and $-12$.

The values for the loop threshold ($t_l \in \{2, 3, 4, 5, 6, 7, 8\}$) were chosen based on our observation for the secondary structures of the 5'L introns and on the distributions of the number of free bases in bulges and internal loops obtained from the SSTRAND database (Figure 4.14).

We ran the modified algorithm for identification of thermodynamically stable zipper stems for all combinations of $t_e$ and $t_l$ values. For each particular pair of $t_e$ and $t_l$ values, the algorithm identified a zipper stem that satisfies the requirements for each of the 5'L introns and calculated a shortened branchpoint distance. To statistically test the difference between the shortened distance distribution for the 5'L introns and the original branchpoint distance distribution for 5'S introns, we used the KS test. The p-values thus obtained are shown in Table 4.2.

For most of the values for $t_e$ and $t_l$, the p-values are less than 0.05, indicating statistically significant differences between the datasets of branchpoint distances. However, for $t_e = -7$ and for $t_l \in \{4, 5\}$, the p-value was 0.055, indicating that the hypothesis that the two compared datasets stem

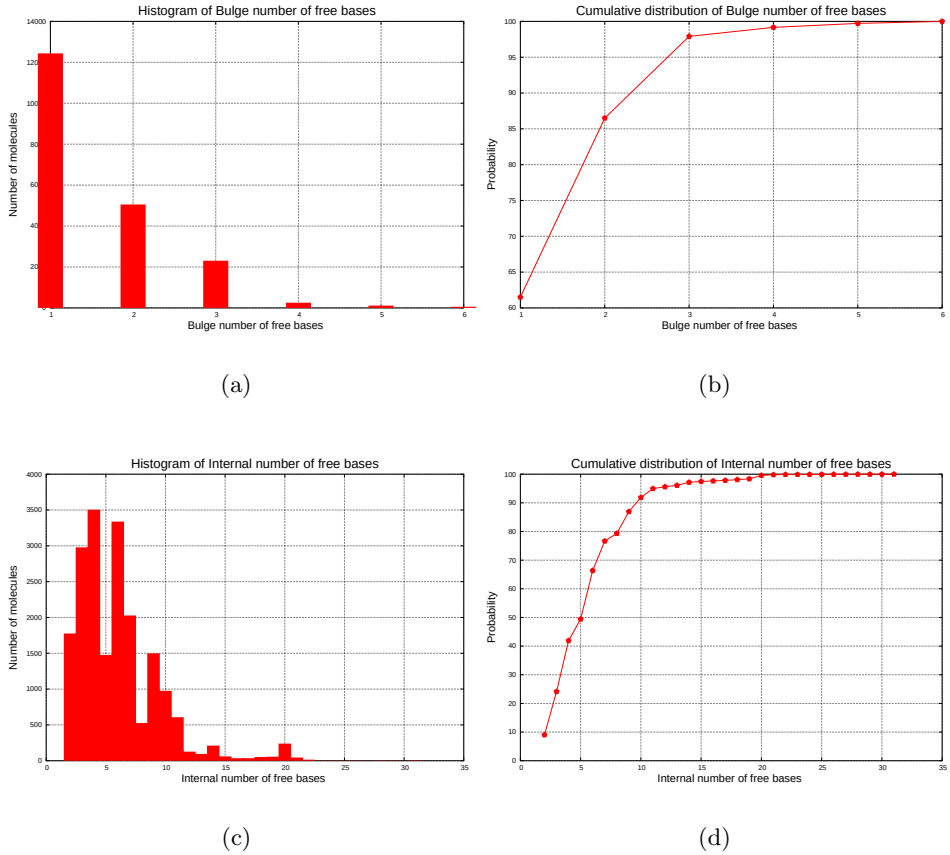



Figure 4.14: Distribution histograms and cumulative distribution plots of the number of free bases in **(a), (b)** bulges and **(c), (d)** internal loops of naturally occurring RNA molecules (obtained from SSTRAND database).

| $t_e \backslash t_l$ | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
|---|---|---|---|---|---|---|---|
| **−5** | $< 0.001$ | $< 0.001$ | $< 0.001$ | $< 0.001$ | $< 0.001$ | $< 0.001$ | $< 0.001$ |
| **−7** | 0.011 | 0.007 | $< 0.001$ | $< 0.001$ | $< 0.001$ | $< 0.001$ | $< 0.001$ |
| **−10** | 0.001 | 0.001 | **0.218** | **0.370** | **0.575** | **0.543** | **0.543** |
| **−12** | $< 0.001$ | $< 0.001$ | 0.002 | 0.002 | 0.009 | 0.032 | 0.042 |

Table 4.3: The p-values from the KS test applied to the dataset of the 5'L branchpoint distances shortened by multiple thermodynamically stable zipper stems and the dataset of the 5'S branchpoint distances. The numbers in the first row are the values for the loop threshold ($t_l$) and the numbers in the first column are the values for the energy threshold ($t_e$). The p-values highlighted in boldface are greater than 0.05 and for these $t_e$ and $t_l$ values the hypothesis that two compared datasets stem from the same distribution cannot be rejected at the standard significance level of $\alpha = 0.05$..

from the same distribution cannot be rejected.

In order to test the significance of these results, we also ran the algorithm on the two control datasets, described in Section 4.2.1. Similar to our analysis with the 5'L introns, we ran the algorithm for all possible pairs of values for the energy and loop threshold. For both control datasets and for any pair of $t_e$ and $t_l$ values the KS test rejected the null hypothesis at the significance level of 0.05 (the p-values are smaller than 0.001 for any $t_e \in \{-5, -7, -10, -12\}$ and for any $t_l \in \{2, ..., 8\}$).

### 4.3.2 Multiple zipper stems

Analogous to our earlier phase of zipper stem analysis, we wanted to explore the possibility that there can be more than one zipper stem that would shorten the branchpoint distance. For this purpose, we modified our algorithm for identifying thermodynamically stable zipper stems to find all the stems that satisfy given thermodynamic criteria and that effectively shorten the distance between the donor site and the branchpoint sequence. The algorithm was run for the same values of $t_e$ and $t_l$ as before. The p-values calculated by the KS test are given in Table 4.3.

As can be seen from this data, for $t_e = -10$ and $t_l \in \{4, ..., 8\}$, the algorithm finds zipper stems that shorten the 5'L branchpoint distance in such a

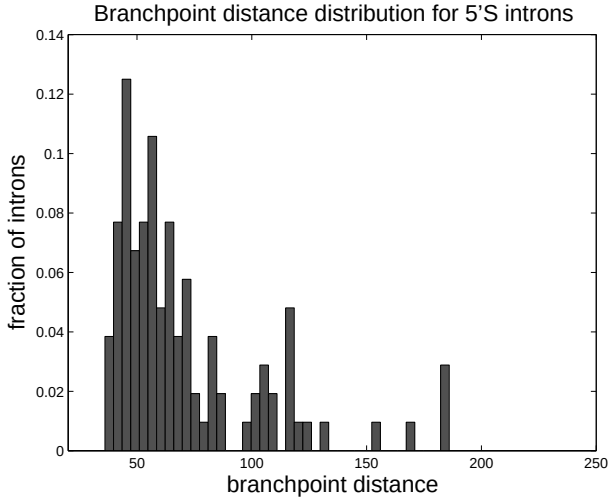
way that the resulting shortened distances closely resemble the optimal distances found in 5'S introns. This is illustrated in Figure 4.15, which shows the distributions of the original branchpoint distances for the 5'S dataset and shortened branchpoint distances for the 5'L dataset, when $t_e = -10$ and $t_l = 6$. The distributions appear very similar, with almost identical ranges and modes around 50 nt. The similarity is even more apparent in the cumulative distribution plots (Figure 4.16), where the respective function curves are nearly identical (KS test: $D = 0.105$, p-value = 0.575). The algorithm still finds only one zipper stem in 79 introns and more than one zipper stem in the remaining 31 introns.

We also identified multiple thermodynamically stable zipper stems for the random and exonic control datasets and compared their shortened branchpoint distances to the 5'S branchpoint distances using the KS test. The test rejected the null hypothesis for all combinations of energy and loop thresholds at a significance level of 0.05 (p-value $< 0.01$).

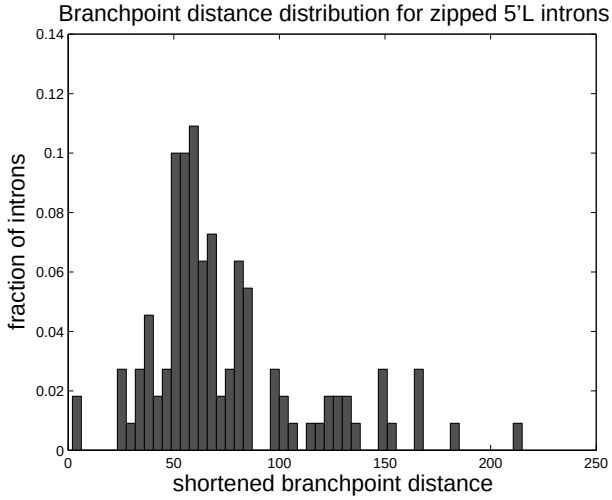
Overall, the identification of zipper stems based on the thermodynamic approach yielded better results than the stem-length based approach. This is encouraging since the former one is based on sound thermodynamic principles. We investigated the effects of single or multiple zipper stems on branchpoint distances in 5'L STRIN introns, and the results indicate that zipping the introns with multiple, more stable zipper stems, with minimum free energy $\Delta G(S) <$ -10 kcal/mol, yields better results. These results were compared with two control datasets, for which a similar effect of zipper stems on the analogue of branchpoint distances was not observed.

## 4.4 Phylogenetic analysis of zipper stems

In order to investigate if the zipper stems that we found in 5'L introns of *S. cerevisiae* are evolutionarily conserved among the *sensu stricto* species, we need to determine secondary structures for the corresponding introns in *S. paradoxus*, *S. mikatae*, and *S. bayanus*, find potential zipper stems and examine if they are at the same location as the stems in *S. cerevisiae*.




(a)




(b)

Figure 4.15: Distributions of (a) branchpoint distances for 5'S introns and (b) branchpoint distances for 5'L introns that were shortened by multiple thermodynamically stable zipper stems ($t_e = -10$ and $t_l = 6$).

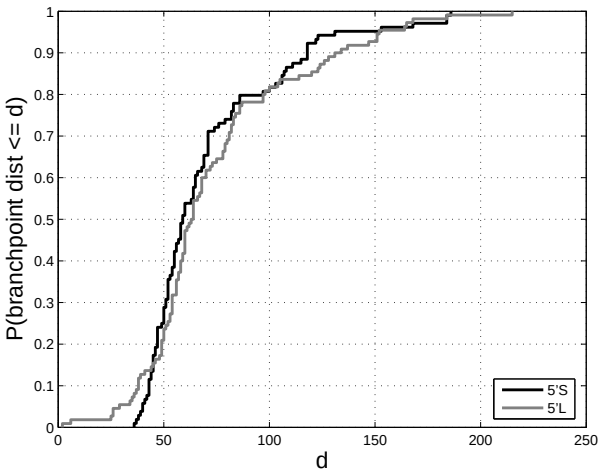



Figure 4.16: Cumulative distributions of the 5' splice site - branchpoint distance for 5'S introns and 5'L introns zipped with one or more stems ($t_e = -10$ and $t_l = 6$).

Another approach is to use comparative (a.k.a. phylogenetic) analysis of structures based on covariation analysis and/or structural alignment. Covariation analysis of RNA secondary structure is based on the assumption that a mutation that disrupts the Watson-Crick base-pairing of a functionally important RNA stem has a deleterious effect, which may be overcome by a second, compensatory mutation that restores base-pairing. An example of compensatory mutations is given in Figure 4.17. There are several approaches for identification of conserved RNA secondary structures, which will be discussed in Section 4.4.2.

Comparative RNA structure analysis is usually crucially dependent on the multiple sequence alignment algorithm used, and often visual inspection is needed to recognize errors in the alignment. This is why, for the initial phylogenetic analysis, we selected a small subset of nine 5'L STRIN introns. These introns were previously proposed to contain a stem-loop structure between the 5' splice site and the branchpoint sequence (Parker and Patterson, 1987). The selected introns are given in Table 4.4.

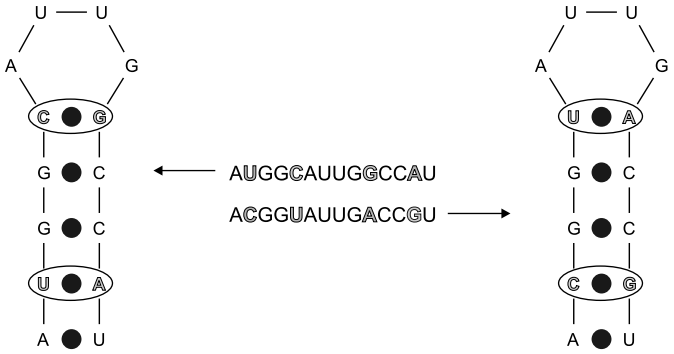



Figure 4.17: An example of compensatory mutations. The mutations in the second sequence are compensatory since they maintain basepairing.

| **ORF name** | **name in Parker and Patterson (1987)** | **intron length** | **intron location** | **branchpoint sequence** |
|---|---|---|---|---|
| YCR031C | rp59 | 307 | 8-314 | UACUAAC |
| YDR447C | rp51B | 314 | 4-317 | UACUAAC |
| YFL039C | ACT1 | 308 | 11-318 | UACUAAC |
| YGL030W | rp73 | 230 | 4-233 | UACUAAC |
| YGL103W | CYH2 | 511 | 50-560 | UACUAAC |
| YKL180W | rpL17A | 306 | 310-615 | UACUAAC |
| YML024W | rp51A | 398 | 4-401 | UACUAAC |
| YNL301C | rp28B | 432 | 113-544 | UACUAAC |
| YOL127W | rpL25 | 414 | 14-427 | UACUAAC |

Table 4.4: Introns used for our comparative RNA structure analysis.

### 4.4.1 Comparative analysis by visual inspection

We obtained optimal structures for the nine selected introns from Table 4.4 and for each of the four species using the mfold Web server (Zuker, 2003). For each structure we marked stem regions that are located relatively close to the 5' splice site and the branchpoint, which makes them possible candidates for zipper stems (Figures 4.19 and 4.20 show an example of this approach applied to intron YCR031C). These complementary sequence regions were subsequently labeled in the multiple sequence alignments. Since the quality of multiple sequence alignments is essential for comparative RNA structure analysis, we did not want to be solely dependent on the ClustalW results provided by Kellis et al. (2003). Therefore, we used another popular alignment Web server, LAGAN (`http://lagan.stanford.edu/lagan_web/index.shtml`, last accessed in June 2006), from which we obtained multiple sequence alignments using the Multi-LAGAN program (Brudno et al., 2003). An example of a labeled LAGAN multiple sequence alignment for intron YCR031C is shown in Figure 4.18.

The ClustalW and LAGAN alignments do not differ significantly, nevertheless LAGAN produces slightly better results for our data – zipper stems were more easily detectable using these alignments and a greater number of conserved stems was detected by the phylogenetic analysis based on LAGAN alignments. Thus, we mostly based our findings on the LAGAN alignments.

Inspecting the labeled alignments we looked for sequence regions where the labels overlap – each column in the alignment is labeled for each of the species. If they exist and are complementary, these regions form a stem which is conserved in *sensu stricto* species (black boxes in Figure 4.18). This stem is selected as a potential zipper stem.

Conserved zipper stems were found for five introns: YCR031C, YFL039C (2 stems), YGL030W, YGL103W, and YOL127W. Introns YKL180W and YML024W have a conserved zipper stem between *S. cerevisiae* and *S. bayanus*, but slightly different base-pairing for *S. paradoxus* (YKL180W) and *S. mikatae* (YML024W). However, when we checked the suboptimal structure predictions for two disagreeing species, we were able to find structures

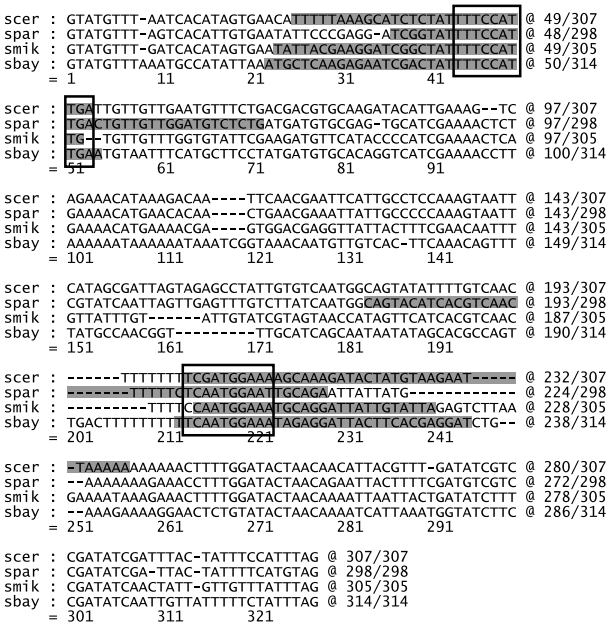



Figure 4.18: LAGAN multiple sequence alignment for intron YCR031C. Potential zipper stem regions are highlighted in *S. cerevisiae*, *S. paradoxus*, *S. mikatae* and *S. bayanus* sequences. Black boxes indicate the location of the conserved zipper stem.

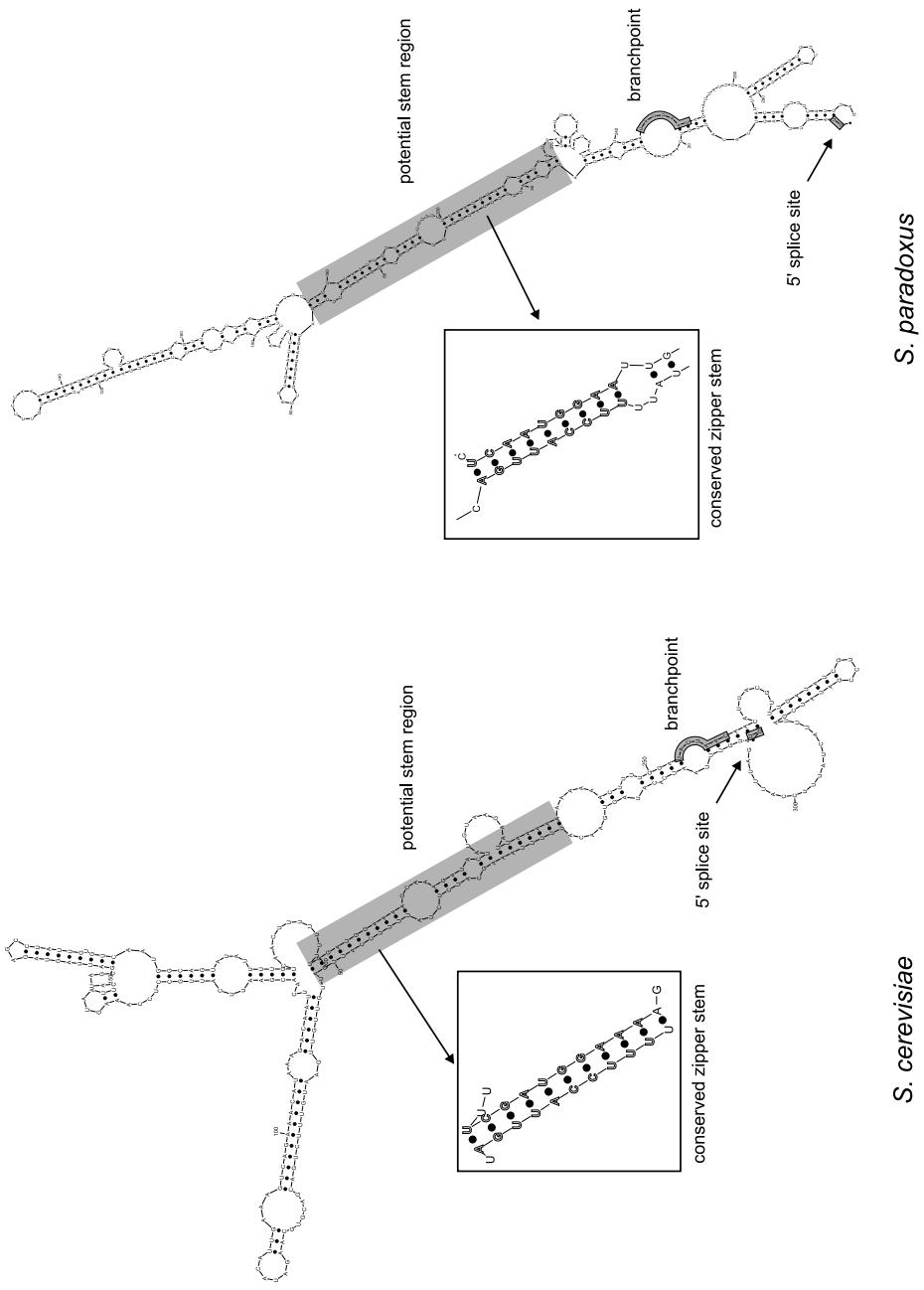



Figure 4.19: Minimum free energy secondary structures for intron YCR031C in *S. cerevisiae* and *S. paradoxus*. The 5' splice site, branchpoint and potential stem region are annotated for each structure. Conserved zipper stems found by comparative analysis are magnified and shown in boxes. Base-pairs conserved among all four species are highlighted.

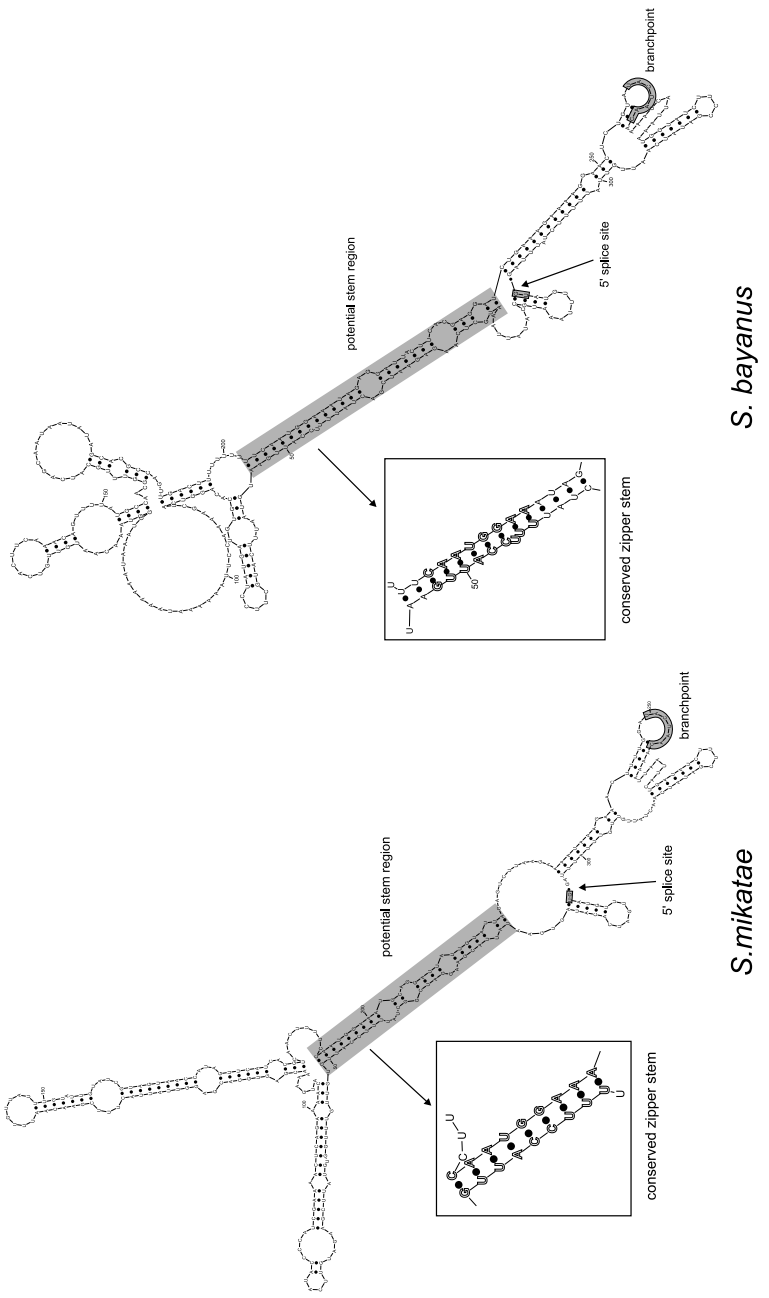



Figure 4.20: Minimum free energy secondary structures for intron YCR031C in *S. mikatae* and *S. bayanus*. The 5' splice site, branchpoint and potential stem region are annotated for each structure. Conserved zipper stems found by comparative analysis are magnified and shown in boxes. Base-pairs conserved among all four species are highlighted.

that have stems at the same location as for the other two species. Intron YDR447C has a conserved zipper stem between *S. cerevisiae* and *S. paradoxus*, but slightly different base-pairing for *S. bayanus*, which remained consistent even for the suboptimal structures (we looked only at the suboptimal structures within 5% from the MFE). Finally, for intron YNL301C we could not find a conserved zipper stem; nevertheless, complementary sequences close to the 5' splice site and the branchpoint sequence were found for each species.

The stems that we identified as conserved do not necessarily have to have conserved basepairing, although this is mostly the case. Since the zipper stem hypothesis states that the role of a zipper stem is to shorten the branchpoint distance to one that is optimal for splicing, the exact basepairing within the stem or the stem structure itself (internal loop and bulges) is not essential. Even if the stem's location is not the same but somewhat shifted, the stem's function would be unchanged. Therefore, the fact that a conserved stem was not found among all four species does not contradict the zipper stem hypothesis, especially considering that we are dealing with imperfect multiple sequence alignments and imperfect secondary structure predictions.

### 4.4.2 Comparative analysis using programs for comparative RNA structure prediction

There are a number of programs available for comparative RNA structure prediction, and they usually implement one of three basic approaches. In the case of relatively high sequence similarity among related RNA sequences, when it is possible to compute a good quality multiple sequence alignment, covariance or compensatory mutation analysis is used to process the alignments and predict common structures for aligned sequences. Examples of the algorithms from this class are Alidot (Hofacker et al., 1998), Pfold (Knudsen and Hein, 1999), ConStruct (Lück et al., 1999), Pfrali (Hofacker and Stadler, 1999) and Alifold (Hofacker et al., 2002). In the case when RNA sequences are more diverged and reliable multiple sequence alignment

cannot be achieved, the secondary structures of the sequences are predicted independently and then structurally aligned. An example of a program that uses this approach is RNA Forester, which uses a tree representation of predicted RNA structures to align them by applying a generalization of sequence alignment techniques (Höchsmann et al., 2003). The third class of comparative RNA structure prediction tools attempts to align given RNA sequences while simultaneously predicting their folding and common structure. This approach is preferred for computing a sequence alignment of RNA sequences since it uses secondary structure information, unlike ClustalW, LAGAN or other multiple alignment programs, which are based solely on primary sequences. It is a well-known biological phenomenon that structure is better conserved than sequence, especially for functional RNAs, and taking the structural information into account is essential for accurate alignment of these sequences. However, computing alignment and secondary structure simultaneously is very computationally expensive, which limits the number and length of sequences that can be aligned. Examples of the third class of programs are Foldalign (Gorodkin et al., 1997), Dynalign (Mathews and Turner, 2002) and Carnac (Touzet and Perriquet, 2004).

The algorithms for comparative RNA structure analysis can also be differentiated depending on whether they compute globally or locally conserved structure of the input RNA sequences. The majority of the programs mentioned attempt to predict a common global secondary structure (Pfold, ConStruct, Alifold, RNA Forester, Dynalign), while others predict only locally conserved structures.

Since in our study we have relatively reliable intron alignments of four closely related *Saccharomyces sensu stricto* species, we used the tools that rely on these multiple sequence alignments. We tested the following programs on our 9-intron dataset: Pfold and Alifold, which predict a consensus structure for all sequences in the input sequence alignment, and Alidot and Pfrali, which detect conserved substructures. Since zipper stems are long-range structural motifs, we should be able to detect them either in a global consensus structure of intron sequences or as a conserved stem.

**Pfold** uses context-free grammars to predict a common structure for

a given alignment of RNA sequences. It does not look for compensatory mutations directly, but estimates a phylogenetic tree from the alignment and uses it for maximum *a posteriori* approximation. The Web version of Pfold can be used for a set of maximum 40 RNA sequences, with a length limitation of 500 nt. The Web server can be accessed at `http://www.daimi.au.dk/~compbio/rnafold/` (last accessed in July 2006).

**Alifold** is a part of the Vienna package for RNA secondary structure prediction (Hofacker, 2003). It uses a modified energy model that integrates thermodynamic and phylogenetic information. It takes ClustalW multiple sequence alignment as its input and computes a consensus structure of the sequences. The maximum size of the input file is 10 KB. The Web server can be accessed at `http://rna.tbi.univie.ac.at/cgi-bin/alifold.cgi` (last accessed in July 2006). We also used a local copy of the program from version 1.5 of the Vienna package.

**Alidot** is also one of the Vienna package structure prediction tools. It starts with a ClustalW alignment of sequences and independently computes the MFE secondary structure of each of them. Using both the sequence alignment and the MFE structure predictions, Alidot aligns the structures and produces a set of candidate basepairs. These basepairs are further filtered to exclude any inconsistencies and to check for compensatory mutations.

**Pfrali** is yet another of RNA structure prediction tool contained in the Vienna package. The approach is very similar to Alidot's with the difference that Pfrali uses basepairing probabilities obtained from McCaskill's partition function algorithm (McCaskill, 1990) instead of MFE structure predictions. Since the basepairing probabilities contain information about a large number of plausible structures, this approach is less likely to miss any conserved structural elements.

Each of the nine introns, with its corresponding LAGAN alignment, was processed by Pfold, Alifold, Alidot and Pfrali. The first two algorithms give the common secondary structure for all of the sequences in the input alignment. The structure is given in dot-bracket notation and for Alifold also in conventional graphical representation. Examples of results for intron

YCR031C are shown in Figures 4.21 and 4.22.

Alidot and Pfrali have identical outputs listing all candidate basepairs and the final secondary structure, containing only the basepairs that pass all the filtering steps, is given in dot-bracket notation. For each basepair, the following information is given: location of interacting bases, number of sequences in the multiple sequence alignment in which the listed basepair is not one of the six standard RNA basepairs and an indicator that shows whether the basepair conflicts with a basepair that is predicted with a higher confidence. Some additional statistics are given but we use none of these for our analysis. The final structure given in dot-bracket notation is an ensemble of all basepairs that are not labeled as conflicting, however some of them may be supported by only one sequence in the alignment. For our study, we wanted to have better basepair support and thus considered only basepairs that were inconsistent with at most one sequence. The output of this post-processing step is a secondary structure in dot-bracket notation that contains all locally conserved substructures.

For each intron and for each program's predictions, we manually searched for a conserved stem that would bring the 5' splice site and the branchpoint sequence into closer proximity. Potential stems were annotated and compared to findings of comparative analysis based on visual inspection. A summary of the results from this analysis is given in Table 4.5. The first column of the table shows if a conserved zipper stem was found by visual inspection; in one case where conservation was found only between two species, those species are given. For each of the programs, two columns are provided: the first of these indicates if a program found a stem that satisfies our zipper stem requirements and also shows the number of these stems in parentheses if it is greater than one. The second column indicates if a predicted stem is present in the minimum free energy structure (obtained by mfold) of each of four species (this is the stem found by visual inspection). The four programs we used predict only uninterrupted stems, while zipper stems found by visual inspection can contain internal loops or bulges. This sometimes results in a zipper stem found by visual inspection matching two or more stems predicted by the programs. If a predicted stem was found only in

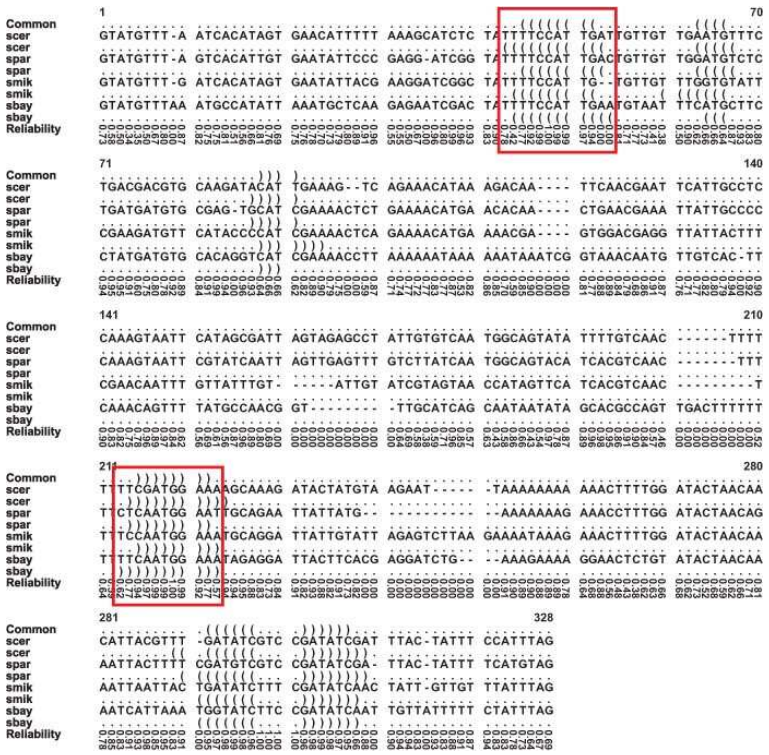


Figure 4.21: Pfold result for the alignment of the YCR031C intron. The first line is a common structure for all the sequences. The individual structures are found by applying the common structure to each sequence and extending stems, if possible. The last line indicates the reliability of prediction for each nucleotide in the alignment. The stem that satisfies the requirements for a zipper stem is enclosed in a box.

```
4 sequences; length of alignment 328

GUAUGUUU_AAUCACAUAGUGAAUAUUCCAAAGGAACCGCUAUUUUCCAUUGA_UGUUGUUGGAUGUUUC
CGAUGAUGUGCAAA_UACAUCGAAAACUCUAAAAACAUAAAAACAA____GUGAACGAAGUUAUUACCUC
CAAACAAAUUCAUAUCAAUUAG___AGU_UAUCGUAUCAACGACAGUACAUCACGUCAAC_______UUU
UUUUCAAUGGAAAUGCAGAAUUACUAUAU_AGAAU_______AAAAAAAAAAACCUUUGGAUACUAACAA
AAUUACUUAUUGAUAUCGUCCGAUAUCAAUUUAC_UAUUUCCAUUUAG
...((((...(((..(((((((...((.....)).)))))))((((((((((...(((((..(((((((
(((((((((.........)))))).....)))))))))).)))))).....................
.....................................(((((((....)))))))....................
..))))))))))..................................................)))...)))))..
..........(((((((....)))))))..................... (-21.74)
```

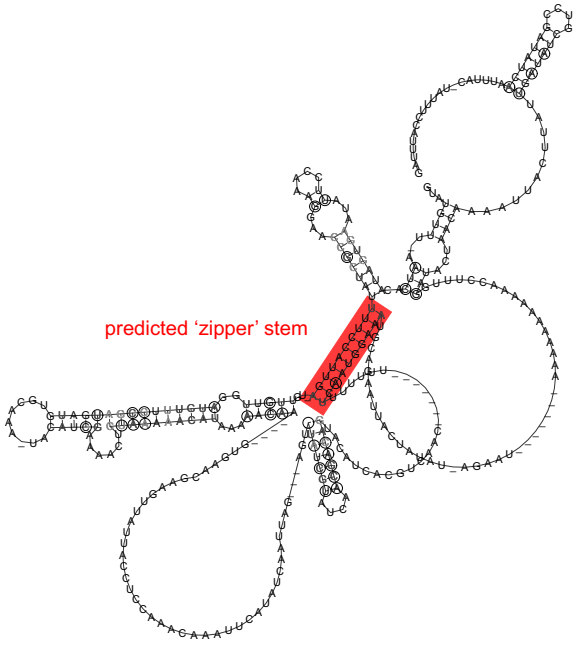



Figure 4.22: Alifold result for the alignment of the YCR031C intron. The output contains the consensus sequence and the optimal consensus structure in dot-bracket notation followed by its energy. A graphical representation of the structure is also given. The stem that satisfies the requirements for a zipper stem is enclosed in a box.

some then the species containing it are indicated.

Pfold found one or more potential zipper stems in 8 introns. One of the introns (YGL103W) was over the Pfold length limit. Alidot also identified zipper stems in 8 introns, failing to identify any conserved stem region for intron YNL301C. Alifold and Pfrali found one or more stems for all of the introns. For introns YCR031C, YGL030W, YGL103W, YKL180W and YOL127W all of the programs identified the zipper stem that was found by visual inspection. In three cases (YDR447C, YFL039C, YML024W) two of four programs did not confirm the stem found by visual inspection. The reason for this might be either that the stems identified by visual inspection did not have conserved basepairing but just overlapping locations or that the programs that missed them did so due to their specific weaknesses (since the stems were identified by the other two programs).

It is interesting that all programs except Alidot found a relatively long stem in YNL301C that was not identified by visual inspection. The identified stem was present only in the MFE prediction for *S. mikatae*, while the MFE predictions for the other three species do not contain it. Considering that the stem seems to be relatively stable, it is very likely that it is contained in some of the suboptimal predictions for the other species.

The zipper stems identified by our analysis shortened the branchpoint distances to 46-81 nt. These distances correspond well to the distances observed for 5'S introns (Figure 4.3(a)).

We believe that the results presented in this section are very encouraging, providing further support for the existence of zipper stems. For seven 5'L introns in our test dataset zipper stems were conserved among all four species; for one intron (YDR447C), conservation was observed only between the two closest species, *S. cerevisiae* and *S. paradoxus*. The high level of conservation observed for detected stems suggests functional significance. Alifold and Pfrali, which are based on compensatory mutations, were able to find conserved stems satisfying zipper stem requirements for all nine introns.

| **intron** | **zipper stem found by vi** | **Pfold** | | **Alifold** | | **Alidot** | | **Pfrali** | |
|---|---|---|---|---|---|---|---|---|---|
| | | stem found | present in MFE struct | stem found | present in MFE struct | stem found | present in MFE struct | stem found | present in MFE struct |
| YCR031C | + | + | + | + | + | + | + | +(2) | + |
| YDR447C | S.cer,S.par | + | − | +(2) | + | + | − | + | + |
| YFL039C | + | + | + | +(2) | − | + | − | + | + |
| YGL030W | + | + | + | + | + | + | + | + | + |
| YGL103W | + | too long | n/a | +(3) | + | +(3) | + | +(3) | + |
| YKL180W | + | + | S.par | + | + | + | S.par | +(3) | + |
| YML024W | + | + | − | + | − | +(2) | + | +(3) | S.cer,S.mik |
| YNL301C | − | + | S.mik | +(3) | S.mik | − | − | +(2) | S.mik |
| YOL127W | + | +(2) | + | +(3) | + | +(2) | S.cer,S.mik | +(3) | + |

Table 4.5: Results of comparative RNA structure analysis. Details are given in the text (S.cer=*S. cerevisiae*, S.par=*S. paradoxus*, S.mik=*S. mikatae*, S.bay=*S. bayanus*).

### 4.4.3 Comparative structure analysis on STRIN 5'L introns

Taking into account what we learned on the small, 9-intron test dataset, we performed an automated phylogenetic analysis on the introns from our phylogenetic dataset (Section 3.3). Based on Table 4.5, it seems that among the programs we tested Alifold and Pfrali were best suited for detection of conserved potential zipper stems. Both programs were able to predict one or more potential zipper stems in all of the nine introns, and in the majority of cases these stems overlapped the zipper stems identified by visual inspection analysis.

Each of these programs has some drawbacks. Alifold's prediction does not exclude basepairs that are inconsistent with some of the sequences in the input alignment, but keeps them as possible cases of sequencing or alignment errors or non-canonical basepairs (Hofacker et al., 2002). Thus, some of the zipper stems identified by Alifold might not be conserved in all four species. Pfrali also includes basepairs that are inconsistent with some of the sequences in the input alignment, however, this information is part of its output and we used it to post-process the final prediction to include only basepairs that are inconsistent with at most one sequence. The post-processing step introduces more free bases in the final structure prediction and results in shorter stems with fewer consecutive basepairs.

For the identification of conserved zipper stems using Alifold and Pfrali, we selected 49 STRIN 5'L introns for which we have all *Saccharomyces sensu stricto* sequences aligned (see Section 3.3). The more sequences there are in the input multiple sequence alignment, the better is the program's performance is. (For optimal performance, Hofacker et al. (2004) suggest at least 5 sequences with pairwise alignment of around 80% ).

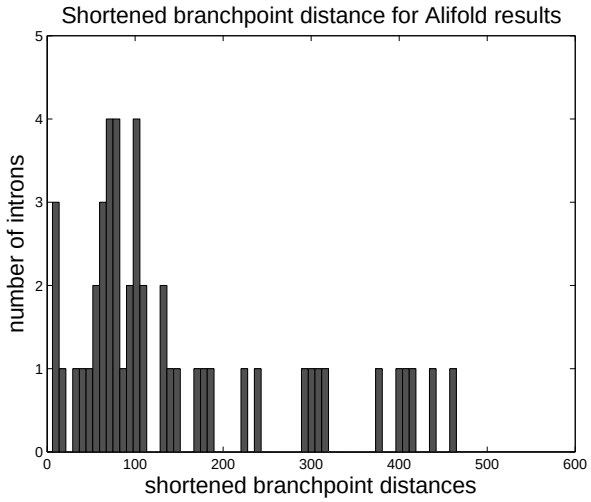
We used Lagan to align the sequences since for our data it produced slightly better alignments than ClustalW. Lagan alignments had to be re-formated into ClustalW format, since the latter is the only acceptable input format for the programs we used. The alignments were processed by Alifold and Pfrali, which produced a common secondary structure (global for Alifold and local for Pfrali) for all of the sequences in the alignment. The

secondary structure was further processed using the algorithm described in Section 4.3.1, which identifies a candidate zipper stem as the stem that maximally zips the intron and calculates the shortened branchpoint distance. Based on some empirical testing, we chose the algorithm's input parameters to be $t_e = -4$ and $t_l = 5$, which ensures that the minimum number of consecutive basepairs in a stem is three (often found in Pfrali predictions).

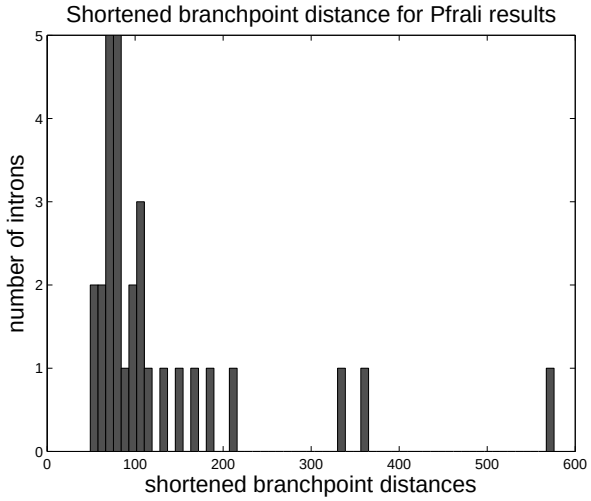
Of the 49 common secondary structures predicted by Alifold, our algorithm found potential zipper stems in 48. Of the 97 Pfrali predictions, stems were found in only 29 cases, due to the previously mentioned post-processing step that we applied to eliminate basepairs that were inconsistent with more than one sequence in the alignment (without pre-processing, stems were found in 45 cases). The distributions of shortened branchpoint distances for introns where zipper stems were found are shown in Figure 4.23; the cumulative distribution plots are shown in Figure 4.24.

When we compare these distributions with the distribution of branchpoint distances for short STRIN introns (Figure 4.15(a)), we observe that the majority of computed branchpoint distances fall within the distance range found for *S. cerevisiae* 5'S introns. For each program there are several exceptions where the branchpoint distances are either longer or shorter than for the 5'S introns. Zipper stems predicted by Pfrali, which are more reliable than those predicted by Alifold since they have to be conserved between at least three species, are better at shortening the branchpoint distance to the optimal range. This suggests that some of the shortened branchpoint distance calculated based on Alifold's predictions that are outside the optimal range might be a result of spurious stems conserved only between two species.

The results of this larger-scale phylogenetic analysis argue in favor of evolutionarily conserved zipper stems among *Saccharomyces sensu stricto* species. In a majority of cases, the conserved stems that were found by this fully automated comparative structure approach reduced branchpoint distances to values thought to be optimal for splicing. There are several possible explanations why this approach failed to identify conserved zipper stems for all of the considered introns (Pfrali) or identified the ones




(a)




(b)

Figure 4.23: Distributions of shortened branchpoint distances for 5'L STRIN introns where the analyzed secondary structure was the consensus structure for all *sensu stricto* species. The consensus structures were produced by **(a)** Alifold and **(b)** Pfrali.

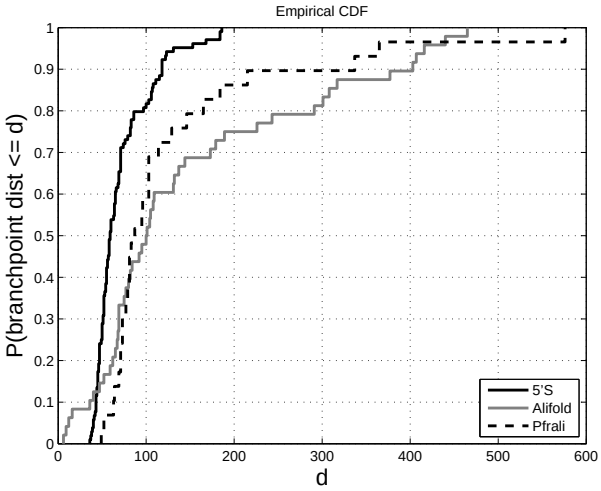



Figure 4.24: Cumulative distributions of shortened branchpoint distances for 5'L STRIN introns where the analyzed secondary structure was the consensus structure for all *sensu stricto* species. The consensus structures were produced by **(a)** Alifold and **(b)** Pfrali.

that might not be the real zipper stems resulting in shortened branchpoint distances outside the optimal range. First, both programs are heavily dependent on the underlying multiple sequence alignments, which may contain errors. Second, both programs depend on RNA secondary structure prediction methods that have limited accuracy (Mathews et al., 1999; Gutell et al., 2002). Errors in either multiple sequence alignment or secondary structure prediction would result in faulty consensus structure prediction. Another possible explanation is that in some cases approximate stem locations are conserved rather than the exact basepairing interactions; Alifold and Pfrali would fail to identify these stems. As discussed before, the zipper stem hypothesis states that the role of a zipper stem is to shorten the branchpoint distance to one that is optimal for splicing thus the exact basepairing within the stem is not essential. Even if the stem's location is not the same but somewhat shifted, the stem's function would be unchanged.

## 4.5 Conclusions

The analyses in this chapter were motivated by biological studies that identified long-range basepairing interactions in a number of long introns that result in shortening of the distance between the 5' splice site and the branchpoint sequence. It is believed that the relatively short branchpoint distance is necessary for spliceosomal assembly and efficient splicing of introns. To complement these biological studies we conducted a computational analysis of yeast introns, which can be classified into two groups based on their branchpoint distance – short (5'S) and long (5'L). The current hypothesis is that the branchpoint distances in the 5'S introns are thought to be optimal for splicing and that 5'L introns achieve this optimal distance by secondary structure formation.

Our computational analysis focused on detecting stems in secondary structures of the 5'L introns that would shorten the branchpoint distance to be within the optimal range. Zipper stems were found in all of the 5'L introns, which is not surprising considering the strong tendency of RNA sequences to form basepairing interactions. However, the shortened branch-

point distances of zipped long introns are distributed similarly to the branchpoint distances of short (5'S) yeast introns and differently than corresponding distances of zipped random and exonic sequences. The distribution modes for 5'S and 5'L introns are located around 50 nt, while the corresponding distances for zipped random and exonic sequences are uniformly distributed, without prominent modes. This was the case for all types of zipper stems that we considered. One possible implication of the mode phenomenon is that the optimal branchpoint distance for splicing of yeast introns is about 50 nt, which in long introns, as our results show, is achieved by shortening of the original distance by formation of zipper stems.

With further refinement of the zipper stem identification process, by considering only thermodynamically stable stems with limited internal loop sizes, the distributions of shortened branchpoint distances for 5'L introns and original branchpoint distances for 5'S introns became almost identical. In contrast, the distributions for random and exonic sequences were significantly different.

We also performed comparative structure analysis to analyze conservation of zipper stems among closely related yeast species. A careful manual analysis on a sample intron dataset of 9 introns found zipper stems that were conserved between four *sensu stricto* species. This is a significant result considering that the sequence conservation is not very high within the introns (50-74%). Similar, more automated analysis on a larger set of STRIN introns identified conserved zipper stems in almost all of them. The resulting shortened branchpoint distances fall within the optimal range observed in 5'S introns. Evolutionary conservation of zipper stem gives further support to their functional significance.

# Chapter 5

# Splicing efficiency and branchpoint distance

The zipper stem analysis in Chapter 4 provided evidence that the long branchpoint distances in STRIN 5'L introns can be shortened to distances presumed to be optimal for spliceosome assembly by one or more stem structures. The more or less ubiquitous presence of zipper stems in the secondary structures of long introns suggest that they might be important for efficient splicing of these introns. We investigate this hypothesis for the case of the RP51B intron and a number of its mutants for which the experimentally measured splicing efficiency results are correlated with the branchpoint distances shortened by the zipper stems.

As we will show in this chapter, the shortened branchpoint distances obtained using the approaches discussed in Chapter 4 do not explain the observed splicing efficiency results. This prompted us to modify our model of the role of intronic pre-mRNA secondary structure in splicing by considering not only MFE structure prediction of intron sequences but also near-optimal predictions and refining our calculation of shortened branchpoint distances. The refined distance calculation takes into account the entire structure of the intron, eliminating the need to search for a particular zipper stem. The zipper stem criteria are thus effectively relaxed, allowing complex stem structures and positioning of the 3' constituent of the stem downstream from the branchpoint sequence.

The refined model is shown to be in better agreement with RP51B experimental results, suggesting its potential to identify introns that have optimal structural conformation for splicing. We test the new approach on STRIN

long introns and random sequences with similar sequence characteristics as 5'L introns.

Finally, we describe the design of several new RP51B mutants based on the refined model of the role of intronic pre-mRNA secondary structure in splicing. We use these mutants to test our computational findings by biological laboratory experiments.

## 5.1 Experimental results for the RP51B intron

As briefly mentioned in Section 3.2.3, the pre-mRNA of the *S. cerevisiae* ribosomal protein rp51b has been used extensively for the analysis of secondary structure within introns and of its role in intron splicing. In 1993, Goguel and Rosbash observed that secondary structure interaction between two sequence segments located downstream of the 5' splice site and upstream of the branchpoint sequence promotes efficient splicing of the RP51B pre-mRNA. To further test the importance of this secondary structure in splicing, Libri et al. (1995) employed a copper resistance gene (CUP1), whose expression is dependent on splicing, as a reporter gene. They inserted the RP51B intron into the coding region of CUP1, which is otherwise intronless. Interrupted in this way, the cup1 protein is going to be functional, i.e., a yeast cell is going to be viable in the copper-containing medium, only if excision of the RP51B intron is successful.

In order to test the sensitivity of splicing to alterations in the stem as proposed by Goguel and Rosbash, Libri et al. introduced mutations in the interacting regions designated UB1 (upstream box 1) and DB1 (downstream box 1). They created 9 mutants: 3mUB1 (3 nt mutation), 4mUB1 (4 nt), 5mUB1 (5 nt), 6mUB1 (6 nt) and 8mUB1 (8 nt) where mutations fall in the UB1 region, 3mDB1 (3 nt) and 5mDB1 (5 nt) where mutations fall in the DB1 region and are compensatory mutations to the mutations in the 3mUB1 and 5mUB1, respectively, and 3mUB1/3mDB1 and 5mUB1/5mDB1, which are double mutants. All of the single mutants are expected to disrupt the secondary structure, while the double mutants are predicted to restore it. The copper sensitivity assay showed that for all single mutants except 8mUB1,

splicing was reduced. Surprisingly, 8mUB1 had a similar level of splicing as the wild type intron. Out of two double mutants, 5mUB1/5mDB1 was able to partially rescue splicing, while for 3mUB1/3mDB1 splicing was severely inhibited. These unexpected results were suggested to be the result of some secondary structure rearrangements; however, the secondary structure of the mutants 8mUB1 and 3mUB1/3mDB1 was not further explored.

Another interesting observation that emerged from the Libri et al. study is that UB1 and DB1 are not the only sequence segments in the RP51B intron whose interaction can facilitate pre-mRNA splicing: if the wild type interaction was disrupted, splicing could be restored by alternative basepairing, where the mutated UB1 sequence would pair with another block of sequence upstream or downstream from the branchpoint sequence. One of the alternative DB1 blocks was even found downstream from the branchpoint sequence. The stem formed using this DB1 sequence still shortens the distance between the 5' splice site and the branchpoint sequence, indicating that the existence, rather than the exact sequence constituents, of the stem is essential for efficient splicing of the RP51B intron.

In a follow-up study, Charpentier and Rosbash (1996) attempted to answer some of the important questions regarding secondary structure in the RP51B intron. Using newly designed mutants of the RP51B intron, they investigated at which step of splicing the stem has a functional role. They concluded that this happens at the time of commitment complex formation, i.e., the recognition of the donor site by the U1 snRNA and its consequent binding to it. The authors propose that the stable stem contributes to this complex formation.

Charpentier and Rosbash also performed structural probing of the stem whose secondary structure had been predicted only computationally. They confirmed that basepairing interactions between the UB1 and DB1 regions are indeed formed *in vitro* and *in vivo*, and that the structure of the stem formed is close to the one predicted (see Figure 5.1). The exact nature of the interaction is still not known since the techniques used for structural probing are not very precise. Their mutational analysis also confirmed that the effects on splicing *in vivo* were parallel to the effect observed *in vitro*,

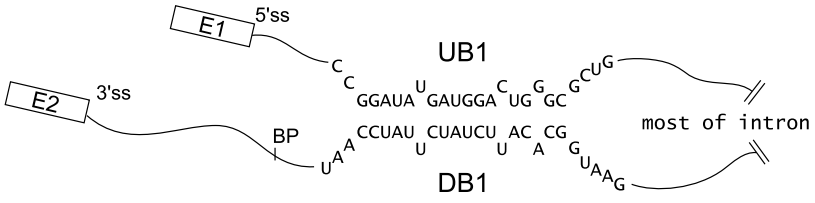



Figure 5.1: Reproduction of Figure 1B from Charpentier and Rosbash (1996): putative interaction between the UB1 and DB1 regions.

which is an important indication that secondary structure plays an essential role in splicing of the RP51B intron in nature.

Even though these studies have shown that the stem structure in the RP51B intron is essential for pre-mRNA splicing, they were unable to explain its real function. Libri et al. (1995) and Charpentier and Rosbash (1996) maintained the hypothesis suggested in Parker and Patterson (1987) that the structure might serve to reduce the distance between the 5' splice site and the branchpoint sequence to a distance optimal for spliceosomal interactions and pre-mRNA splicing.

## 5.2 Structural and branchpoint distance analysis of RP51B mutants

As mentioned earlier, splicing efficiency results for some of the RP51B mutants were different than expected. The assumption behind the mutant design in Libri et al. (1995) and Charpentier and Rosbash (1996) was that any mutation within the zipper stem, a stem bringing the donor and branchpoint sequence closer together, would disrupt the stem and change the intron secondary structure in such a way that the resulting branchpoint distance would be greater than for the wild type intron. However, the resulting

| **mutant** | $\bar{d}_{length=5}$ | $\bar{d}_{t_e=-10,t_l=6}$ | **splicing efficiency** |
|---|---|---|---|
| **wt** | 10 | 55 | efficient |
| **3mUB1** | 42 | 42 | slightly reduced |
| **5mUB1** | 42 | 42 | slightly reduced |
| **8mUB1** | 42 | 42 | efficient |
| **3mDB1** | 42 | 42 | inhibited |
| **5mDB1** | 133 | 46 | inhibited |
| **3mUB1/3mDB1** | 42 | 42 | inhibited |
| **5mUB1/5mDB1** | 133 | 83 | slightly reduced |
| **6mUB1** | 188 | 64 | inhibited |
| **4mUB1** | 188 | 86 | reduced |

Table 5.1: Correlation of the shortened branchpoint distance ($\bar{d}$) with splicing efficiency for Libri's mutants. Shortened branchpoint distances were calculated by two versions of algorithms for zipper stem identification: one uses the stem length to select stable zipper stems (first column), and the other uses thermodynamics criteria for stem selection (second column). Levels of splicing efficiency were inferred from Libri et al. (1995)

structures and branchpoint distances were never tested experimentally or computationally. This prompted us to compute intron secondary structures and shortened branchpoint distances of these mutants in an attempt to explain the experimental splicing efficiency results.

We first employed our algorithms for zipper stem identification, which we described in Chapter 4, to calculate the shortened branchpoint distance for the wild type RP51B gene and for all of the mutants described in Libri et al. (1995). We ran both versions of our algorithm, one of which uses stem length to select stable zipper stems and the other of which uses stem thermodynamics for the selection. In both cases, we ran the respective algorithm with the parameters that have been shown to produce the best results as explained in Chapter 4. For the first version of the algorithm, only stems longer or equal to 5 bp were considered; for the thermodynamics version, the energy threshold ($t_e$) was set to $-10$ kcal/mol and the loop threshold ($t_l$) was set to 6.

Table 5.1 shows the shortened branchpoint distances calculated by our two algorithms along with the results of the splicing efficiency assay per-

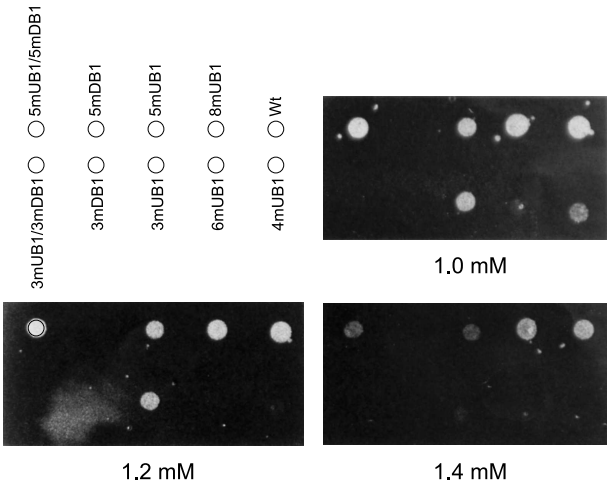



Figure 5.2: Reproduction of Figure 2 (C) from Libri et al. (1995): copper growth assay from Libri's mutants. In the upper left corner of the figure is a schematic drawing showing the locations of the mutants. $CuSO_4$ concentrations (in $10^{-3}$ moles/litre) are indicated under each panel.

formed by Libri et al. (1995). The splicing efficiency levels were inferred from Figure 2 (C) in Libri et al. (1995) (reproduced in Figure 5.2), which shows the results of the copper growth assay for wild type and the mutants. Figure 5.2 shows pictures of three plates containing different concentrations of copper sulfate ($CuSO_4$), where mutant cells have been growing in colonies (white spots). The size and intensity of the white spots indicate colony size and the level of copper resistance. Unfortunately, this type of measurement is not very precise and cannot be accurately quantified, so we approximated it with four different levels of splicing efficiency: efficient, slightly reduced, reduced and inhibited.

It is clear from Table 5.1 that the shortened branchpoint distances calculated by our zipper stem prediction algorithms do not explain the splicing efficiency levels for the Libri mutants. For example, the rows for 5mUB1, 8mUB1 and 3mDB1 have the same shortened distance of 42 for both versions of the algorithm, but different splicing efficiency levels.

One potential problem with our approach is that we assume that the sec-

ondary structure of an intron is independent of its flanking exonic or 5' UTR sequences (it is often the case that introns in *S. cerevisiae* are located right after the first codon). This is most probably an unrealistic assumption, since an intron is just a part of pre-mRNA, which, as with any RNA molecule, will tend to fold into its minimum free energy structure. However, pre-mRNAs are not free molecules but associate with many different proteins, protein complexes and other RNAs. It has been shown that a large number of processing factors associate co-transcriptionally with nascent RNA: 5' end-capping processing factors associate with the emerging transcript when it is only 20-40 bp long (Neugebauer, 2002), spliceosomal RNAs (U1, U2, U4, U5 and U6 snRNPs) along with a large number of splicing factors are co-transcriptionally recruited (Görnemann et al., 2005; Kotovic et al., 2003), and 3' end cleavage and polyadenylation factors are also bound throughout the length of the nascent RNA (Yu et al., 2004; Bentley, 2005). There are also many proteins and protein complexes responsible for transcription regulation, RNA editing and quality control, nuclear export, localization in the cytoplasm, translation regulation and degradation that have been shown to associate with pre-mRNA either during or after transcription (Neugebauer, 2002; Jensen et al., 2003; Hieronymus and Silver, 2004; Yu et al., 2004). Since these interactions will have an effect on the structure formation of a pre-mRNA molecule, predicting the secondary structure of the entire pre-mRNA, using current computational approaches that do not take into account such interactions, is unlikely to be successful.

Another reason why we should not consider predicting secondary structure of the entire pre-mRNA is the existence of co-transcriptional splicing: it has been shown that splicing often occurs during transcription while RNA polymerase II is still transcribing the downstream portion of a gene. This phenomenon has been observed for multicellular eukaryotes (Osheim et al., 1985; Beyer and Osheim, 1988; Baurén and Wieslander, 1994; Wuarin and Schibler, 1994; Tennyson et al., 1995; Wetterberg et al., 1996), but also more recently for *S. cerevisiae* (Elliott and Rosbash, 1996; Kotovic et al., 2003; Görnemann et al., 2005). Since the purpose of our computational structure prediction is to determine which secondary structure elements are essential

for splicing, we are only interested in the portion of the nascent pre-mRNA that has been synthesized when splicing occurs. However, the precise part of the nascent pre-mRNA that serves as the splicing substrate is not known.

Finally, the accuracy of computational RNA secondary structure prediction decreases with increased RNA sequence length and is considered unreliable for sequences longer than several hundred nucleotides (Morgan and Higgs, 1996; Mathews et al., 1999). Considering that the average ORF size in the STRIN dataset is 1065 nt, the average pre-mRNA size for the STRIN dataset is about 1300 nt (the average combined 5' UTR and 3' UTR length in yeast is about 250 nt (Hurowitz and Brown, 2003)), which means that predicting accurate secondary structures of STRIN pre-mRNAs would be difficult.

Based on these arguments, we believe that folding only intronic sequences is a reasonable approximation of the secondary structure within an intron, but we also repeated the same analysis where introns were folded with short flanking sequences on one or both sides. Since the splicing efficiency of the RP51B intron and its mutants was analyzed using the CUP1 gene as a reporter gene, we have considered the RP51B intron and its flanking sequences in this context. The RP51B intron was inserted into the genomic CUP1 gene after the first codon (Stutz and Rosbash, 1994), thus the 5' flanking sequence mainly consists of the 5' UTR region, which is at most 68 nt long (Karin et al., 1984; Zhang and Dietrich, 2005). We used RNAfold to fold the RP51B intron and all of Libri's mutants, including the first ATG codon and the 68-nt region upstream of the gene start. The shortened branchpoint distances were computed as before. The same analysis was performed with both 5' upstream and 50 nt downstream regions (beginning of the second exon). The results are shown in Table 5.2.

Analysis of the shortened branchpoint distances obtained by the two versions of our algorithm suggests that including flanking sequences in computation of the intronic secondary structure still does not explain the splicing efficiency results from Libri et al. (1995). In the case where only the 5' UTR of the CUP1 gene was included for predicting secondary structure, an overall trend seems to exist suggesting that mutants that are more efficiently spliced

| mutant | with 5' flanking seq | | with both flanking seq | | splicing efficiency |
|---|---|---|---|---|---|
| | $\bar{d}_{length=5}$ | $\bar{d}_{t_e=-10,t_l=6}$ | $\bar{d}_{length=5}$ | $\bar{d}_{t_e=-10,t_l=6}$ | |
| **wt** | 35 | 84 | 256 | 155 | efficient |
| **3mUB1** | 50 | 163 | 243 | 156 | slightly reduced |
| **5mUB1** | 50 | 112 | 255 | 155 | slightly reduced |
| **8mUB1** | 50 | 143 | 256 | 155 | efficient |
| **3mDB1** | 46 | 46 | 256 | 155 | inhibited |
| **5mDB1** | 229 | 117 | 256 | 155 | inhibited |
| **3mUB1/3mDB1** | 202 | 163 | 243 | 156 | inhibited |
| **5mUB1/5mDB1** | 233 | 105 | 255 | 155 | slightly reduced |
| **6mUB1** | 229 | 102 | 256 | 155 | inhibited |
| **4mUB1** | 50 | 85 | 256 | 155 | reduced |

Table 5.2: Correlating shortened branchpoint distance ($\bar{d}$) with splicing efficiency for Libri's mutants where short flanking regions were folded with intronic sequences. Shortened branchpoint distances were calculated by two versions of algorithms for zipper stem identification: one which uses the stem length to select stable zipper stems ($\bar{d}_{length=5}$) and the other which uses thermodynamics criteria for stem selection ($\bar{d}_{t_e=-10,t_l=6}$). Levels of splicing efficiency were inferred from Libri et al. (1995)

have lower values for $\bar{d}_{length=5}$. However, there are a few contradictory examples that do not support this theory: mutants 3mUB1 and 5mDB1, which both show inhibited splicing, have very different $\bar{d}_{length=5}$ values (46 and 229 nt, respectively). Also, for the double mutant 5mUB1/5mDB1, which has slightly reduced splicing, the shortened distance $\bar{d}_{length=5}$ equals 233 nt, which is more than for any other mutants, including those with apparently non-existent splicing. When both flanking sequences are included, the 68-nt-long 5' UTR of the CUP1 gene and 50 nt from the beginning of the second CUP1 exon, the resulting shortened distances, $\bar{d}_{length=5}$ and $\bar{d}_{t_e=-10,t_l=6}$, are roughly equal for all of the mutants, and fail to discriminate between the mutants that are spliced and those that are not.

Since we are aware that our shortened distance calculation is just an approximation with respect to the real distance within the folded molecule, we also wanted to look at the predicted secondary structures of these mutants and see if we can detect any structural differences that would explain their different splicing results. We computed the structures using the mfold Web server at `http://www.bioinfo.rpi.edu/applications/mfold/old/rna/form1.cgi` (last accessed in April 2006).

Our secondary structure analysis of the intron mutants, which basically involved visual inspection and determination of structural differences and similarities, mainly focused on the part of the structure that included the donor site and the branchpoint sequence, since we are interested in the distance between these two sites. The observed structural domain was almost identical for the 3mUB1, 5mUB1, 8mUB1, 3mDB1, 3mUB1/3mDB1 and 5mUB1/5mDB1 mutants, some of which have very different splicing efficiency levels. Moreover, the entire secondary structures of the 3mUB1 and 3mUB1/3mDB1 mutants were almost identical with only three basepairs difference, while the copper resistance experiments showed that the first one is spliced with only slightly reduced efficiency and that the second one is not (Figure 5.3). A similar analysis of secondary structures of Libri's mutants folded with their flanking regions showed that with the addition of flanking sequences, the secondary structure of the mutants tends to be even less susceptible to short mutations; in the majority of cases, there weren't any

major differences between the overall secondary structures of the mutants. Thus, visual comparison of the minimum free energy secondary structures of Libri's mutants folded with or without flanking sequences failed to explain the observed differences in the splicing efficiency.

Assuming that the splicing efficiency results from Libri et al. (1995) are accurate and that MFE prediction of secondary structure is reliable, we were not able to detect any correlation between the experimental results and the distance between the donor site and the branchpoint sequence. There is still a possibility that the distance between the two sites in the tertiary structure of the RP51B intron is quite different from our approximations and that this real distance would determine the efficiency of intron excision. Nevertheless, we believe that the more likely explanation for our inability to find any correlation between splicing efficiency and branchpoint distance is that our analysis was limited only to a single, minimum free energy prediction of the secondary structure of the mutants.

## 5.3 Refinement of zipper stem hypothesis

In order to remedy this limitation, we modify our approach by considering not only MFE structure prediction of intron sequences but also near-optimal predictions and refining our calculation of shortened branchpoint distances.

### 5.3.1 Including suboptimal structures in the analysis

According to the 'thermodynamic hypothesis', which was initially established for protein chains (Anfinsen, 1973), the tertiary structure of a protein or RNA molecule in its normal physiological environment is the one in which the Gibbs free energy of the whole system is the lowest. However, in the case of RNA, there are many alternative structures whose free energy is very close to the optimal one. It is possible that an RNA molecule can oscillate between these different structures or that different RNA molecules with the same nucleotide sequence can fold into these different, but energetically similar structures.

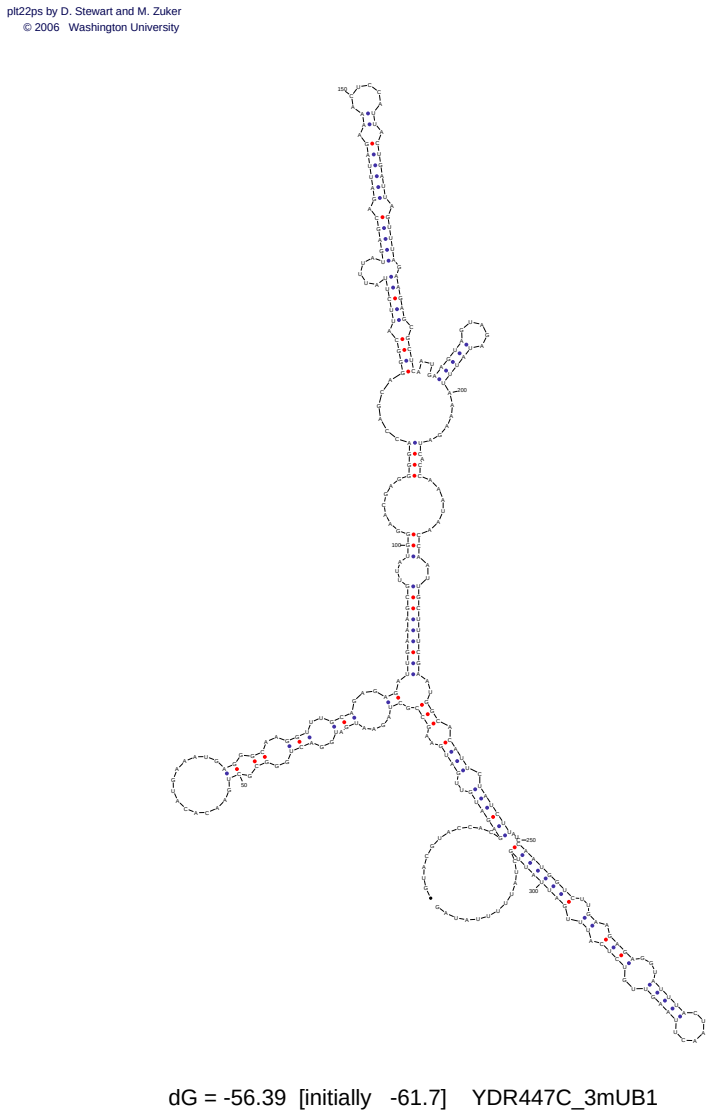



Figure 5.3: Minimum free energy secondary structure of 3mUB1 mutant predicted by mfold.

For functional, non-coding RNAs, such as tRNAs and rRNAs, there is a strong evolutionary pressure to maintain the unique, functional structure; mRNAs, on the other hand, do not have functional constraints on their global structure, and it is likely that they exist in a population of structures. Some evidence for this has been described in Christoffersen and Mcswiggen (1994), Betts and Spremulli (1994) and Freyhult et al. (2005).

Another reason why it would be beneficial to consider RNA secondary structures other than the MFE structure, especially when using computational prediction methods, is the known inaccuracy of the RNA secondary structure prediction algorithms (see also Section 2.2).

Therefore, it is possible that the native structure is not equivalent to the predicted MFE structure but to one of the suboptimal ones that still have free energies very close to the MFE. Some examples of this phenomenon have been noted in the literature (Wuchty et al., 1999). Considering this possibility, we have decided to include in our analysis all of the predicted suboptimal structures whose free energy is within 5% of the minimum free energy (this is the default value for mfold predictions; other values of suboptimality percentage are used in later analysis). Technically, this is easy to do since the folding programs that we use allow the user to specify this percentage.

### 5.3.2 A new way of calculating the branchpoint distance

The calculation of distance is very important for our analysis. The zipper stem hypothesis that we are testing implies that relatively short branchpoint distance is required for efficient splicing of an intron. Therefore, we need to be able to approximate this distance as closely as possible. For a direct calculation of the actual branchpoint distance in three-dimensional space, it is necessary to have accurate tertiary structure prediction. As currently there are no reasonably reliable algorithms for predicting RNA tertiary structure, we have to base our distance calculation on the RNA secondary structure. Even at this level it is hard to decide how to calculate the branchpoint distance.

In Chapter 4 we calculated this distance as follows: once the zipper stem is predicted, we summed the number of bases from the donor site to the zipper stem and from the zipper stem to the branchpoint sequence (see Definition 4). This method assumes that there is no secondary structure formation from the zipper stem towards the ends of the sequence. While somewhat consistent with our hypothesis that only the zipper stem is essential for shortening the distance, this assumption is obviously unrealistic. There is no reason to believe that any part of the intronic sequence will remain unfolded unless it is previously bound by some other molecule.

We thus decided to refine our model such that the distance between donor and branchpoint sites is calculated in the context of an entirely folded intron. The need to identify zipper stems, as defined in Definition 3, is thereby eliminated, allowing more flexible structure (arbitrary stem free energy and internal loop/bulge size) and location of a stem that brings the 5' splice site and branchpoint into close proximity (second complementary sequence can be located downstream from the branchpoint sequence, as observed by Libri et al. (1995)). This refinement also implies that the entire structure will be important for determining the branchpoint distance, rather than just a single stem. Even with this assumption, there is no unique way to calculate branchpoint distance.

To the best of our knowledge, there have been no previous attempts to compute the distance between a pair of nucleotides in RNA secondary structure. If we assume that helices are rigid, then calculating the distance between two nucleotides located in the same helix is relatively straightforward, assuming that we will use a number of nucleotides or basepairs as an abstract distance measurement. The task becomes more complicated if two nucleotides are found in different helices, or in general are separated by any loop (multi-loop or internal loop), because the mutual position of the two helices is not known: if the helices are closer to each other, the distance should be shorter, and if they are farther apart, for example because the angle between them is larger, the distance should be larger. However, this mutual positioning of the secondary structure elements can be considered a part of the tertiary structure of an RNA molecule and presently, we cannot

approximate distances between them. Instead, we have based the distance calculation solely on the basepairing information: the distance calculation within a stem is performed by counting the number of basepairs separating two sites, while in a loop it is done by counting the number of free bases. To perform this calculation, we employed Dijkstra's shortest-path algorithm (Dijkstra, 1959).

**Dijkstra's algorithm**

Dijkstra's algorithm, named after its creator, Edsger Dijkstra, is an algorithm that solves the single-source shortest path problem for a directed graph with non-negative edge weight. The input to the algorithm consists of a directed graph $G$ with associated edge weights, a source vertex $s$, and a target vertex $t$. Each edge of the graph is an ordered pair of vertices $(u, v)$ representing a connection from vertex $u$ to vertex $v$. Weights of edges are given by a weight function $w : E \to [0, \infty)$, so that $w(u, v)$ is a non-negative cost of moving from vertex $u$ to vertex $v$. The weight of an edge can be thought of as the distance between the two vertices. Then, the distance between two vertices in a graph, i.e., the cost of the path between the two edges, is equivalent to the sum of the costs of the edges in the path. For a given pair of vertices $s$ and $t$ in $V$, the algorithm finds the path from $s$ to $t$ with the lowest cost (i.e., shortest distance).

The algorithm works by keeping for each vertex $v$ the cost $d(v)$ of the shortest path found from the source vertex $s$ to $v$. Initially, $d(s) = 0$ and $d(v) = \infty$ for all other vertices $v \in V$ except $s$. The basic operation in Dijkstra's algorithm is *edge relaxation*: if there is an edge from $u$ to $v$, then the shortest known path from $s$ to $u$ can be extended to a path from $s$ to $v$, whose cost is $d(u) + w(u, v)$. If the result is less than the current $d(v)$, we can replace this value with the new value. Edge relaxation is applied to edge $(u, v)$ only once, when $d(u)$ has reached its final value.

The algorithm maintains two sets of vertices: $S$ contains only vertices for which the shortest path is known, and $Q$ contains all the other vertices. At the beginning, set $S$ is empty, and at each step the vertex with the lowest

value of $d(u)$ is moved from $Q$ to $S$. The time complexity of the algorithm is $O(n^2)$, where $n$ is the number of vertices in a graph.

#### Computing branchpoint distance using Dijkstra's algorithm

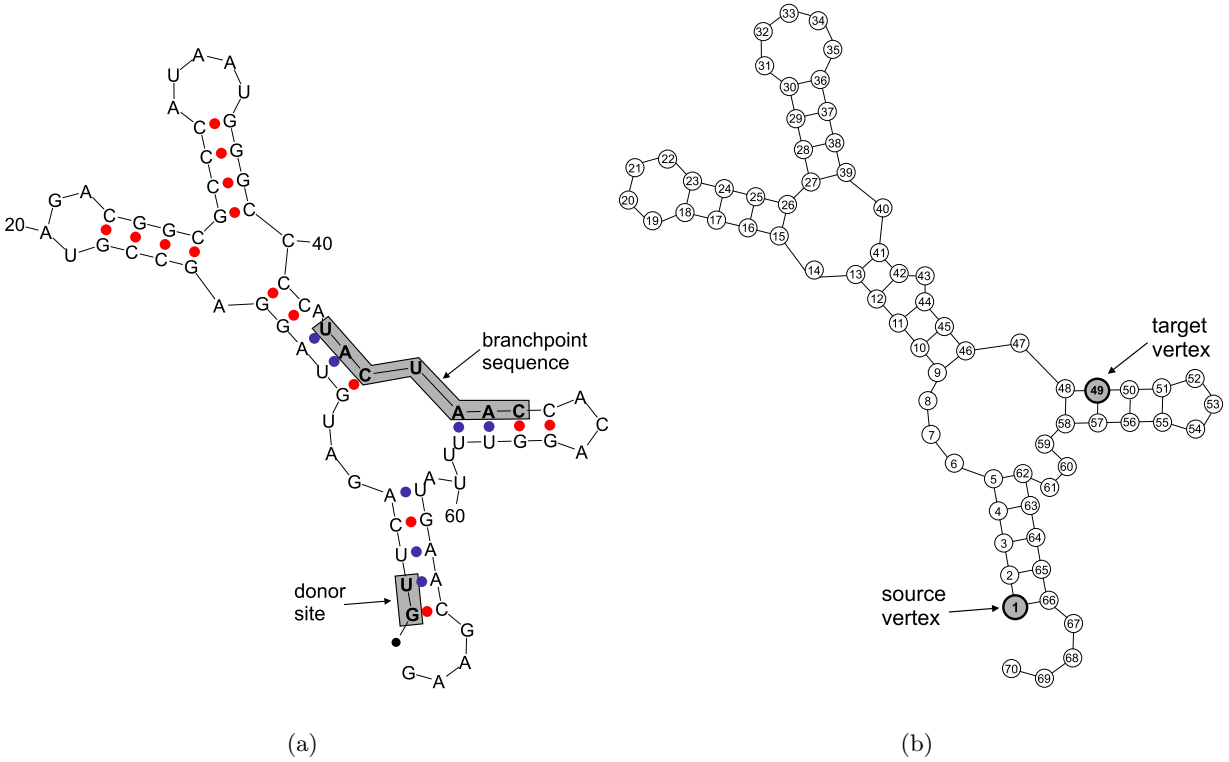
To calculate the branchpoint distance, we consider a predicted secondary structure of the intronic pre-mRNA as an undirectional graph where nucleotide bases are vertices of the graph and edges are bonds between the nucleotides. These bonds can be either sugar-phosphate bonds between the nucleotides in the RNA chain or the hydrogen bonds between paired bases in a given RNA secondary structure. Figure 5.4 shows the conversion from an RNA secondary structure to the graph representing it. Since Dijkstra's algorithm requires a directed graph, we will represent each non-directed edge $(u, v)$ as two directed edges, $(u, v)$ and $(v, u)$. All edges in the RNA graph have a uniform weight $w(u, v) = 1$.

In our implementation of the algorithm, the inputs to the program are an RNA structure in dot-bracket notation and the locations of two bases for which the distance needs to be calculated. These bases are the first nucleotide of the intron and the bulging $A$ in the branchpoint sequence (UACUA<u>A</u>C). The output of the program is the shortest distance between these two bases, which we are going to consider as the branchpoint distance for the given secondary structure.

### 5.3.3 Computation of structural characteristics of introns

In order to do the branchpoint distance analysis based on our refined model, we wrote a program that runs all of the analyses needed to obtain the desired structure information. Given a file with RNA sequences and a file with corresponding branchpoint distances, for each sequence in the input file the program produces the shortened branchpoint distance, the structure's free energy and the probability of each suboptimal structure within 5% from the MFE.

Our program calls two RNA secondary structure prediction programs: mfold and RNAfold. We use both of these programs because we need fea-




(a) (b)

Figure 5.4: Conversion from the RNA secondary structure to the graph representing it. **(a)** Graphical representation of the secondary structure of an intron (circles represent basepairing interactions, i.e., hydrogen bonds. **(b)** Graph representing the RNA structure in (a).

tures that are not available in a single program: mfold does not have an option for computing the partition function, and RNAfold does not have an option for selecting a set of suboptimal structures that are within a certain percent of suboptimality from the minimum free energy.

The program first runs RNAfold to calculate the equilibrium partition function $Q$ for each of the sequences in the input file (see also Section 2.2). RNAfold computes the partition function when run with the option '-p', but its value is not a part of RNAfold's regular output, so we had to slightly modify the source code to have it printed out. Once the partition function is computed for a given RNA sequence we can calculate the probability of each predicted structure given its free energy.

Since the partition function is computed by RNAfold and is thus based on the energy model used by the Vienna package, the free energy of a structure also needs to be computed using the same energy model in order to be consistent. Although both RNAfold (along with all of the other programs from the Vienna package) and mfold use Turner's energy model, they use different versions of it, which sometimes result in slightly different free energy values for the same RNA sequence. Thus, we first run mfold to predict the MFE structure and suboptimal structures within 5% from the MFE and then run RNAeval, a program which is a part of the Vienna package, to compute the free energies of the structures predicted by mfold.

The *percentage of suboptimality* is one of the parameters of mfold. We also used default values for the other mfold parameters, including the *window* parameter. This parameter controls how many structures will be computed and how different they must be from one another. It takes on positive integer values; a smaller value results in more computed structures that may be quite similar to one another, while a larger value results in fewer, more varied structures. If this parameter is not chosen by the user, a default value is selected according to the length of the input sequence. After experimenting with different values for this parameter, we decided to use the default value.

Using the partition function value computed by RNAfold, $Q$, and the free energy value of the structure $R_{ij}$, $\Delta G(R_{ij})$, computed by RNAeval, we can now calculate the probability of that structure in the ensemble of all

Procedure StructureAnalyze

**input:** file with $n$ RNA sequences $S_1, \ldots, S_n$, file with corresponding branchpoint distances $d_1, \ldots, d_n$
**foreach**$(S_i, where\ i \in \{1, \ldots, n\})$
  run *RNAfold* to obtain the value for $Q$
  run *mfold* to predict all structures $R_{ij}$, $j = 1, \ldots, m$ within 5% from MFE
  **foreach**$(R_{ij}, where\ j \in \{1, \ldots, m\})$
    run *RNAeval* to calculate $\Delta G(R_{ij})$ (free energy of $R_{ij}$)
    calculate $P(R_{ij})$
    run *dijkstra*$(R_{ij}, d_i)$ to calculate branchpoint distance $\bar{d}_{ij}$
    output $\Delta G(R_{ij})$ from *mfold*, $\Delta G(R_{ij})$ from *RNAfold*, $\bar{d}_{ij}$ and $P(R_{ij})$
  **end foreach**
**end foreach**

Figure 5.5: Pseudo-code for the procedure StructureAnalyze described in Section 5.3.3.

possible structures $(R_{i1}, R_{i2}, \ldots)$ for the given sequence $S_i$:

$$P(R_{ij}) = \frac{e^{-\Delta G(R_{ij})/RT}}{Q} \tag{5.1}$$

Finally, using Dijkstra's algorithm we calculate the branchpoint distance for each of the computed suboptimal structures. At this stage we also count the number of bases of the branchpoint sequence that remained unbound in the given secondary structure. We will use this value later to test if the pairing status of branchpoint sequence is important for splicing efficiency.

The output of the program include mfold's free energy value, RNAeval's free energy value, the branchpoint distance and the computed structure probability for each suboptimal structure predicted by mfold (all of them being within 5% of the MFE). The top-level pseudo-code for the described procedure, which we call StructureAnalyze, is given in Figure 5.5.

Considering that the time complexity of RNAfold and mfold is $O(l^3)$ and the time complexity of RNAeval and Dijkstra's algorithm is $O(l^2)$, where $l$ is the length of the given input sequence, the time complexity of the procedure

StructureAnalyze is $O(n(l^3 + m \cdot l^2))$. The value of $m$, which is the number of predicted sub-optimal structures within 5% of the MFE, is typically in the order of 10 (the maximum value for STRIN dataset is 34). Even when the percentage of suboptimality is increased to 20% the value for $m$ remains similar (the maximum value for STRIN dataset is 44), due to mfold's *window* parameter. Since $l$ is greater than $m$, the overall time complexity for the procedure is $O(n \cdot l^3)$.

**Post-processing of the results**

The output of the procedure StructureAnalyze is piped into a post-processing procedure that computes some additional values for each predicted structure, $R_{ij}$, of a given sequence, $S_i$, and also calculates summary statistics for each of the given RNA sequences. $P(R_{ij})$ is the computed probability of a structure in the ensemble of all possible structures for a given sequence; since the number of all possible structures grows exponentially with the length of the sequence, this probability is usually a very small number. Since we are considering only the structures that are within 5% from the MFE we also compute a normalized (or relative) probability in the following way:

$$P_{norm}(R_{ij}) = \frac{P(R_{ij})}{\sum_{j=1}^{m} P(R_{ij})} \tag{5.2}$$

Another value that is calculated in this post-processing phase is $b\_weight_{ij}$, which is dependent on the number of free bases in the branchpoint sequence. It is defined in the following way:

$$b\_weigth_{ij} = \begin{cases} P_{norm}(R_{ij}) \times 5 & : \quad \#\ of free\ bases\ in\ branchpoint > 3 \\ P_{norm}(R_{ij}) & : \quad otherwise \end{cases} \tag{5.3}$$

The purpose of $b\_weight_{ij}$ is to take into account the hypothesis that structural accessibility of the branchpoint sequence (e.g., being located in a loop) has a positive effect on splicing efficiency. We will discuss this

topic further in Section 6.2. The structures where four or more branchpoint nucleotides are in a loop are more heavily weighted than the structures where the branchpoint is mostly in a stem. The multiplication factor of five was arbitrarily chosen.

Once all predicted structures for a given sequence $S_i$ have been processed, the program computes several summary statistics:

**average distance** – this is the average branchpoint distance for all of the structures predicted by mfold with the percentage of suboptimality equal to 5. If there are $m$ predicted structures, the $average\ distance_i$ is computed in the following way:

$$average\ distance_i = \frac{\sum_{j=1}^{m} d_{ij}}{m} \tag{5.4}$$

If our hypothesis that a relatively short branchpoint distance is required for efficient splicing were correct, lower values of *average distance* would indicate higher splicing efficiency.

**r_weight** – this is a heuristic measure, which is based on the hypothesis that introns with highly probable sub-optimal structures that have short branchpoint distance are spliced more efficiently. It is defined in the following way:

$$r\_weight_i = \sum_{j=1}^{m} \frac{P_{norm}(R_{ij})}{d_{ij}} \tag{5.5}$$

The addends in the above sum are computed for each predicted secondary structure and are directly proportional to the relative (normalized) probability of that structure $P_{norm}(R_{ij})$ (thus 'r' in $r\_weight$) and inversely proportional to its branchpoint distance. Higher values of $r\_weight_i$ should indicate more efficient splicing of the sequence $S_i$.

**r_b_weight** – this value is analogous to $r\_weight_i$ but also takes into account the number of free bases in the branchpoint sequences. It is defined as follows:

$$r\_b\_weight_i = \sum_{j=1}^{m} \frac{b\_weight_{ij}}{d_{ij}} \tag{5.6}$$

*b_weight* is defined in Equation 5.3. Higher values of $r\_b\_weight_i$ should indicate more efficient splicing.

## 5.4 Branchpoint-distance analysis of RP51B mutants using the refined model

We used our updated approach for analyzing intronic secondary structures and branchpoint distances to analyze the wild type RB51B intron and all of the mutants described by Libri et al. (1995) and by Charpentier and Rosbash (1996). The new approach to the structural analysis considers not only the MFE structure, but also all of the suboptimal structures that are within 5% from the minimum free energy and that are significantly different from each other according to mfold's criteria. The branchpoint distance is also calculated differently than before, using Dijkstra's shortest path algorithm, described in Section 5.3.2.

We included all of Libri's mutants in the analysis: 3mUB1, 5mUB1, 8mUB1, 3mDB1, 5mDB1, 3mUB1/3mDB1, 5mUB1/5mDB1, 6mUB1, and 4mUB1. The mutants were also analyzed using our original structure and branchpoint-distance analysis described in Section 5.2.

In addition, we analyzed all of the RP51B intron mutants that were described by Charpentier and Rosbash (1996). These are mut-UB1i, which has an inverted UB1 sequence (upstream box 1; 5' complementary region), mut-DB1i, which has an inverted DB1 sequence (downstream box 1; 3' complementary region), mut-UB1iDB1i, which has both UB1 and DB1 sequences inverted to make them complementary to each other, mut-5, which reduces the consecutive pairing region to 5 basepairs, mut-12, which improves pairing to 12 consecutive basepairs (eliminating one one-nucleotide bulge), and mut-18, which extends pairing to 18 consecutive basepairs (eliminating all three bulges in the pairing region, see Figure 5.1).

Similar to the approach of Libri et al. (1995), Charpentier and Rosbash (1996) inserted the mutated introns into the CUP1 gene. The insertion was made after the first codon (Stutz and Rosbash, 1994), thus the 5' flanking sequence consisted mainly of the 5' UTR region, which is at most 68 nt long (Karin et al., 1984; Zhang and Dietrich, 2005). The 3' flanking sequence is part of the CUP1 coding sequence. Analogous to the analysis done in Section 5.2, we considered both Libri's and Charpentier's mutant introns in isolation, and also including one (5') or both flanking sequences. As previously mentioned, the 5' flanking sequence consisted of the first CUP1 codon and 68 nt of its 5' UTR, and the 3' flanking sequence consisted of 50 nt CUP1 coding sequence downstream from the inserted intron.

The splicing efficiency of the mutants was not directly quantified by Charpentier and Rosbash (1996), but they can be inferred from some of the figures in their paper. Unlike the splicing efficiency analysis conducted by Libri et al. (1995), which was done by copper growth assay, Charpentier and Rosbash used gel electrophoresis of formed spliceosomal complexes. The pre-mRNAs, lariat intermediate complex and lariat product complex were resolved based on size, and the splicing efficiency level was approximated from the intensity of the bands. Thus, levels of splicing efficiency for the wild type pre-mRNA and mutant pre-mRNAs can only be approximated based on the published gel images.

Another complication in assessing splicing efficiency levels for Charpentier's mutants is that there is not a direct comparison of all of the mutant results with the wild type results. In some cases, authors used the wild type intron and the mutated introns with the donor site modified to correspond to the consensus donor sequences. With these modifications, introns are in general spliced more efficiently than in the original RP51B context. The splicing efficiency levels can still be approximated relative to one another based on Figures 2 and 3 from the article by Charpentier and Rosbash (1996). Four different levels can be observed: normal for wt, reduced for mut-UB1i, mut-DB1i, and mut-5, slightly improved for mut-UB1iDB1i and improved for mut-12 and mut-18. These levels cannot be compared directly to splicing efficiency levels for Libri's mutants.

All of the described mutants were processed using the program StructureAnalyze described in Section 5.3.3. The detailed output of the post-processing procedure is given in Appendix C.

The summary statistics for all of Libri's mutants are reported in Table 5.3. From the table we can see that there is an interesting correlation between the average branchpoint distance and the splicing efficiency levels: sequences that are more efficiently spliced (wild type, 3mUB1, 5mUB1, 8mUB1, 5mUB1/5mDB1, and 4mUB1) have lower values for the average distance than those that are poorly spliced. If we assign numerical values to the descriptive splicing efficiency labels (efficient = 1, slightly reduced = 2, reduced = 3 and inhibited = 4) we can compute the Pearson correlation coefficient ($r = 0.95$).

Looking at the detailed output given in Appendix C.1, we can observe that all of the sequences that are spliced efficiently or with slightly reduced efficiency have one or more predicted structures (MFE or suboptimal) for which the branchpoint distance is very short, $\bar{d}_{ij} = 5$. Analysis of the secondary structures of these sequences reveals that this distance corresponds to a structural conformation where the donor site and the branchpoint have two basepairing interactions between them. The part of the RNA secondary structure that shows this contact conformation is illustrated in Figure 5.6. Even more intriguing is the fact that the number of predicted structures that have this contact conformation between the donor site and the branchpoint seems to be proportional to the level of splicing efficiency. The wild type intron has 5 of these structures (efficient splicing), 3mUB1 and 5mUB1 mutants have one structure each where $\bar{d}_{ij} = 5$ (slightly reduced splicing), the 8mUB1 mutant has 3 of these structures (efficient splicing), and the 5mUB1/5mDB1 mutant has one such structure (slightly reduced splicing). Mutant 4mUB1, which has reduced but not completely inhibited splicing, does not have a structure with contact conformation within 5% from the MFE; however, if the percentage of suboptimality is increased to 10%, one of the structures predicted has a structure with a branchpoint distance of 5.

The value of $r\_weight_i$, which is supposed to be higher for the sequences that are more efficiently spliced, seems to have some predictive power, since

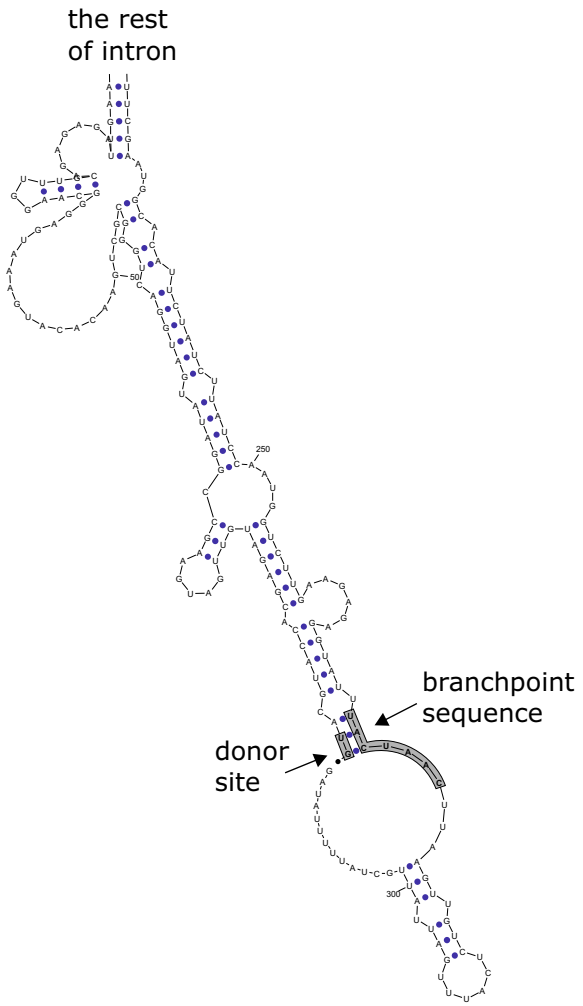



Figure 5.6: A part of the RP51B wild type intron secondary structure that shows basepairing between the donor site and the branchpoint sequence.

| mutant | avg | r_weight | r_b_weight | splicing efficiency |
|---|---|---|---|---|
| **wt** | 20 | 0.1494 | 0.1494 | efficient |
| **3mUB1** | 28 | 0.0319 | 0.0573 | slightly reduced |
| **5mUB1** | 29 | 0.0302 | 0.0509 | slightly reduced |
| **8mUB1** | 26 | 0.0283 | 0.0283 | efficient |
| **3mDB1** | 41 | 0.0244 | 0.0244 | inhibited |
| **5mDB1** | 41 | 0.0296 | 0.0631 | inhibited |
| **3mUB1/3mDB1** | 43 | 0.0243 | 0.0243 | inhibited |
| **5mUB1/5mDB1** | 34 | 0.0364 | 0.1218 | slightly reduced |
| **6mUB1** | 48 | 0.0208 | 0.1042 | inhibited |
| **4mUB1** | 38 | 0.0494 | 0.2458 | reduced |

Table 5.3: Summary statistics for Libri's mutants. Levels of splicing efficiency were determined from Figure 5.2.

there is a good correspondence between the values and the efficiency of splicing ($r = -0.52$). In general, efficiently spliced sequences have higher values than those with poor splicing. The other summary statistics, $r\_b\_weight_i$, that takes into account the number of free bases in the branchpoint sequence, does not correlate well with splicing efficiency ($r = -0.14$).

For Libri's mutants with flanking sequences, there seems to be no characteristic structures or branchpoint distances that would distinguish efficiently spliced sequences from ones that are not. Likewise, none of the summary statistics correspond well with the observed splicing efficiency levels (see Appendices C.2 and C.3).

The branchpoint distance results for Charpentier's mutants are similar to the results for Libri's mutants: the average branchpoint distances are lower for the sequences that are efficiently spliced (wild type, mut-UB1iDB1i, mut-12, and mut-18). If we assign numerical values to the descriptive splicing efficiency labels (improved = 1, slightly improved= 2, normal = 3 and reduced = 4), we can compute the correlation coefficient ($r = 0.92$). This, again, corresponds to the ability of these sequences to fold in such a way as to bring the donor site and the branchpoint sequences very close to each other: each of the efficiently spliced sequences has a number of contact-conformation structures (see Appendix C.4). The mutants that have reduced splicing do not fold into this type of structure, except for mut-DB1i, which

| mutant | avg | r_weight | r_b_weight | splicing efficiency |
|---|---|---|---|---|
| **wt** | 20 | 0.1494 | 0.1494 | efficient |
| **mut-UB1i** | 38 | 0.0454 | 0.2240 | reduced |
| **mut-DB1i** | 30 | 0.0424 | 0.0978 | reduced |
| **mut-UB1iDB1i** | 13 | 0.1569 | 0.1578 | slightly improved |
| **mut-5** | 35 | 0.0491 | 0.2455 | reduced |
| **mut-12** | 13 | 0.1569 | 0.1578 | improved |
| **mut-18** | 13 | 0.1569 | 0.1578 | improved |

Table 5.4: Summary statistics for Charpentier's mutants. Levels of splicing efficiency were inferred from Figures 2 and 3 and Table 1 in the article by Charpentier and Rosbash (1996).

has one suboptimal structure (of relatively low probability) with $\bar{d}_{ij} = 5$. This does not necessarily contradict the general trend, since the splicing efficiency measurements are very imprecise and it is possible that mut-DB1i is spliced more efficiently than mut-UB1i and mut-5, which would explain the presence of the short-branchpoint-distance structure.

The value of the **r_weight** summary statistic also correlate well with the splicing efficiency levels: $r\_weight_i$ values are significantly higher for efficiently spliced sequences, indicating good prediction potential ($r = -0.89$). This is not the case for **r_b_weight**, whose values do not correlate well with the splicing efficiency levels ($r = 0.29$).

For Charpentier's mutants with flanking sequences, there appear to be no characteristic structures or branchpoint distances that would distinguish between sequences that are spliced efficiently from ones that are not. Also, none of the summary statistics correlate well with the splicing efficiency level (see Appendices C.5 and C.6).

To conclude, structural and branchpoint distance analysis of the RP51B introns described by Libri et al. (1995) and by Charpentier and Rosbash (1996) has demonstrated the benefits of our new approach. Considering not only MFE structure but also a certain subset of suboptimal structures, as well as improving the calculation of the branchpoint distance, allowed us to identify some structural characteristics of RP51B intron mutants that may be responsible for their differential splicing. Namely, for all of the 16

sequences analyzed, the ability to form highly probable secondary structures with short branchpoint distance ensures their efficient splicing. In addition, it seems that the number of these structures and their probability correlates well with the splicing efficiency levels. This observation is also reflected in the average branchpoint distance value, which is always lower for the sequences that are more efficiently spliced. At this point it is not clear if it is the short distance itself or the specific contact conformation between the donor site and the branchpoint sequences that is important for splicing.

Our previous branchpoint distance analysis, based on zipper-stem identification, was not able to detect these differences, since it considered only the MFE structure of the mutants. Also, the definition of a zipper stem (Definition 3) excluded stems that contained either the donor site or the branchpoint sequence and thus would not be able to identify the stem found in the RP51B intron and its efficiently spliced mutants that bring the donor and branchpoint sequence into contact conformation.

The new branchpoint distance analysis yielded good results only when intron sequences without any flanking regions were folded. Adding flanking regions to the introns eliminates any recognizable differences between efficiently spliced mutants and those that are poorly spliced.

### Calculation of the probability of close branchpoint distance using the partition function

Instead of searching for structures that have short branchpoint distances or, in the case of the RP51B gene, that have basepairing interactions between the donor site and the branchpoint, the probability of contact conformation for a given sequence can be obtained directly. This probability corresponds to the basepairing probabilities of the paired nucleotides in question, and is calculated by the RNAfold program using the partition function. Each basepair probability reflects a sum of all weighted structures in which the chosen basepair occurs. Thus, these basepairing probabilities also take into account the structures that were not within 5% from the MFE, eliminating the necessity to chose an arbitrary suboptimality percentage value.

| mutant | G-C probability | U-A probability | splicing efficiency |
|---|---|---|---|
| **wt** | 0.40 | 0.40 | efficient |
| **3mUB1** | 0.33 | 0.33 | slightly reduced |
| **5mUB1** | 0.31 | 0.31 | slightly reduced |
| **8mUB1** | 0.34 | 0.34 | efficient |
| **3mDB1** | 0.01 | 0.01 | inhibited |
| **5mDB1** | $< 0.01$ | $< 0.01$ | inhibited |
| **3mUB1/3mDB1** | 0.01 | 0.01 | inhibited |
| **5mUB1/5mDB1** | 0.11 | 0.11 | slightly reduced |
| **6mUB1** | 0.05 | 0.05 | inhibited |
| **4mUB1** | 0.18 | 0.18 | reduced |

Table 5.5: Basepairing probabilities of contact conformation (Figure 5.6) for Libri's mutants. The probabilities were calculated by the RNAfold program.

Figure 5.6 shows the relevant basepairs formed between the donor site and the branchpoint sequence when the predicted structure contains what we call 'contact conformation'. The first basepair is formed between the first intron base (G) and the third base of the branchpoint sequence (C), and the second basepair is between the second base in the intron (U) and the second base of the branchpoint sequence (A). The probabilities for these particular basepairs can be obtained when the RNAfold program is run with the '-p' option, which invokes computation of the partition function and basepairing probabilities.

The basepair probability values for the wild type RP51B intron and all of Libri's mutants are given in Table 5.5. It can be observed that all of the sequences that are efficiently spliced have higher values for the basepair probabilities than the sequences that are poorly spliced ($r = 0.92$). The correlation is not strictly linear since, for example, the mutant sequence 8mUB1 has almost the same basepair probability value as 3mUB1 and 5mUB1, although it is more efficiently spliced than these two. Similarly, the mutant 5mUB1/5mDB1 is more efficiently spliced than 4mUB1, but this is not reflected in the basepair probability values.

For Charpentier's mutants, the basepair probabilities are also higher for the sequences that are more efficiently spliced (Table 5.6): all of the sequences that are efficiently spliced (wild type, mut-UB1iDB1i, mut-12, and mut-18) have basepair probabilities of 0.40, while the other sequences

| mutant | G-C probability | U-A probability | splicing efficiency |
|---|---|---|---|
| **wt** | 0.40 | 0.40 | normal |
| **mut-UB1i** | 0.04 | 0.04 | reduced |
| **mut-DB1i** | 0.25 | 0.25 | reduced |
| **mut-UB1iDB1i** | 0.40 | 0.40 | slightly improved |
| **mut-5** | 0.04 | 0.04 | reduced |
| **mut-12** | 0.40 | 0.40 | improved |
| **mut-18** | 0.40 | 0.40 | improved |

Table 5.6: Basepairing probabilities of contact conformation for Charpentier's mutants.

have lower values ($r = 0.79$). The mutant mut-DB1i has a relatively high basepair probability value with respect to the other two mutants, possibly for the same reason as given in the previous section – the splicing efficiency measurements in Charpentier and Rosbash (1996) lack precision, and it is possible that mut-DB1i is spliced more efficiently than mut-UB1i and mut-5. Another reason might be the imprecision of the energy model on which the basepair probabilities are based.

Overall, based on the results for Libri's and Charpentier's mutants it seems that the basepair probabilities can be good indicators for splicing efficiency.

## 5.5 Branchpoint-distance analysis on the STRIN dataset

According to the preceding analysis of RP51B mutants, it appears that intron sequences that can significantly shorten their branchpoint distance by forming secondary structures are spliced more efficiently. The predicted secondary structures with short branchpoint distances have to be highly probable structures, i.e., structures with close-to-optimal free energies.

Having this model of splicing in mind, we analyzed all of the long introns in the STRIN dataset to see if they exhibit structural characteristics similar to the wild type RP51B intron. The STRIN dataset contains 110 5'L introns, where the 'linear' distance between the donor site and the branchpoint sequence is greater than 200 nt. However, 12 of these introns are 5'

UTR introns (i.e., are found in the 5' untranslated region of genes), which we decided to exclude from this analysis for two reasons. The first reason is that in this analysis we also want to consider the flanking regions of introns, and for the 5' UTR introns information about their exact location is not available. For the introns found in the coding sequence, the location of splice sites is calculated from the first coding nucleotide. This cannot be done for the 5' UTR introns, thus in the Ares database from which we extracted intron locations these introns are assigned location 0, and in the SGD database the 5' UTR introns are not annotated at all. The second reason for not including the introns located in 5' UTRs is that, although they are removed by the same spliceosome machinery, it is possible that their requirements for splicing efficiency are different. It has been shown that the cap-binding complex (CBC), which binds to the 5' end of the nascent pre-mRNA, plays a role in the recognition of the 5' splice site of the first intron by the U1 snRNP during the formation of the spliceosomal commitment complex (Lewis et al., 1996a,b). Another study on the ACT1 intron found a correlation between intron position and splicing efficiency (Klinz and Gallwitz, 1985), and showed that splicing efficiency decreases with increased distance between the RNA cap site and the intron. Thus, it is possible that the proximity of 5' UTR introns to the pre-mRNA cap may have an effect on the splicing efficiency of these introns, which would diminish the role of the branchpoint-distance-shortening structure formation.

For each of the remaining 98 long introns, we computed all of the secondary structures within 5% from the MFE (with mfold's *window* parameter set to default value). For each sequence, we computed the minimum and the average branchpoint distance for all of the structures predicted and the most probable branchpoint distance, which is the distance calculated for the structure that has the highest probability, i.e., minimum free energy structure. We also computed the minimum, average and most probable branchpoint distance for all of the long introns that were folded with 50 nt flanking sequence on both ends. As control datasets, we generated two sets of 98 random sequences (one corresponding to intron sequences only and one corresponding to introns with flanking sequences) that have the same

length distribution and the same GC content as the intron datasets. Similar to the analysis of STRIN introns, we use the procedure StructureAnalyze to compute the distances corresponding to branchpoint distances in STRIN introns for each sequence in a control dataset. The distances are computed between the first nucleotide in the sequence and the nucleotide found at the same location as the start of the branchpoint sequence in the STRIN intron with the same length as a control sequence. For brevity, we still call these distances branchpoint distances. The distribution histograms and the cumulative distribution plots are shown in Figures 5.7 and 5.8.

From Figure 5.7 it can be observed that yeast long introns tend to fold into structures that have shorter branchpoint distances than folded random sequences of the same length and GC content: approximately one-third (33 sequences) of STRIN long introns have a minimum branchpoint distance of 5, the same as for the RP51B intron, while this is the case for only 6 of the random sequences. This tendency of yeast introns to have shorter branchpoint distances than random sequences can also be seen in the plots for average and most probable branchpoint distances. To test the statistical significance of these differences, we performed a Kolmogorov-Smirnov test, which determines if two distributions differ significantly (see also Chapter 4).

We computed $D$ statistics and p-values for all three pairs of datasets plotted in Figure 5.7. For STRIN and random datasets of minimum and average branchpoint distances, the null hypothesis of no difference was rejected (with p-values of 0.008 and 0.012, respectively). This was not the case for the datasets of most probable branchpoint distances (p-value = 0.13).

Figure 5.8 implies similar conclusions for intronic sequences with flanking regions. Although adding upstream and downstream sequences did not seem beneficial for the branchpoint distance analysis of the RP51B intron (Section 5.4), the analysis on all of the STRIN long introns indicate that even when folded with the flanking sequences, long yeast introns have different structural characteristics than random sequences of the same length and GC content. The KS test rejected the null hypothesis for all three pairs of datasets (for minimum branchpoint distance p-value = 0.0009, for average branchpoint distance p-value $< 10^{-4}$, and for most probable branchpoint

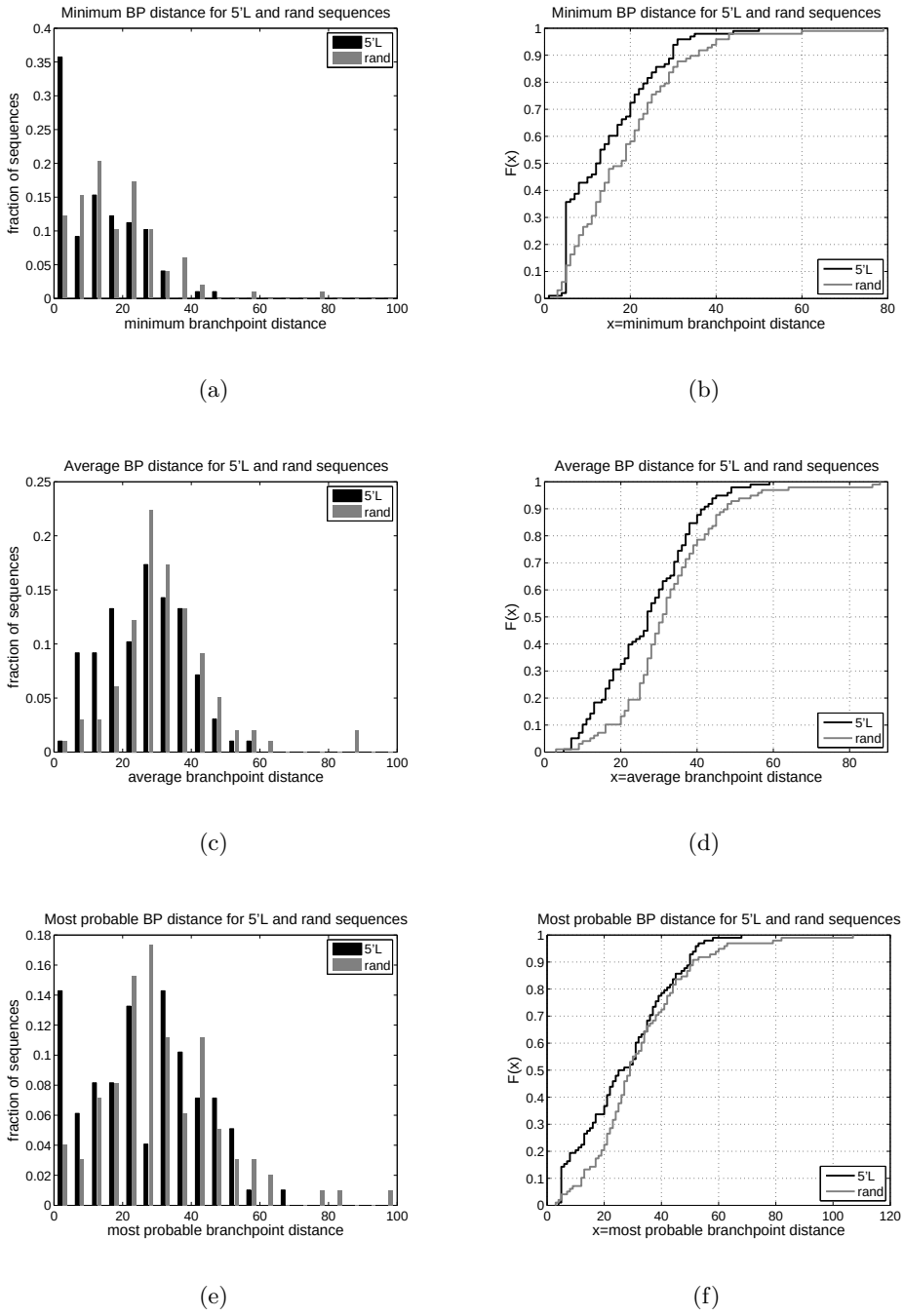



Figure 5.7: Comparing branchpoint distances for STRIN long introns and corresponding random sequences: distribution of minimum branchpoint distances as explained in the text (**a** – distribution histogram and **b** – cumulative distribution); **c, d:** distribution of average branchpoint distances; **e, f:** distribution of the most probable branchpoint distances.

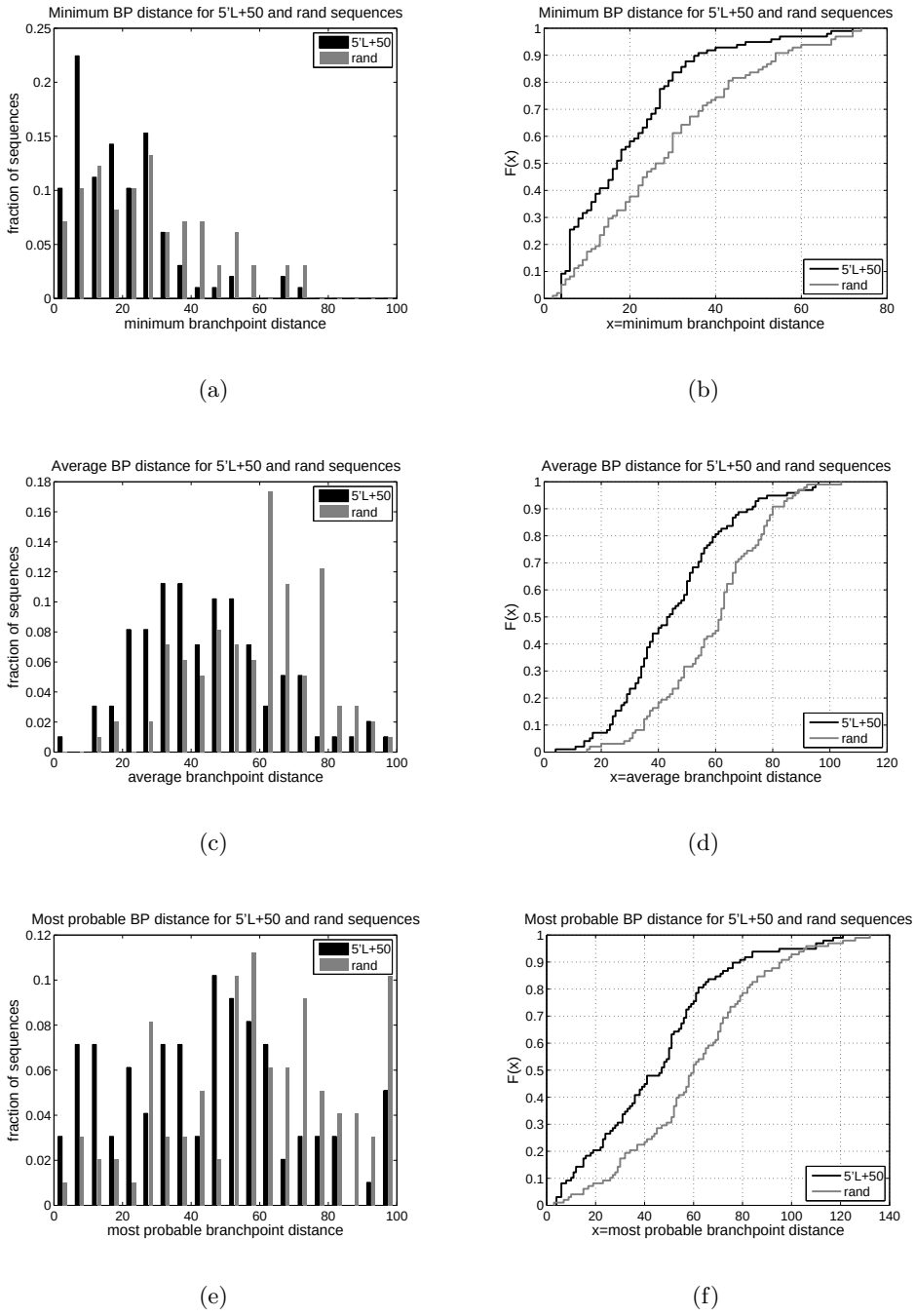



Figure 5.8: Comparing branchpoint distances for STRIN long introns with flanking regions and corresponding random sequences: distribution of minimum branchpoint distances as explained in the text (**a** – distribution histogram and **b** – cumulative distribution); **c, d:** distribution of average branchpoint distances; **e, f:** distribution of the most probable branchpoint distances.

distance p-value = 0.0001).

Since the most common branchpoint distance for the STRIN long introns is 5, we wanted to explore further if this distance always corresponds to the contact conformation observed for the RB51B intron and the extent of contribution of the branchpoint sequence and the canonical dinucleotide at the donor site to the formation of the basepairs between these two sites. Analyzing all of the suboptimal structures of the introns that have a minimum branchpoint distance of 5, we found that out of 33 introns that have this distance, only 13 have the same contact conformation as the RP51B intron. This does not necessarily mean that the rest of the introns cannot form this contact between the two sites, since for some of them the basepairing probabilities are relatively high (see next section).

In order to test the contribution of the GU sequences at the donor site and the branchpoint sequence (UACUAAC), we generated two new datasets of random sequences with the same characteristics as the previous ones but which have the canonical GU dinucleotide at the donor site (beginning of the sequence for the random sequences that have the same length distribution as STRIN introns and 50 nt from the beginning of the sequence for the random sequences that correspond to introns with flanking regions), dinucleotide AG at the acceptor site and canonical branchpoint sequence UACUAAC at the same location as the intron with the corresponding length. Thus, these sequences are very much like the real yeast long introns except that the sequences that are not essential for splicing are randomized. We computed the minimum, average and the most probable branchpoint distance for each random sequence in the dataset and compared these distributions to the corresponding distributions for the STRIN introns.

For introns and random sequences without flanking regions, the distribution differences are not as prominent as for the first type of random sequences: the Kolmogorov-Smirnov test failed to reject the null hypothesis of no difference for all three types of distribution. However, there is still a significant difference with respect to very short branchpoint distances: while the STRIN dataset has 33 sequences with a minimum branchpoint distance of 5, there are only 6 such sequences in the random dataset (5 have contact

conformation) and another 12 that have even shorter distances. The results for the sequences with flanking regions are similar to the results in Figure 5.8. For all three pairs of datasets we can reject the hypothesis that they stem from the same underlying distributions.

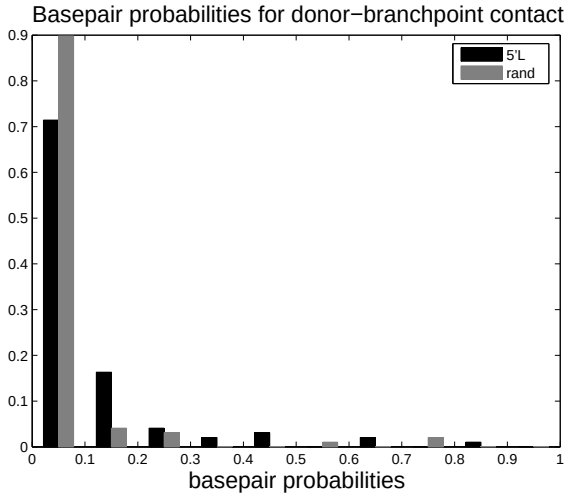
### Calculation of the probability of close branchpoint distance using partition function

Instead of searching for structures that have short branchpoint distances, we can use basepair probabilities that take into account all possible suboptimal structures. Basepair probabilities are calculated using the partition function (running RNAfold with '-p' option), with each probability reflecting a sum of all weighted structures in which the chosen basepair occurs.

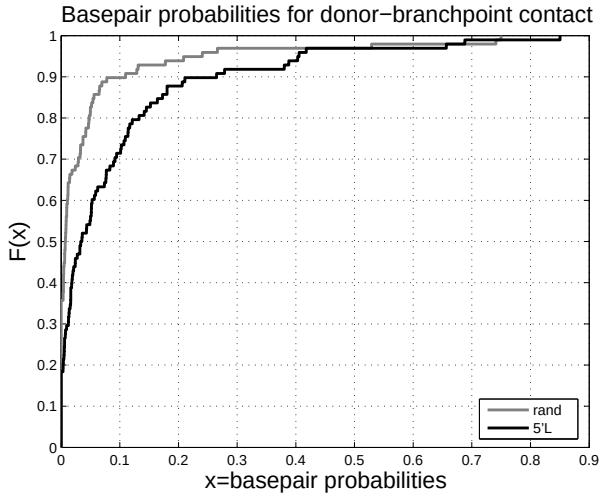
As we previously did for the RP51B mutants, we can calculate the probabilities of contact conformation for all of the STRIN long mutants and compare them with probabilities for random sequences. For this analysis, we used the random sequences that have the same length distribution and GC contents as the STRIN long introns but also have authentic splicing signal sequences (see previous section). For each of the sequences in these two datasets (for this analysis we did not consider sequences with flanking regions), we computed the basepairing probability of interaction between the donor site and the branchpoint sequence, as shown in the Figure 5.6. The distributions of obtained values are plotted in Figure 5.9.

From the figure it can be observed that the STRIN dataset, when compared to the random dataset, has fewer sequences that have very low probability of contact conformation ($p < 0.1$) and more sequences that have higher probability. The KS test rejects the null hypothesis that underling distributions are identical ($D = 0.36$ and p-value $< 10^{-4}$).

We can also compute the probability of a short branchpoint distance using basepair probabilities. Instead of extracting the specific probability of basepairing between the first and second nucleotides of the donor sequence with the third and second nucleotides (respectively) of the branchpoint sequence, we can choose the highest probability of basepairing between the




(a)




(b)

Figure 5.9: Comparing probabilities of the basepairing between the donor site and the branchpoint sequence for STRIN long introns and corresponding random sequences. **a:** distribution histogram of the basepairing probabilities; **b:** cumulative distribution of the basepairing probabilities.

donor dinucleotide and any dinucleotide within a certain window from the branchpoint sequence. This basepairing interaction will determine the maximum branchpoint distance for that sequence. We experimented with different window sizes up to 25 nt; for the sizes less than 20 nt, the KS test yielded p-values below 0.05, indicating that we can reject the null hypothesis that the two datasets stem from the same underlying distribution.

In summary, the results presented in Section 5.5 imply that yeast long introns, when compared to randomly generated sequences that resemble real introns in many respects, are more likely to fold into structures that have short branchpoint distances. There is a significant class of yeast long intron (1/3 of them) that can fold into secondary structures that have shortened branchpoint distance of 5, while this is the case for only a few random sequences. This observation is also supported by analysis of basepairing probabilities between the donor site and the branchpoint sequence.

## 5.6 Validation by biological experiments

Based on our structural and branchpoint distance analysis of Libri's and Charpentier's mutants (Section 5.4), we modified our previously proposed hypothesis of the role of secondary structure on intron splicing as follows: The existence of highly probable secondary structures (whose free energy is within 5% from the minimum free energy) that have short branchpoint distance (calculated by Dijkstra's algorithm, see Section 5.3.2) is required for efficient splicing of a yeast intron. In order to test the validity of this model of splicing, we needed to test it on introns that have not been used in the derivation of the model. We decided to design additional RP51B mutants whose splicing efficiency would be tested by laboratory experiments.

Validation of computational prediction by laboratory experiments is a very important component of bioinformatics research. It can provide additional support to computational results and make them more significant to the biological community, thus contributing to the research in both fields. We performed our laboratory experiments in collaboration with Dr. Philip

Hieter's group at UBC's Michael Smith Laboratories, whose focus is the molecular biology of *Saccharomyces cerevisiae*. The RP51B intron mutants that we designed were assembled and tested by Dr. Hieter's doctoral student Ben Montpetit.

Our experimental approach differs from that used by Libri et al. (1995) and by Charpentier and Rosbash (1996) in several ways. The mutated intron sequences are inserted back into the RP51B gene, instead of the CUP1 gene, which allows us to analyze the splicing of this intron in its endogenous environment. Another difference is that we estimated the splicing efficiency directly from the protein expression levels that we quantified using a state-of-the-art fluorescence imaging system. This makes our measurements more precise than those of Libri et al. (1995) and of Charpentier and Rosbash (1996). Ideally, splicing efficiency should be measured by the relative ratio of pre-mRNAs and mRNAs in a cell (Pikielny and Rosbash, 1985); however, our laboratory environment was not suitable for RNA isolation and quantification.

### 5.6.1 Verification of the experimental system

To verify that our experimental system works and that we can obtain the same results as in Libri et al. (1995), we synthesized some of Libri's mutants and tested their protein expression levels. The sequences tested were the wild type RP51B intron, and the 5mUB1, 3mUB1, 8mUB1, 5mDB1, and 3mDB1 mutated introns. The experimental procedure is described in Appendix B.

Figure 5.10 displays the results of our experiments. For each sequence, between four and six sample protein abundance measurements were obtained that were normalized with respect to the wild type protein levels. The normalized values were used to generate the plot. The shaded boxes represent the mean values for all the samples and the error bars represent $+/-$ 1 standard deviation. The error bar for the wild type intron comes from comparison of two different wild type samples.

The results in Figure 5.10 are not exactly the same as the splicing efficiency results in Libri et al. (1995), but the difference between the mutants

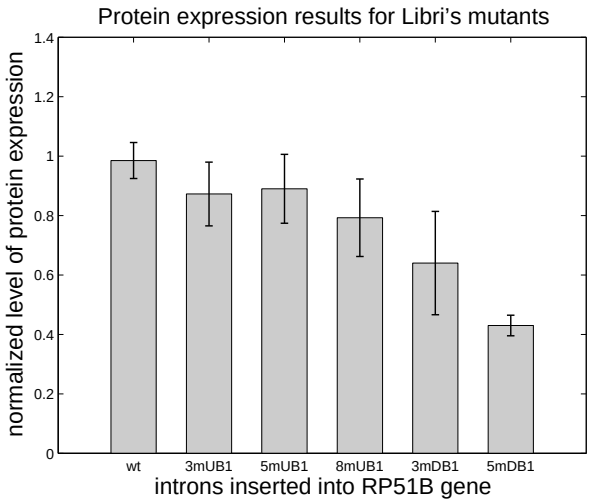



Figure 5.10: Protein expression results for the RP51B gene containing some of Libri's mutant introns obtained by our experimental approach. Protein expression level is normalized with respect to wild type expression level. Shaded boxes represent the mean value for several different samples and error bars represent $+/-$ 1 standard deviation for these samples. The error bar for the wild type intron comes from comparison of two different wild type samples.

that are more efficiently spliced and the ones that are not is obvious. Mutants 3mUB1 and 5mUB1, which showed slightly reduced levels of splicing in Libri et al. (1995) are still reduced compared to the wild type expression levels. The difference is that 8mUB1 mutant now seems to be less efficiently spliced than 3mUB1 and 5mUB1, which does not agree with the results in Libri et al. (1995). Expression levels for mutants 3mDB1 and 5mDB1 might also seem different than in the previous study, since their protein level appears not to be significantly reduced. However, we cannot assume that there is a perfect correlation between the rp51b copper resistance and splicing efficiency levels (see Figure 5.2). Thus, the lack of spots on the copper sulfate plates for mutants 3mDB1 and 5mDB1 does not necessarily imply that their splicing is completely inhibited. If we apply the KS test to determine the statistical significance of the differences between the wild type and mutant expression levels, we get p-values of 0.07 for 3mUB1, 5mUB1, and 8mUB1, indicating no significant difference between the samples, and a p-value of 0.005 for 3mDB1 and 5mDB1, which indicates that the hypothesis of no difference between the samples should be rejected.

In general, it is not realistic to expect identical results from two different experimental procedures, especially since the changes in expression levels of two different proteins (cup1 and rp51b) were measured. Overall, we consider our experimental results to be fairly consistent with Libri's results, which gives us confidence to use our experimental system for further analysis.

### 5.6.2 Mutant design

We designed 10 additional RP51B intron mutants for the purposes of testing our current model of the role of intronic pre-mRNA secondary structure on splicing. Five of the mutants were designed to have structurally unfavorable characteristics for efficient splicing ('bad' mutants) and the other five mutants have structural characteristics that are supposed to ensure efficient splicing ('good' mutants). The most important structural characteristic used for mutant design was branchpoint distance, calculated according to our model. In general, the main criterion for the design was that 'good' mu-

tants have short branchpoint distances and that 'bad' mutants have longer distances. More specifically, the requirement for the 'good' mutants was that they have multiple contact conformations (Figure 5.6), short average branchpoint distance and higher $r\_weight_i$ values. On the other hand, the 'bad' mutants were not supposed to have any structures with contact conformation or otherwise short branchpoint distance. Their average branchpoint distance should be significantly higher than for the wild type intron and the 'good' mutants and $r\_weight_i$ values should be lower. We also looked at the basepairing probabilities for the contact conformation and the number of free bases in the branchpoint sequence. The importance of having unstructured branchpoint sequence was studied and discussed by Hall et al. (1988), Stephan and Kirby (1993), Mougin et al. (1996), and Chen and Stephan (2003).

We carried out the design process manually, guided by the MFE structure of the wild type intron: for 'bad' mutants the main goal was to disrupt any structural elements (stems) that bring the donor site and the branchpoint closer together, and for the 'good' mutants these structures were stabilized. All the mutations are single-block mutations up to 20 nt long, where sequences of contiguous nucleotides were substituted by new sequences designed to change the secondary structure. Four of the five 'bad' mutants were obtained by mutating the original RP51B intron sequence, and the bad4 mutant was obtained by mutating the sequence of Libri's 8mUB1 mutant. Similarly, one of the 'good' mutants, good4, was obtained by mutating Libri's 3mDB1 mutant. Also, to test if contact conformation itself is important for splicing or just the resulting short branchpoint distance is, we created one of the 'good' mutants (good3) such that it would not have any structures with contact conformation but would have many structures where $\bar{d}_{ij} = 9$. The sequences of the mutated introns are given in Appendix D. Table 5.7 shows values for various quantities and characteristics used in the design process.

| **mutant** | **# of cc** | **p1** | **p2** | **avg** | **r_weight** | **BP** |
|---|---|---|---|---|---|---|
| **wt** | 5 | 0.70 | 0.40 | 20 | 0.1494 | loop |
| **bad1** | 0 | 0.0 | 0.0 | 42 | 0.0243 | loop |
| **bad2** | 0 | 0.0 | 0.0 | 45 | 0.0227 | loop |
| **bad3** | 0 | 0.0 | 0.0 | 43 | 0.0226 | stem |
| **bad4** | 0 | 0.0 | 0.21 | 36 | 0.0494 | stem |
| **bad5** | 0 | 0.0 | 0.005 | 33 | 0.0303 | stem |
| **good1** | 7 | 1.0 | 0.99 | 5 | 0.2000 | loop |
| **good2** | 6 | 0.83 | 0.40 | 13 | 0.1721 | loop |
| **good3** | 0 | 0.0 | 0.03 | 9 | 0.0968 | loop |
| **good4** | 5 | 0.61 | 0.80 | 17 | 0.1361 | loop |
| **good5** | 8 | 0.76 | 0.70 | 7 | 0.1615 | loop |

Table 5.7: Characteristics of newly designed RP51B mutants: **# of cc** – number of structures with the contact conformation within 5% from the MFE; **p1** – sum of normalized probabilities $P_{norm}(R_{ij})$ for all of the structures $R_{ij}$ that contain contact conformation; **p2** – basepairing probability of interaction between the donor site and the branchpoint sequence based on the partition function; **avg** – average branchpoint distance as given in Equation 5.4 (p. 112); **r_weight** – summary statistics defined in Equation 5.5 (p. 112); **BP** – structural configuration of branchpoint sequence (loop or stem).

### 5.6.3 Experimental procedure

The selected mutants were first synthesized using designed primers and PCR reaction and then inserted into the RP51B gene, which has a deleted intron. The cells were then allowed to grow and produce rp51b protein. The protein expression levels were measured using Western blotting analysis and then quantified using a specialized imaging system. The details of the experimental procedure are given in Appendix B.

As mentioned before, quantifying the level of protein expression is not an ideal measurement of splicing efficiency: the first assumption that we are making is that the level of protein abundance is proportional to the mRNA abundance in the cell. However, there are a number of post-transcriptional and post-translational events that can affect this proportionality on the global level. Although it has been shown that the general trend is that abundant mRNAs encode for abundant proteins and that the average protein per mRNA ratio is relatively constant through the full range of mRNA abundances (2500-4800 protein molecules per mRNA molecule), the results for individual genes can be very different: genes that have similar levels of mRNA abundance can have 30-fold variation in protein levels and vice versa (Gygi et al., 1999; Ghaemmaghami et al., 2003; Greenbaum et al., 2003; Beyer et al., 2004; Moore, 2005).

Since in our experiments we are dealing with only one gene, it is relatively safe to assume that post-transcriptional and post-translational events will have the same affect on all of the mutants tested. Consequently, the observed differences in the protein abundance level should reflect differences in the mRNA abundance level. However, the opposite may not necessarily be true: there may be changes in the mRNA level due to splicing deficiency or enhancement that are not going to be reflected in the protein abundance level (since post-splicing events can regulate the protein expression level).

The second assumption that we are making is that any change in splicing efficiency will be reflected in the mRNA levels, which is not necessarily true: Pikielny and Rosbash (1985) observed that for some of the mutants they tested, the levels of pre-mRNA were significantly increased, while there

were no changes in the mRNA level. Similar conclusions were drawn from genome-wide analysis of yeast splicing where authors were using the *splice junction index* $SJ$ = mRNA/(pre-mRNA + mRNA) and *intron accumulation index* $IA$ = pre-mRNA/(pre-mRNA + mRNA) to analyze the effects of mutations on splicing (Clark et al., 2002). Using only one of these two indexes or using only the pre-mRNA to mRNA ratio failed to detect all of the cases in which splicing was modified.

Overall, if the protein abundance levels for different mutants are different, we can conclude that that is a consequence of changes in splicing efficiency. However, if the protein abundance levels for the wild type intron and a mutant intron appear to be the same, then we still cannot exclude the possibility of modified splicing efficiency.

### 5.6.4 Results and discussion

The results of protein level abundance for the RP51B gene with our new mutated introns are given in Figure 5.11. For each mutant, between four and six sample protein abundance measurements were obtained that were normalized with respect to the wild type protein levels. The shaded boxes represent the mean normalized values for all the samples and the error bars represent $+/-$ 1 standard deviation. The results for mutants bad2, bad5, and good1 are missing because for bad2 and bad5 we were not able to insert the designed intron mutants into the RP51B gene, and for good1 there was no observable protein expression. We were not able to resolve the cause of these problems.

From Figure 5.11, we can see that mutants bad1 and bad3 have reduced splicing efficiency (or more precisely, protein expression levels) when compared to the wt as expected (the KS test applied on the distributions of protein expression levels rejected the null hypothesis of no digfference; p-value = 0.005 for both mutants).

Mutant bad4 has somewhat reduced splicing efficiency but not as much as the other two bad mutants. There are two possible reasons for this:

- The probability of basepairing interaction between the donor site and

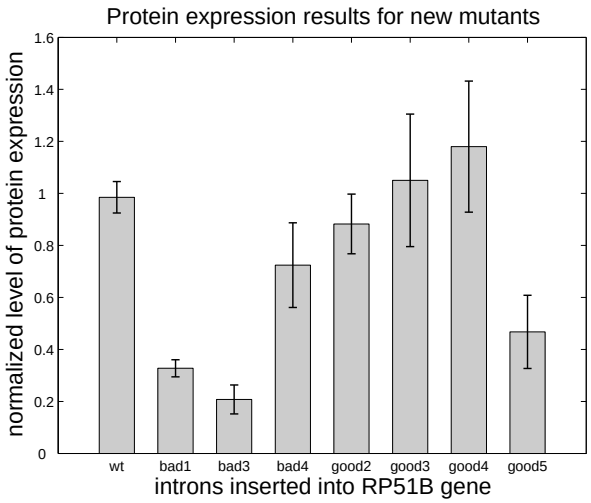



Figure 5.11: Protein expression results for the RP51B gene containing our new mutant introns. Protein expression level is normalized with respect to wild type expression level. Shaded boxes represent the mean value for several different samples and error bars represent $+/-$ 1 standard deviation for these samples.

the branchpoint sequence is 0.21, even though there were no structures within 5% from the MFE that had the contact conformation. This means that there are probably many less probable structures that have this conformation. If we analyze all of the suboptimal structures predicted by mfold (with default *window* parameter) that are within 20% from the MFE, we can see that this assumption is true – there are suboptimal structures in this free energy range with $\bar{d}_{ij} = 5$ and one of them is only 70 times less probable than the MFE structure. Thus, these slightly less probable structures may be sufficient to ensure relatively efficient splicing.

- The minimum free energy structure for this mutant has a branchpoint distance of 20, which still may be short enough for relatively efficient splicing.

The distribution of protein expression levels for bad4 is still significantly different from that for the wild type intron (p-value = 0.002).

Mutants good2, good3, and good4 are all spliced efficiently, as predicted. Mutant good3 does not have the contact conformation in any of the predicted structures. However, many of these structures have a short branchpoint distance of 9, suggesting that a specific structural arrangement between the donor site and the branchpoint sequence is not required for efficient splicing.

Mutant good5 shows reduced levels of protein abundance, which is in disagreement with our prediction. A possible explanation for this phenomenon may be the existence of a very stable stem (the free energy of the stem is $\Delta G = -36.6$ kcal/mol) that holds the 5' splice site and the branchpoint together (Figure 5.12). This zipper stem may be too stable to be disrupted, but a disruption would be needed in order to allow the spliceosome to bind to the splice signals.

Since good1 has an even more stable zipper stem than good5 ($\Delta G = -46.6$ kcal/mol), it is possible that this could cause total inhibition of splicing. Another reason could be the proximity of the mutated block to the 5' splice site (9 nt apart). However, judging by the other results it seems unlikely that there would be no detectable amounts of protein in the cell.

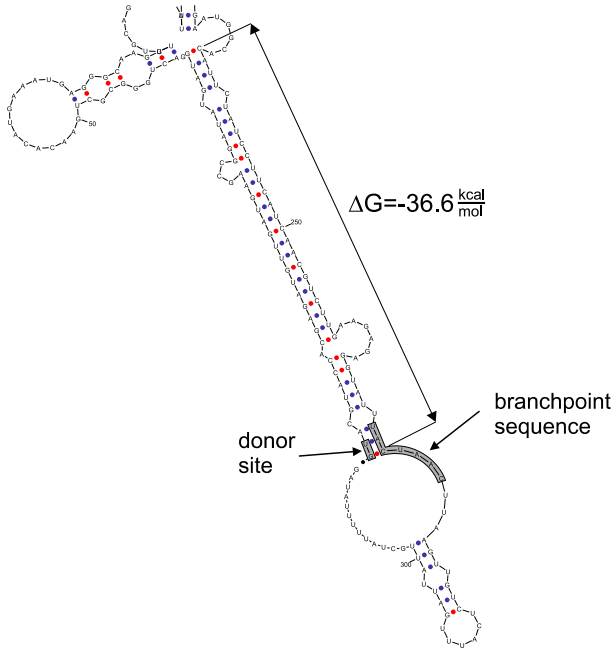



Figure 5.12: Part of the MFE secondary structure prediction for mutant good5 that shows the donor site and branchpoint sequence basepairing as well as the very stable zipper stem that stabilizes their interaction.

Thus, we still suspect that this result is due to a possible experimental error.

Overall, the results on the new RP51B intron mutants are consistent with our model of the role of intronic secondary structure in gene splicing.

## 5.7 Conclusions

In this chapter we extended our previous approach for calculating shortened branchpoint distances in 5'L yeast introns in two ways: (1) by considering not only the MFE structure but also a subset of suboptimal structures with free energies close to the MFE; and (2) by improving the calculation of the shortened branchpoint distance, taking into account the entire structure of an intron. This new approach allowed us to identify some structural characteristics of the RP51B intron and its mutants that seem to be responsible for their differential splicing. We observed that the mutants with a very short branchpoint distance corresponding to specific contact conformation between the donor site and the branchpoint sequence are spliced more efficiently.

We applied the new model to the STRIN dataset and observed that yeast long introns, when compared to randomly generated sequences (which resemble real introns in many respects) are more likely to fold into structures that have short branchpoint distances. This tendency is especially strong for very short branchpoint distances ($\sim 5$), resulting in a relative abundance of some specific structural conformations between the donor site and the branchpoint sequence. The contact conformation itself might imply that the canonical branchpoint sequence is maintained not only because it is complementary to a sequence in the U2 snRNA that interacts with the branchpoint sequence, but also because it is important for the secondary structure of the intron.

Obviously, having a contact conformation or a very short branchpoint distance is not a requirement for splicing, since there are many STRIN introns that do not have these characteristics. Still, it seems that there is a significant class of yeast introns that can fold into secondary structures with very short branchpoint distances (one-third of STRIN introns have a

branchpoint distance of 5). Judging by the experimental results on one of these (Section 5.6), this short distance is important for efficient splicing. For the remaining STRIN introns, there are four possible explanations:

- Their branchpoint distance is still relatively short and sufficient for efficient splicing.
- Their predicted branchpoint distance is long due to inaccuracy of the secondary structure predictions or branchpoint distance calculation.
- Their predicted branchpoint distance is long, but they have some other mechanism to achieve an optimal conformation for spliceosome assembly, or splicing efficiency is stimulated in another way (e.g., by protein factors).
- Their predicted branchpoint distance is long resulting in reduced splicing efficiency that is optimal for proper functioning of the gene in question; one example is the yeast gene YRA1 (Preker and Guthrie, 2006).

Finally, we computationally designed new RP51B intron mutants and predicted their splicing efficiency levels based on our new model of structural requirements for efficient splicing. The predictions were verified by laboratory experiments: the designed mutants were synthesized and inserted into the RP51B gene and then the expression level of the rp51b protein was quantified. The obtained measurements, which are thought to be proportional to splicing efficiency levels, were found to be consistent with our computational predictions.

# Chapter 6

# Structural characteristics of yeast introns

In the previous two chapters we focused on identification and analysis of a specific structural formation within yeast introns whose role is to shorten the large distance between the 5' splice site and the branchpoint sequence, and thereby enabling proper spliceosome assembly. However, it is possible that pre-mRNA secondary structure has other functions related to gene splicing in yeast.

In this chapter we conduct various analyses on the STRIN dataset, with the goal of identifying any structural characteristics that might be important for splicing. First, we investigate if intron sequences are more structurally stable than random sequences with the same sequence characteristics. We also analyze the stability of local structures in the vicinity of and at the splice signals, and investigate the stability of basepairing interactions between splice sites and snRNAs and its contribution to intron identification. Finally, we look for any conserved structural motifs in the vicinity of the splice signals.

## 6.1 Structural stability of introns vs. random sequences

There have been several attempts to assess the 'foldability' of naturally occurring RNA sequences – their tendency to fold into more stable secondary structures than expected by chance. In 1999, Seffens and Digby examined 51 mRNA sequences from several different organisms to determine if their

minimum free energies are more negative than for randomized mRNA sequences with the same composition and length (Seffens and Digby, 1999). They employed six different mRNA randomization procedures, randomizing either entire mRNAs, coding regions or untranslated regions and preserving either base composition or codon composition. For each native mRNA and each randomization procedure, they generated 10 randomized mRNA sequences, calculated their minimum free energies using the mfold algorithm and analyzed the differences, using what they call 'segment score'. Segment score for an mRNA sequence is number of standard deviations the mean of the randomized set is away from the native free energy. This value is more often called Z-score. In general, Z-score of a number $x$ with respect to a set $s_1 \dots s_N$ of numbers is defined by:

$$Z = \frac{x - \mu}{\sigma}, \tag{6.1}$$

where $\mu = \frac{s_1 + \ldots + s_N}{N}$ and $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(s_i - \mu)^2}{N-1}}$ are the mean and standard deviation, respectively, of the numbers $s_1, \dots, s_N$. If we want to compare the free energies of RNA sequences, $x$ would be the minimum free energy of the native RNA and $s_1, \dots, s_N$ would be the minimum free energies for $N$ randomized RNA sequences. In a recent comparison study of six RNA folding measures that estimate how well an RNA sequence folds, Z-score was found to be the most sensitive measure (Freyhult et al., 2005).

The study by Seffens and Digby found that the native mRNAs have significantly lower free energies than the randomized sequences: the average Z-score for the whole-randomized set (whole mRNA, base composition preserved) was $-1.23$, and for the coding-random set (only coding regions randomized) it was $-0.87$. The Z-score values were slightly higher for other randomization procedures.

This study was later challenged by Workman and Krogh, who re-examined the same set of mRNAs and concluded that the apparent higher stability of native mRNAs could be explained by differences in dinucleotide compositions between native and randomized sequences (Workman and Krogh, 1999). Their rationale was that since most of the algorithms for RNA

secondary structure prediction (including mfold) use a nearest neighbour thermodynamic model, which assumes that the stability of a specific basepair depends on the neighbouring bases, the RNA folding thermodynamic is strongly dependent on basepair stacking interactions. Therefore, in order to have a fair comparison between structural stability of native and randomized RNA sequences, dinucleotide content needs to be preserved.

Workman and Krogh (1999) used 46 mRNAs from the original set of Seffens and Digby (1999), and generated 10 randomized sequences for each of them preserving either mononucleotide or dinucleotide composition. When only mononucleotide content was preserved, they obtained similar Z-scores as did Seffens and Digby (1999) (Z-score $= -1.59$). When dinucleotide content was preserved, the Z-score values were much higher ($-0.20$ for the first order Markov random sequences). They concluded that mRNA sequences, in general, do not form more stable extended structure than random sequences. They pointed out that the analysis they performed applies only to global secondary structure of the molecules, while more local structural interactions, such as hairpin loops, could not be detected using this approach.

Workman and Krogh (1999) also discussed the calculation of p-values associated with obtained Z-scores, which is important for determining the statistical significance of the scores. The distribution of Z-scores for random sequences can be approximated by a normal distribution with mean 0 and standard deviation 1. Assuming this, the significance of Z-scores can be approximated using an extreme value distribution that captures the likelihood that the given Z-score of a biological RNA sequence is larger than the maximum Z-score from a collection of randomized versions of the same sequence. In other words, p-values associated with Z-scores can be computed as the ratio of random sequences with a Z-score lower than that of the native sequence. This analysis was further extended by Rivas and Eddy (2000), who estimated that in the case where 100 random sequences are generated for each native sequence, the Z-score for a single sequence has to be at least of the order of $-3.8$ to be considered significant (at the 0.01 significance level). For the significance level of 0.05 this value would approximately be $-2$.

For a set of native sequences we can compute the upper bound for the

average Z-score in order for it to be significant at the 0.05 significance level. Based on the assumption that the distribution of Z-scores for random sequences can be approximated by a normal distribution with mean 0 and standard deviation 1, the Z-score for a single sequence has to be at most $-1.64$ in order to be significant at the 0.05 level (this is the point on x-axis below which the area under the bell curve is 0.05). Since the distribution of average Z-scores for $n$ sequences can be approximated by a normal distribution with mean 0 and standard deviation $1/\sqrt{n}$, an average score of less than $-1.64/\sqrt{n}$ would be significant.

Clote et al. (2005) conducted a similar study of secondary structure stabilities of various types of RNAs, but with an improved randomizing procedure. While Workman and Krogh used a heuristic to perform a dinucleotide shuffle, Clote et al. implemented the provably correct procedure of Altschul and Erickson (1985), which guarantees a dinucleotide composition exactly identical to that of the native sequence. Their results are similar to those of Workman and Krogh (1999) and they show that for the entire mRNA, as well as in 5' UTR, 3' UTR and the coding region of mRNA, the folding energy is approximately that of random RNA of the same dinucleotide composition (the average Z-scores are: $-0.18$, $-0.11$, 0.17, $-0.14$, respectively).

All of the described approaches analyze the stability of global structure of mRNA and/or its main parts. However, local stable secondary structure, such as hairpin loops, would remain undetected by these approaches. Local secondary structure interactions are known to play a role in many different cellular processes, such as transcription, mRNA stability and localization, RNA processing and translation. Examples include formation of a stem within *S. cerevisiae* L32 pre-mRNA that serves as a binding site for the L32 protein, which in turn regulates the splicing of its own pre-mRNA (Eng and Warner, 1991); a stem-loop in the coding sequence of the yeast ASH1 gene that can localize ASH1 mRNA to the bud tip (Grover et al., 1999); secondary structure elements in bacterial 5' UTRs that reduce the rate of mRNA degradation through the inhibition of nuclease activity (Diwa et al., 2000); a stem-loop structure, called an iron response element, that is found

in the 5' UTR of ferritin and transferrin receptor mRNAs, and binds a specific protein that blocks the translation of this mRNA under low iron conditions (Casey et al., 1988; Hentze et al., 1988).

Stability of local secondary structures in the coding regions of mRNAs was analyzed by Katz and Burge (2003). They applied a new shuffling protocol that randomizes mRNAs, preserving dinucleotide composition, amino acid sequence and codon usage. They analyzed thousands of coding regions from 28 different organisms, including *S. cerevisiae*. For each native mRNA sequence, they generated 20 randomized sequences using their *DicodonShuffle* algorithm, and then folded native and randomized sequences in sliding windows of 50 bases (with step size of 10 bases). The folding free energy over all windows for native and randomized sequences is used to compute Z-scores. They obtained a Z-score of $-0.25$ for the coding regions of *S. cerevisiae*, for which they claim to indicate significant bias in favor of local RNA secondary structure. The authors also compared folding potential between intron-containing and intronless genes in yeast and found that the mean Z-score was significantly lower (p-value = 0.004) for the former (Z-score = $-0.50$ versus $-0.24$). This result suggests that the secondary structure in yeast exons might play a role in splicing.

### 6.1.1 Z-score analysis of global intron structure

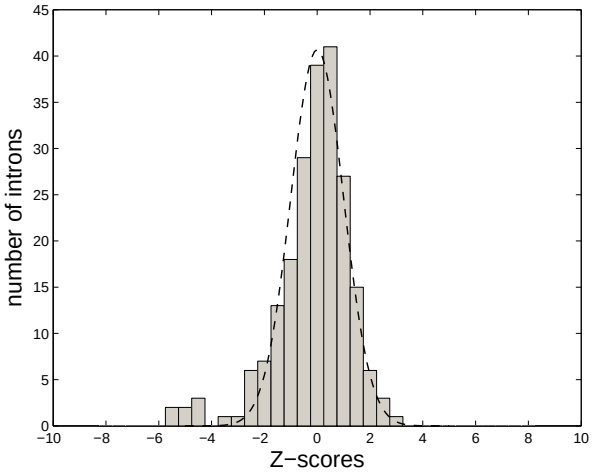
Since all of the previously described studies were done with coding mRNA sequences, we wanted to investigate if there is any bias towards secondary structures in yeast introns, both on a global and local level. For this purpose, we downloaded the Altschul-Erikson dinucleotide shuffling algorithm (Altschul and Erickson, 1985; Clote et al., 2005) from `http://clavius.bc.edu/~clotelab` (last accessed in April 2006), which is guaranteed to preserve the dinucleotide content of the native sequence. We analyzed the following three datasets separately: the entire STRIN intron dataset, long STRIN introns and short STRIN introns.

For each intron sequence in a dataset, we generated 100 randomized sequences using the Altschul-Erikson algorithm. We then folded all of the

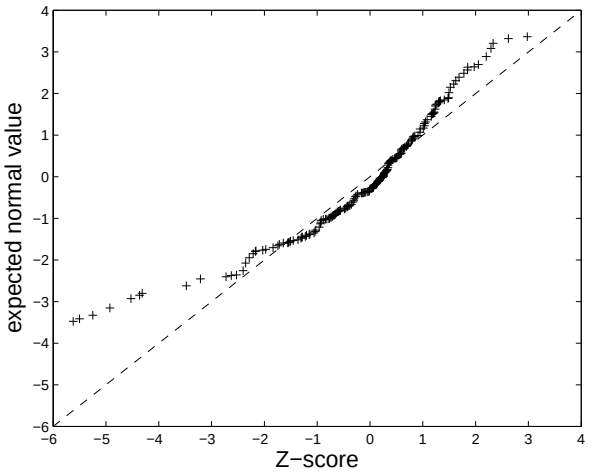
native intron sequences and all of the corresponding random sequences using the RNAfold program from the Vienna RNA secondary structure package (Hofacker et al., 1994). The computed minimum free energies for the native sequence and 100 corresponding random sequences with the same dinucleotide distribution were used to compute the Z-score for that sequence (see Equation 6.1). This was done for each sequence in a dataset. For each dataset, the distribution of Z-scores was plotted and the mean Z-score was calculated.

The Z-score distributions for all STRIN introns, long STRIN introns and short STRIN introns are shown in Figures 6.1, 6.2, and 6.3, respectively. Since Z-scores are expected to be normally distributed with mean 0 and standard deviation 1, we also superimposed the standard normal distribution over the histograms. If there was no bias for the global secondary structure in yeast introns the Z-score histograms would be expected to follow the plotted bell curve. We can observe that this is almost the case for the whole STRIN dataset and the short introns, while there is a slight deviation for the long introns. This result is also in agreement with the mean Z-scores, which are $-0.14$, $-0.30$, and 0.02 for STRIN introns, long STRIN introns and short STRIN introns, respectively. Based on our discussion in the previous section, the mean Z-scores for all STRIN introns, long STRIN introns and short STRIN introns have to be less than $-0.11$ ($-1.64/\sqrt{214}$), $-0.16$ ($-1.64/\sqrt{110}$) and $-0.16$ ($-1.64/\sqrt{2104}$), respectively, in order to be significant at the 0.05 level. Thus, STRIN long introns and consequently, all STRIN introns have a significant bias towards more stable secondary structures.

We also used quantile-quantile plots (q-q plots) to compare distributions of Z-scores with respective normal distributions (which have the same mean and standard deviation as the corresponding Z-score distributions). If Z-scores are distributed normally, the data points in the plot should fall approximately along the 45-degree reference line. The q-q plots in Figures 6.1, 6.2, and 6.3 confirm our previous conclusions. For the long introns there are many data points above the reference line, indicating that Z-score values are lower than expected if distributed normally. Data points for the short

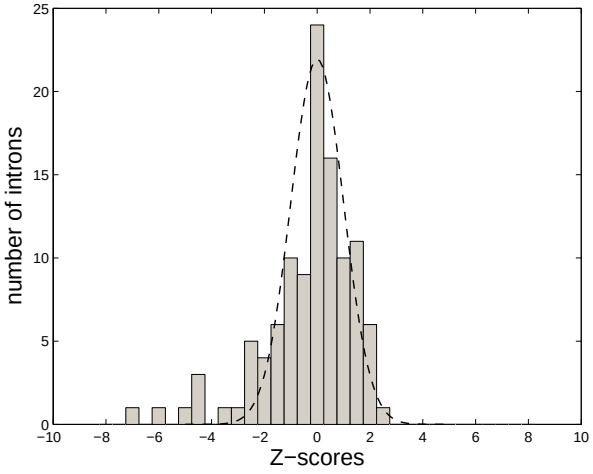



(a)

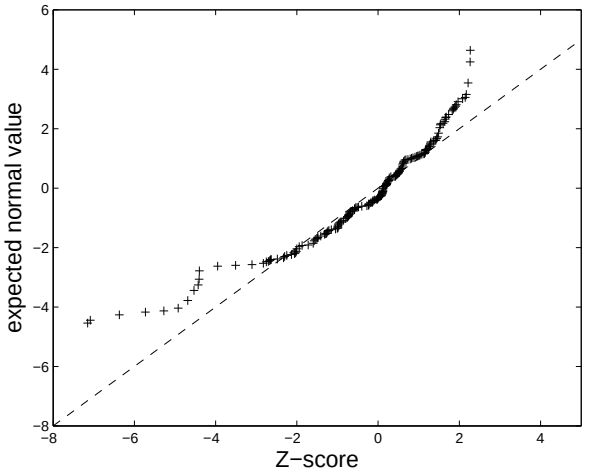



(b)

Figure 6.1: **(a)** Distribution of Z-scores for all STRIN introns. Standard normal distribution (mean = 0 and standard deviation = 1), that is expected distribution for Z-scores, is shown in dashed line. **(b)** Quantile-quantile plot of Z-scores against standard normal distribution.

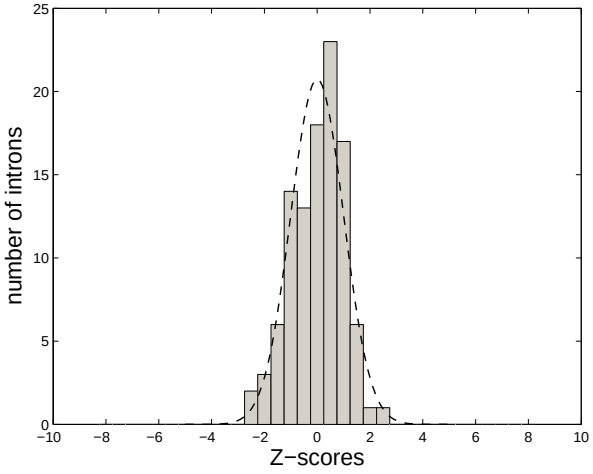



(a)

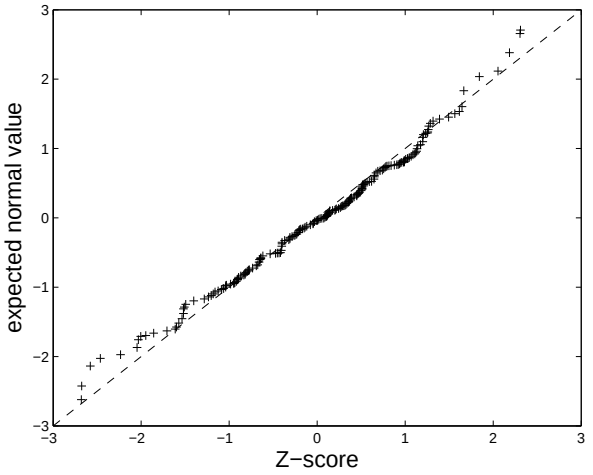



(b)

Figure 6.2: **(a)** Distribution of Z-scores for long STRIN introns. Standard normal distribution is shown in dashed line. **(b)** Corresponding quantile-quantile plot.




(a)




(b)

Figure 6.3: **(a)** Distribution of Z-scores for short STRIN introns. Standard normal distribution is shown in dashed line. **(b)** Corresponding quantile-quantile plot.

| **name** | **Z-score** | **molecular function** |
|---|---|---|
| YDR064W | −4.60 | ribosome component |
| YGL103W | −4.40 | ribosome component |
| YJL191W | −7.15 | ribosome component |
| YNL112W | −4.41 | RNA helicase |
| YNR053C | −3.58 | GTPase activity |
| YOL120C | −5.85 | ribosome component |
| YPL081W | −5.03 | ribosome component |

Table 6.1: STRIN long introns that have very low Z-scores.

introns lie very close to the reference line, indicating that the Z-score values are normally distributed as expected in the case with no structural bias.

The distribution of Z-scores for long STRIN introns has a longer left tail than the normal distribution, indicating that a number of STRIN long introns have very low Z-scores. This can also be observed for all STRIN introns (Figure 6.1), where the same long introns are outside the bell curve. We extracted all of the introns from the long STRIN intron dataset whose Z-score is lower than −3.5, which is close to the threshold for the statistically significant Z-scores (−3.8) calculated by Rivas and Eddy (2000). We looked at the annotated molecular function for each intron to see if there are any patterns. The results are shown in Table 6.1.

From the table, we are tempted to conclude that an unusual proportion of these introns are found in ribosomal proteins (5 out of 7). Since there are 91 ribosomal-protein introns in the STRIN long intron dataset, however, this is not unexpected. We also looked to see if there are any small nuclear RNAs encoded within these introns and found only one snoRNA (snR191) within the YNR053C intron. Small nucleolar RNAs are the class of small non-coding RNA molecules that guide chemical modifications of ribosomal RNAs and are frequently encoded in the introns of ribosomal proteins (which, interestingly, YNR053C is not). It is also interesting to observe that the YNR053C intron has the highest Z-score among the seven tabulated introns. Thus, the presence of a snoRNA does not seem to contribute significant structural stability to the YNR053C intron.

In summary, we can conclude that STRIN long introns show statistically significant bias towards more stable secondary structures and this is mostly

due to the presence of several introns with very stable secondary structures. This is also reflected in the mean Z-score for the entire STRIN dataset. STRIN short introns do not differ from random sequences in this respect. These finding are consistent with our hypothesis that secondary structure within long yeast introns is functionally important.

### 6.1.2 Z-score analysis of local structure

We also analyzed the stability of local structures for yeast introns, but our approach differs from Katz and Burge (2003) in that we do not scan the entire length of an intron but rather the 100-nt windows positioned at the 5' splice site, 3' splice site and the branchpoint sequence. The motivation behind this approach is that since it is known that the splice signals interact with snRNAs, we would expect to see a preference for structure-free regions around these sites. On the other hand, it is possible that local structural elements exist in the vicinity of the splice signals that serve as additional identifiers of intron location.

For each STRIN long intron, we isolated three 100-nt windows centered around the 5' splice site, 3' splice site and the branchpoint sequence. In each of these windows we slid a 50-nt window, with a step size of 10 nt, and for each of the sliding window positions we generated 100 random sequences using the Altschul-Erikson dinucleotide shuffling algorithm (Altschul and Erickson, 1985; Clote et al., 2005). Thus, for each sliding window position we extracted 98 sequence windows from the real long introns and 9800 random sequences (as in Section 5.5, we had to exclude 5' UTR introns from this analysis due to the problems with their annotated location). Next, we folded all of these 50-nt windows using the RNAfold algorithm, and used their predicted minimum free energies to calculate average Z-score for the entire 100-nt region, average Z-score for each sliding window position, and average values of free energies for all native and random sequences in the 100-nt region. The results are shown in Table 6.2.

The p-values in the table are calculated for the difference between the native and random average free energies using the Wilcoxon rank-sum test,

| **region** | **average Z-score** | **average** $\Delta G_{native}$ | **average** $\Delta G_{random}$ | **p-value** |
|---|---|---|---|---|
| **5' ss** | 0.31 | $-4.6$ | $-5.24$ | $6.38 \cdot 10^{-9}$ |
| **3' ss** | 0.12 | $-4.85$ | $-5.13$ | 0.12 |
| **branchpoint** | $-0.09$ | $-3.76$ | $-3.64$ | 0.15 |

Table 6.2: Average values for Z-score and free energies for native and random sequences in the vicinity of splice sites. The values are calculated for sliding windows of size 50 nt and then averaged over all sliding window positions. The p-values are calculated using the Wilcoxon rank-sum test.

which assesses whether the difference in medians between two observed distributions is statistically significant. Our results indicate that there is a slight, but statistically significant bias against stable local secondary structures in the region of $+/-$ 50 nt around the 5' splice sites. A similar bias is not evident for the 3' splice sites and the branchpoint sequences. We also calculated mean Z-scores for each sliding window position and plotted them in Figure 6.4.

We can observe that certain sliding window positions have relatively low or high scores compared to other values. The examples are the first four window positions for the 5' splice site, especially the second one (from $-40$ to $+10$ nt w.r.t. the 5' splice site), first and fourth window positions for the 3' splice site (from $-50$ to 0 nt and from $-20$ to $+30$ nt w.r.t. the 3' splice site, respectively) and the last window position for the branchpoint sequence (from 0 to $+50$ nt w.r.t. the branchpoint sequence). The null hypothesis that the datasets of the average MFEs for native and random sequences stem from the same underlying distribution is rejected by KS test for all of these window positions (p-values are 0.005, 0.0002, 0.002, and 0.002, respectively). We also plotted the distributions of minimum free energies for these window locations (Figures 6.5 and 6.6). It should be noted that the two windows centered at 25 nt upstream from the 3' splice site and at 25 nt downstream from the branchpoint sequence overlap for the majority of introns in the STRIN dataset since the distance between the branchpoint sequence and the 3' splice sites is usually shorter than 100 nt, and is often about 50 nt (see Figure 3.4). This explains, at least partially, why the average Z-score

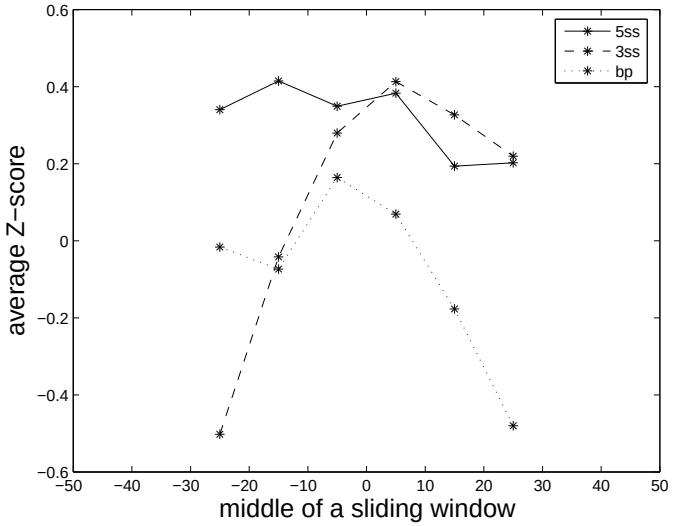



Figure 6.4: Average Z-scores for each sliding window position around the splice signals for 5'L STRIN introns. The size of the sliding window is 50 nt and the step size is 10 nt. The Z-values are plotted for the middle position of each sliding window (6 sliding window positions in total).

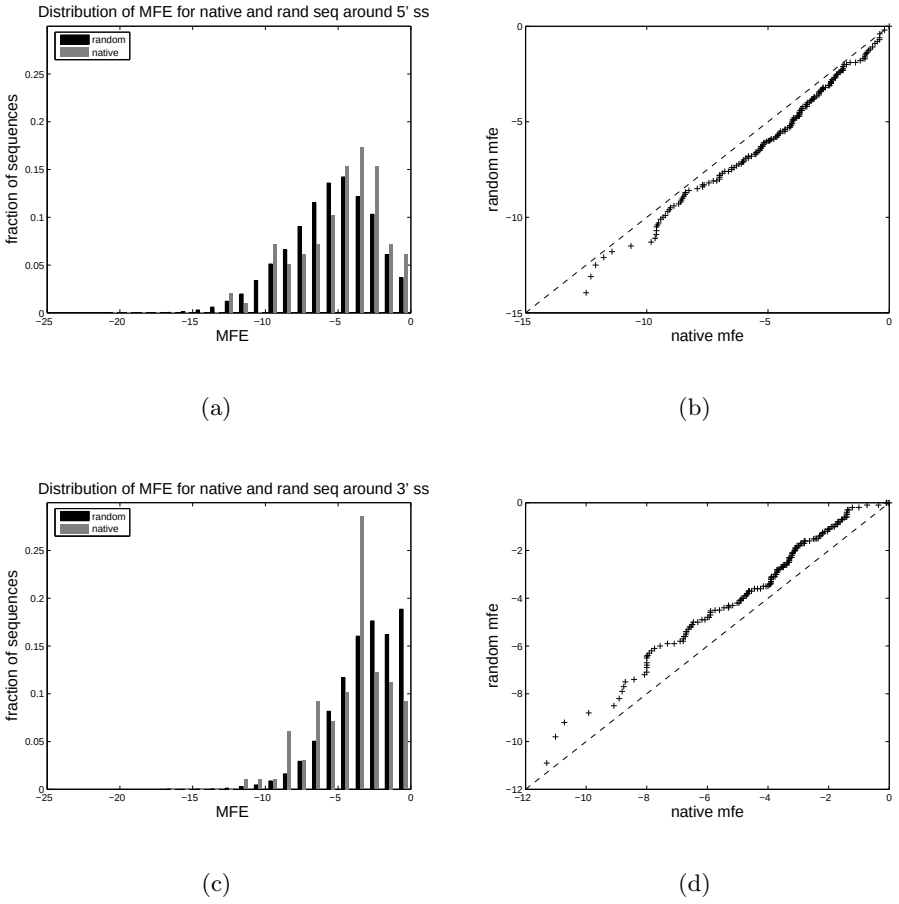
values for these two window locations are almost identical.

The plots in Figures 6.5 and 6.6 indicate slight biases of native sequences towards lower or higher minimum free energies when compared to random sequences. However, these biases are weak and could not be used in isolation as signals for computational identification of yeast introns. The reason for this is that in each case the distribution for native sequences almost entirely overlaps the distribution for the random sequence, and there is no threshold value which could differentiate the real from pseudo sites.

We repeated the same analysis for a smaller sliding window size of 20 nt and with the same step size of 10 nt. None of the statistical values calculated for this window size is statistically significant, indicating that true splice signals in long STRIN introns do not exhibit any strong bias towards or against very short local structures when compared to the random sequences.

Finally, we did the same analysis for all of the STRIN introns, including both long and short introns. The results are somewhat similar to the results for the long introns: the sequences in the vicinity of authentic 5' splice sites show statistically significant bias towards less local structure formation: the average minimum free energy of the native sequences is $\Delta G_{native} = -5.02$, while $\Delta G_{random} = -5.42$ (p-value $= 1.4 \cdot 10^{-9}$). We also calculated mean Z-scores for each sliding window position and plotted them in Figure 6.7. This plot is very similar to that for the 5'L STRIN introns (Figure 6.4). The window positions that have the highest absolute values of average Z-scores are the second and fourth window positions for the 5' splice site (from $-40$ to $+10$ nt and from $-20$ nt to $+30$ nt w.r.t. the 5' splice site), the first window position for the 3' splice site (from $-50$ to 0 nt w.r.t. the 3' splice site) and the last window position from the branchpoint sequence (from 0 to $+50$ nt w.r.t. the branchpoint). The null hypothesis that the datasets of average MFEs for native and random sequences stem from the same distribution is rejected by the KS test for the two window positions around the 5' splice site (p-values are 0.01, 0.0001, 0.2, and 0.09, respectively).

These results are in agreement with the molecular biology of the splicing reaction: recognition of the 5' splice site is the essential first step of the




(a) (b)

(c) (d)

Figure 6.5: Distributions of average MFE for 50-nt-sliding window from native long STRIN introns and generated random sequences: **(a)** from $-40$ to $+10$ nt with respect to the 5' splice site (p-value $= 0.005$) **(b)** q-q plot comparing distributions of native and random sequences **(c)** from $-50$ to 0 w.r.t. the 3' splice site (p-value $1.9 \cdot 10^{-4}$) **(d)** corresponding q-q plot.

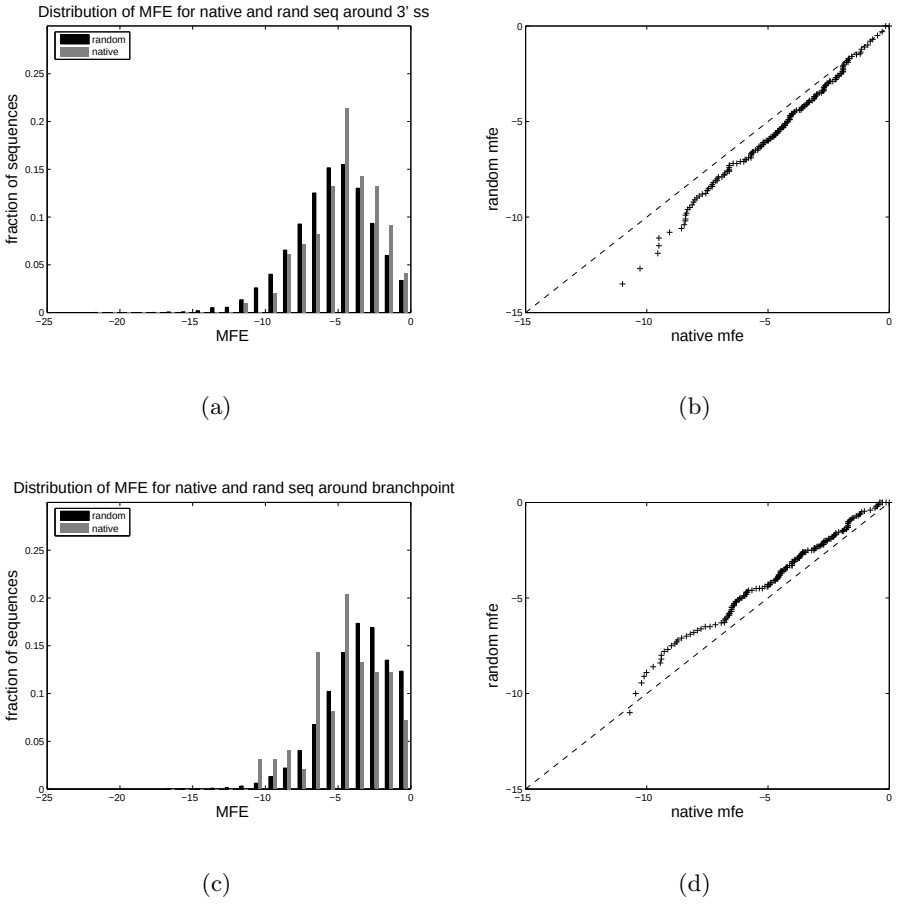



Figure 6.6: Distributions of average MFE for 50-nt-sliding window from native long STRIN introns and generated random sequences: **(a)** from $-20$ to $+30$ w.r.t. the 3' splice site (p-value = 0.002) **(b)** q-q plot comparing distributions of native and random sequences **(c)** from 0 to $+50$ w.r.t. the beginning of branchpoint sequence(p-value = 0.002) **(d)** corresponding q-q plot.

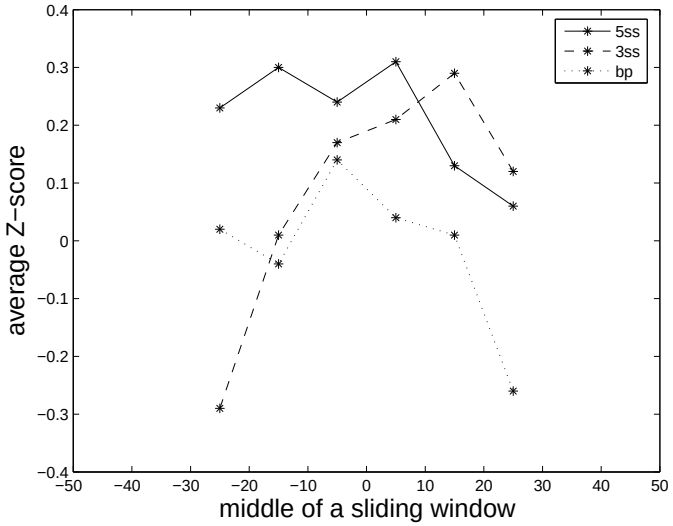



Figure 6.7: Average Z-scores for each sliding window position around the splice signals for all STRIN introns. The size of the sliding window is 50 nt and the step size is 10 nt. The Z-values are plotted for the middle position of each sliding window (6 sliding window positions in total).

process during which the 5' splice site is recognized multiple times: by U1, U6 and U5 snRNAs, as well as by a number of spliceosomal proteins that bind to RNA sequences in the close vicinity of the splice site. Thus, having the region around the donor site relatively free of secondary structure will allow these interactions to happen more easily.

## 6.2 Free bases at splicing signals

The structural analysis of the RP51B intron in Chapter 5 revealed that the branchpoint sequence was usually located in a loop for the mutants that were efficiently spliced (Figure 5.6). This loop structure was also observed for a number of STRIN long introns (see Section 5.5). Some of the earlier studies of pre-mRNA secondary structure also identified the branchpoint sequence to be located in single-stranded regions and showed that this structural configuration tends to be maintained in orthologous introns (Hall et al., 1988; Stephan and Kirby, 1993; Mougin et al., 1996; Chen and Stephan, 2003). Motivated by this observation, we used the branchpoint structural conformation as another indicator of efficient splicing when designing the mutants in Section 5.6.2.

In this section, we investigate whether the hypothesis that the branchpoint sequences tend to reside in loop regions has any statistical support. For this purpose, we considered the secondary structures of all STRIN introns. The introns were folded globally as for the branchpoint distance analysis in Chapter 5. We looked at two classes of subsequences: 7-nt real branchpoint sequences and all of the other 7-nt intronic sequences that do not overlap with the real sequences (non-branch sequences). The sequences themselves were extracted from the secondary structure predictions in dot-bracket notation (see Definition 1). For each sequence we counted the number of unpaired (free) bases specified by the '.' symbol. The distribution of free bases for real branchpoint sequences and non-branch sequences of the same length are plotted in Figure 6.8.

Although the two distributions completely overlap, the tendency of real branchpoint sequences to have more free bases is easily observable. If we

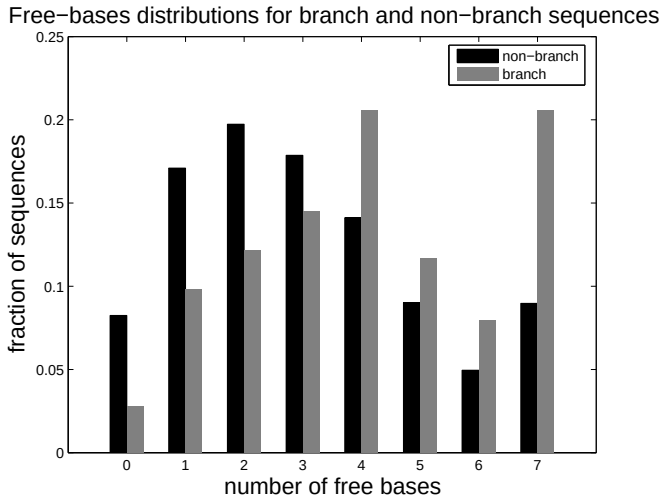



Figure 6.8: Distributions of the number of unpaired nucleotides in branchpoint and non-branch sequences when folded in global intronic secondary structures. The non-branch sequences are intronic sequence windows of length 7 nt that do not overlap with real branchpoint sites.

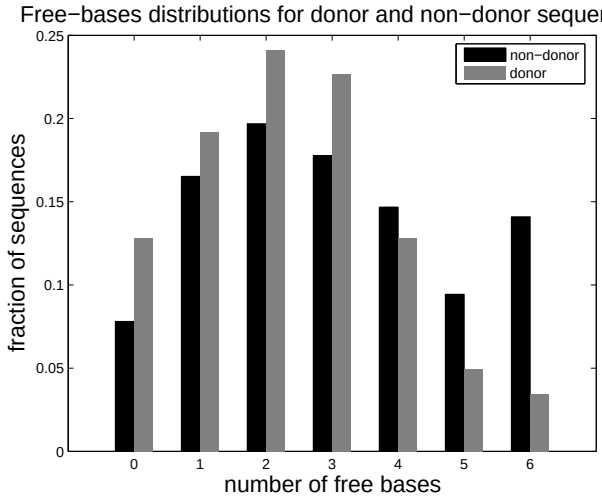
compare the two distributions using the Kolmogorov-Smirnov test, we obtain $D = 0.24$, with corresponding p-value $= 5.34 \cdot 10^{-11}$, rejecting the null hypothesis that the two sets of data stem from the same underlying distribution.

This finding is not in disagreement with the results from Section 6.1.2, where it was found that extended branchpoint sequences (+/− 100 nt flanking regions) resemble random sequences with respect to stability of local folding. In this section, we compute the global folding of introns but focus on structural conformations of intron subsequences. When the stability of 50-nt windows located at specific distance from the branchpoint sequences was considered, the closest ones (at location −30 to +20 w.r.t. the start of the branchpoint sequence) exhibited a very slight preference for less stable structures when compared to random sequences (Figure 6.4).

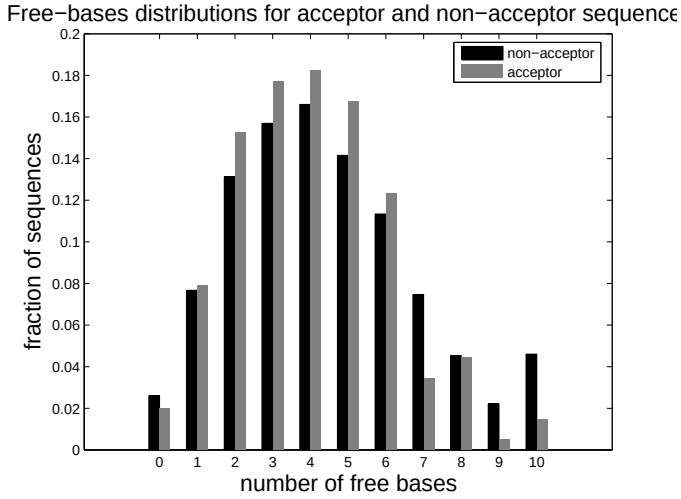
From the biological standpoint, it makes perfect sense not to have branchpoint sequences enclosed in very stable structures, which could hinder binding of the U2 snRNA. Currently, there are no experimental results supporting this conjecture, but it would be interesting to see if by enclosing a branchpoint sequence in a stable stem without changing any other structural characteristics of an intron we would observe decreased efficiency of splicing.

The same line of reasoning is applicable to the other two splicing signals – donor and acceptor sites. To check if there is the same statistical support as for the branchpoint sequence, we repeated the computational experiment but this time folded STRIN introns with 50 nt flanking regions so that the splice sites are found in a larger sequence context. For real donor sites, we used the first 6 nucleotides from the beginning of an intron (consensus sequence GUAUGU). Non-donor sites were all 6-nt intronic sequence windows that do not overlap with the real sites. Similarly, for acceptor sites we chose the last 10 intronic nucleotides to be the real sites (this includes a part of the polypyrimidine tract and conserved AG dinucleotide at the intron/exon boundary) and all 10-nt intronic sequence windows to be non-acceptor sites. The distributions of free bases are plotted in Figure 6.9.

The distribution for donor sites indicates a slight preference of real donor




(a)




(b)

Figure 6.9: Distributions of free bases for donor **(a)** and acceptor **(b)** sites. The length of real and pseudo donor sites is 6 nt, while the length of real and pseudo acceptor sites is 10 nt.

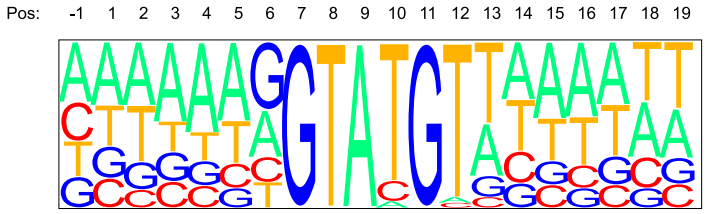
sequences to have less free bases than other intronic sequences of the same length. This difference between distributions is statistically significant when measured using the KS test at 0.05 confidence level ($D = 0.17$, p-value $= 1.28 \cdot 10^{-5}$). The histograms for acceptor and non-acceptor sites look almost identical, indicating no significant difference between the real and pseudo sequences ($D = 0.09$, p-value $= 0.07$).

Therefore, it seems that real donor sequences are enclosed in more stable secondary structures when folded globally with intron sequences, including 50-nt flanking regions. This might be due to the presence of some structural motifs that are recognized by spliceosomal proteins, or may be just a consequence of an unrealistic folding window.

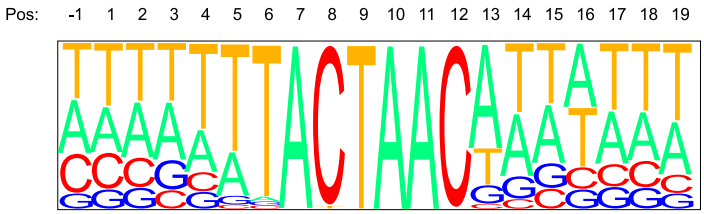
## 6.3 MFE of basepairing between splicing signals and snRNAs

As discussed earlier, the recognition of splice sites by the spliceosome is intriguing: it is done with great specificity, even though the known splicing signals, the 5' and 3' splice sites and the branchpoint sequence are relatively weak. This is especially true for higher organisms, where these signals are highly variable (Sun and Chasin, 2000). Even in *S. cerevisiae*, where the splicing signals are relatively well conserved, there are still numerous pseudo sites that resemble the real ones but are not functional (the pictograms in Figure 6.10 show the conservation of splicing signals in yeast). It is for this reason that current computational methods for splice site prediction and gene-finding have to use statistical properties of coding and non-coding sequences to improve the accuracy of intron detection (see Section 2.1). However, the number of false positive predictions still remains significant, especially when the search is performed on larger genomic regions.

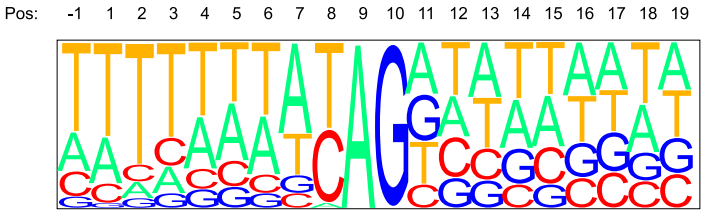
The question remains: how do cells achieve remarkable accuracy in recognizing real splice site without calculating the statistical properties of the adjacent regions? In cells, the conserved splicing signals are recognized by snRNAs, which bind to them, and these basepairing interactions between




(a)




(b)




(c)

Figure 6.10: Pictograms of donor (**a**), branchpoint (**b**) and acceptor (**c**) sites in *S. cerevisiae*. The pictograms were generated using the Web tool Pictogram available at http://genes.mit.edu/ pictogram.html (accessed in May 2006). The height of the letters in the pictogram is proportional to their relative frequencies at each position in the input sequences, considering the background distribution of bases (A=31%, C=19%, G=19%, and T=31% for all yeast ORFs with 1000-nt flanking regions). The input data used is generated from the STRIN dataset.

two RNA molecules follow the laws of thermodynamics, as do basepairing interactions within one RNA molecule. Therefore, it is plausible that the stability of these binding interactions is another indicator of splice site fidelity. This idea was explored in Roca et al. (2005), who examined the correlation between the splicing efficency of the human $\beta$-globin gene and wild type and mutated sequences of the first intron's 5' splice site. The authors showed that the metric that best explains the correlation is the MFE of pairing between the 5' splice site and the U1 snRNA. This result motivated us to analyze the structural stability of snRNA binding to splice signals in yeast.

### 6.3.1 snRNA-pre-mRNA interactions during splicing

The recognition of the 5' splice site is the first and essential step in gene splicing. The fidelity of the recognition is assured by the sequential binding of three snRNAs as well as by interactions with a number of spliceosomal proteins. First, a U1 snRNA binds immediately downstream of the 5' splice site. This basepairing typically involves the first six intron nucleotides, which are highly conserved in yeast but less so in humans (Brow, 2002). In yeast, the large majority of introns contain the sequence GUAUGU at the 5' border. This sequence is only partially complementary to the 5' region of the U1 snRNA, which interacts with the 5' splice site (Figure 6.11; the figure shows the secondary structure of human U1 snRNA, but the 5' end that interact with the 5' splice site is the same in both species). The mismatch between the fourth nucleotide in the GUAUGU sequence and U in the 5' end of the U1 snRNA, which is not present in higher eukaryotes where the fourth intron nucleotide is typically A (Figure 6.12), was shown to stabilize the interaction between the two RNA molecules, probably by allowing protein factors to bind to the distorted helix (Libri et al., 2002).

The 5' splice site is then recognized by U6 snRNA, which enters the spliceosome as a complex U4/U6.U5 tri-snRNP, but then after extensive RNA-RNA rearrangements it forms two helices with a U2 snRNA. The U6 snRNA basepairs with intron sequences at positions +4 to +6 (sequence

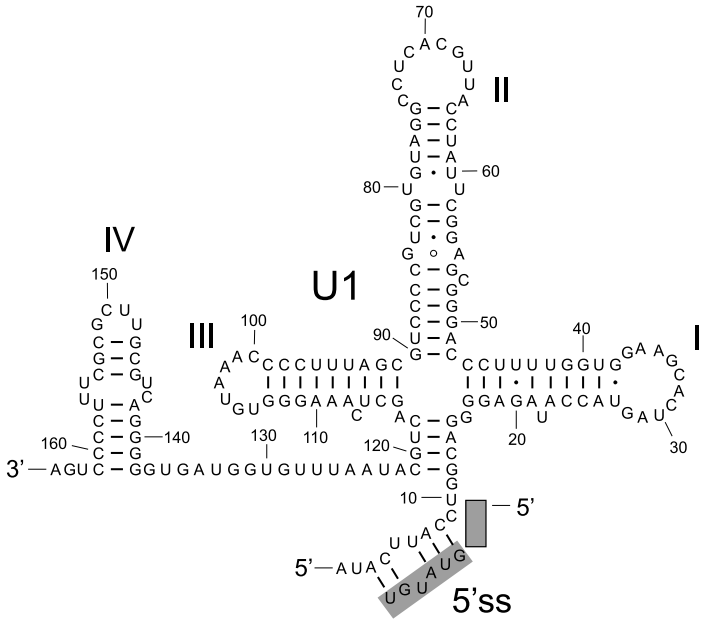



Figure 6.11: Structure of U1 snRNA (human) and its basepairing interaction with the first 6 intronic nucleotides (yeast). The pre-mRNA sequences are shaded, with the 5' exon shown as a shaded box.

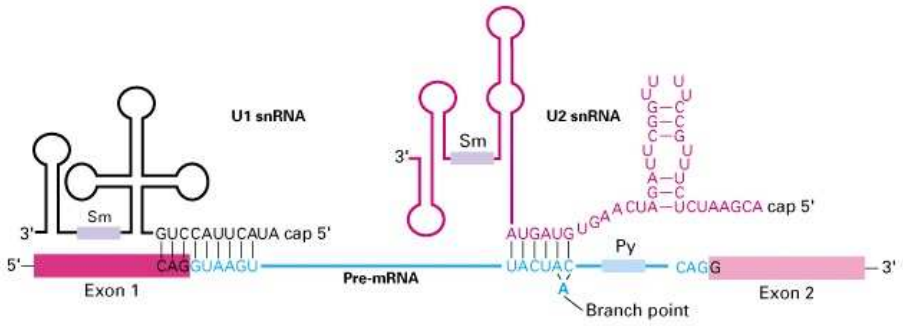



Figure 6.12: U1 and U2 snRNA interactions with a pre-mRNA molecule. The 5' splice sequence is typical for higher eukaryotes where the binding between U1 snRNA and the 5' splice site is more extensive than in yeast. The branchpoint sequence shown is typical for *S. cerevisiae* (http://www.library.csi.cuny.edu, May 2006).

UGU). The U6-U2 snRNA interaction as well as the basepairing between the 5' splice site and the U6 snRNA in yeast are shown in Figure 6.13. There has also been some indication that the basepairing between the U6 snRNA and the 5' splice site may be more extensive (Sawa and Abelson, 1992; Johnson and Abelson, 2001).

The U5 snRNA, as a part of the U4/U6.U5 tri-snRNP, also interacts with the sequences in the vicinity of the 5' splice site. Phylogenetic analysis of the U5 snRNA has revealed that it has an invariant 9-nt sequence within an 11 nucleotide loop (loop I in Figure 6.14). This loop interacts with both, the 5' and 3' splice site sequences, aligning them for the second step of splicing (Newman and Norman, 1992). These interactions are not precisely characterized but it has been suggested that the U5 snRNA forms basepairing interactions with the exon sequences immediately upstream of the 5' splice site before the first step of splicing (Newman and Norman, 1992), and later also forms non-specific interactions with the sequences around the 5' splice site (Alvi et al., 2001; McConnell and Steitz, 2001). The proposed interactions between the U5 snRNA and the 5' and 3' splice sites is shown in Figure 6.13.

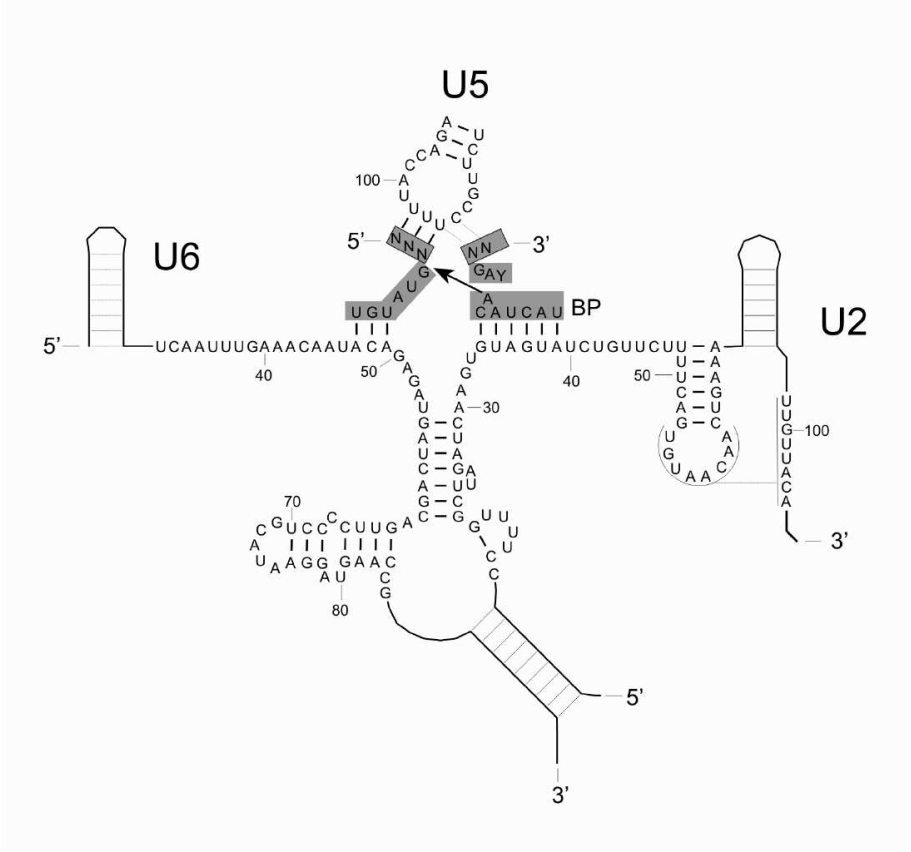



Figure 6.13: Secondary structure interactions within the tri-snRNP complex U4/U6.U5 and with a pre-mRNA in *S. cerevisiae*. The pre-mRNA sequences are shaded (GUAUGU sequence at the donor site, UACUAAC sequence at the branchpoint, YAG sequence at the acceptor site), with exons shown as shaded boxes. The arrow depicts the 'nucleophile attack' by branchpoint A, a chemical reaction that initiates the the cleavage at the 5' exon/intron junction.

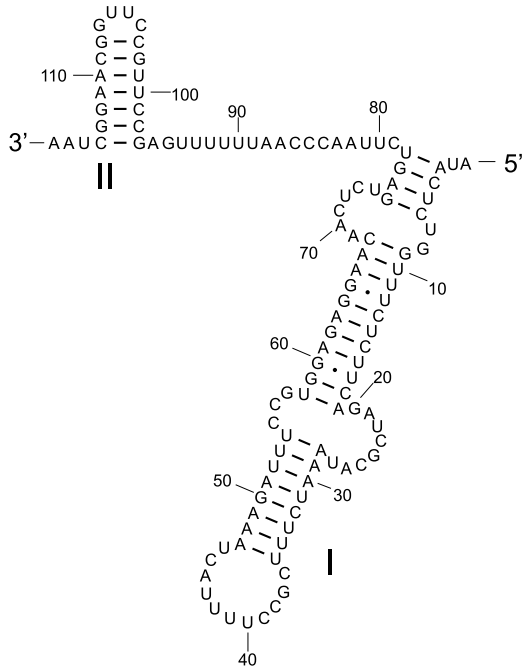



Figure 6.14: Structure of the human U5 snRNA. Two conserved stem/loops are labeled I and II. 11-nt loop I is the one that interacts with the 5' and 3' splice sites.

Similar to the 5' splice site, the branchpoint sequence is recognized through the basepairing interaction with another snRNA – the U2 snRNA. The branchpoint sequence is highly conserved in *S. cerevisiae*, with a consensus sequence UACUAAC. In other organisms, such as humans, the branchpoint sequence is less conserved. The branchpoint sequence basepairs with the sequences in the U2 snRNA, leaving the branchpoint adenosine (underlined A) unpaired and bulged out (Query et al., 1994), thus enabling the first transesterification reaction. The interaction of U2 snRNA with the branchpoint sequence is illustrated in Figures 6.12 and 6.13.

### 6.3.2 PairFold experiments using arbitrary pseudo sites

To test if the minimum free energy of folding between the U1 snRNA and the 5' splice site can serve as an additional statistical signal to aid in splice site identification, we first analyzed the average MFE of U1-snRNA binding in the +/− 100 nt region around the 5' splice site. For all intron sequences in the STRIN dataset, excluding the 5' UTR introns (because of the previously mentioned problems with annotation), we extracted donor sequences with their 100-nt flanking regions on both sides. For the U1 snRNA, we selected only the region that is known to interact with the 5' splice site: AUACUUACCUU. The underlined nucleotides are known to basepair with the first 6 intron nucleotides (Figure 6.11; the figure shows human U1 snRNA, whose 5' end has slightly different sequence – AUACUUACCUG). The selected U1 snRNA subsequence is larger than six nucleotides since it might be possible that there are additional interactions between the two RNA molecules that are important for binding stability but are not always present and thus are not observed in the donor consensus sequence (Figure 6.10 (a)). The 11-nt U1 snRNA subsequence is its 5' end, and it does not interact with the rest of the U1 molecule (see Figure 6.11).

For folding of two RNA sequences, we used the program PairFold (Andronescu et al., 2005), which predicts the minimum free energy pseudoknot-free secondary structure of two nucleic acid molecules. The selected U1-snRNA window was folded with a sliding window from the extended donor

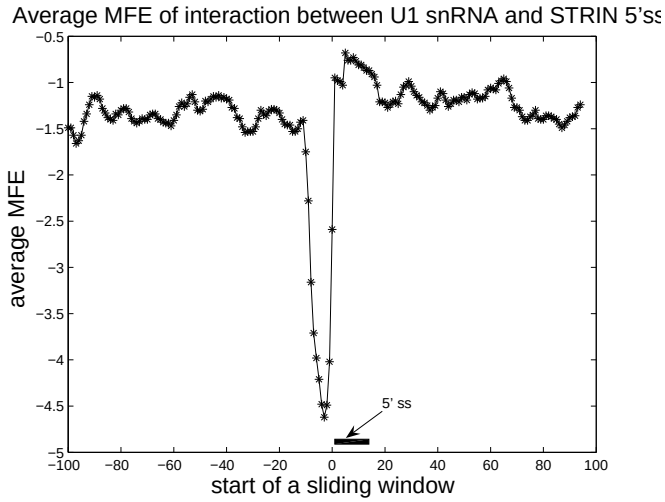



Figure 6.15: The average MFE of folding between U1 snRNA and sliding sequence window from the +/− 100 nt region around the donor site. For each sliding window position (window size = 11 nt, step size = 1 nt) the MFE of folding is averaged over all STRIN sequences.

sequence. The sliding window is the same size as the U1 snRNA window (11 nt) and the step size is 1 nt. For each sliding window position we calculated the average MFE of folding over all STRIN sequences and plotted the results in Figure 6.15.

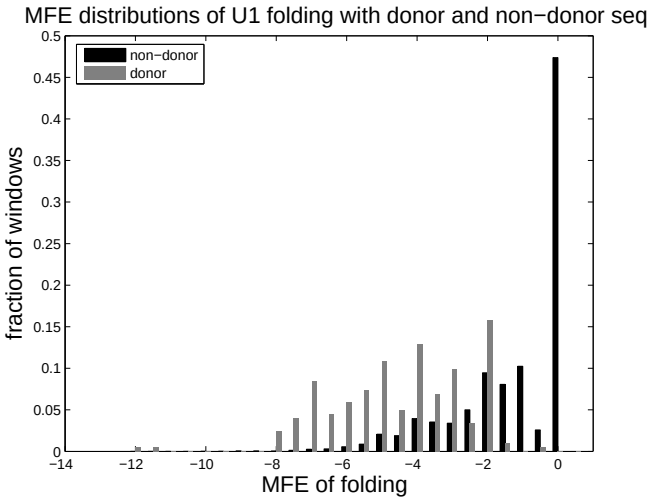
The prominent downward spike, representing the global minimum of average minimum free energies in the extended donor region, is located at position 98, corresponding to the pre-mRNA sequence window NNN| GUAUGUNN, where GUAUGU is the 6-nt consensus sequence at the 5' splice site (the vertical line indicates the exon-intron boundary). Thus, it appears that the most stable binding of U1 snRNA in the extended donor region is precisely at the 5' splice site. If we look at the minimum free energy distributions of U1-snRNA folding with donor (only sliding window position 98 considered) and non-donor (all the sliding windows that do not overlap with the 11-nt window centered around the conserved GUAUGU) sequences, we can observe that the real donor sequences have significantly more negative

MFEs than do pseudo donor sequences (see Figure 6.16; the KS test yielded $D = 0.68$, with corresponding p-value $= 1.23 \cdot 10^{-81}$). However, there is no clear separation between the range of the two distributions.

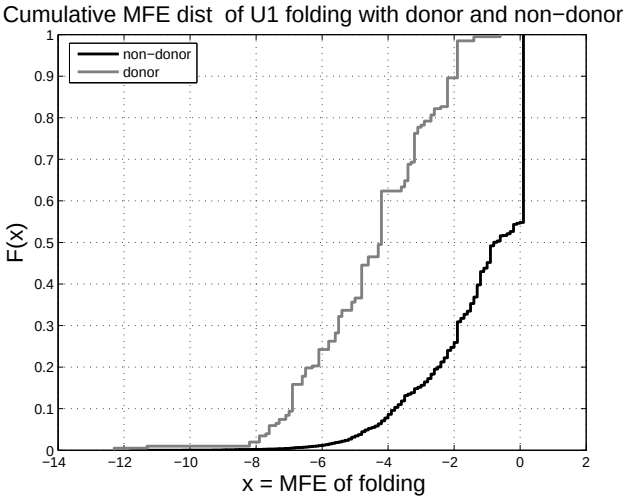
It is interesting to note that the 6-nt consensus sequence at the 5' splice site is not the ideal basepairing partner for the 11-nt segment of the U1 snRNA. When we calculated the MFE of folding between the U1 snRNA and all possible hexamers using the PairFold program, there were exactly 600 hexamers that had lower binding energy than the GUAUGU sequence (the MFE for GUAUGU was $\Delta G = -1.9$; the lowest observed MFE was $\Delta G = -7.4$ for GGUGAG). Therefore, it seems that the conserved sequence at the 5' splice site has not been evolutionarily optimized for U1-snRNA binding, but this is not surprising since the U1 snRNA is not its only basepairing partner. It has been shown that extending the basepairing between the U1 snRNA and the 5' splice site decreases the efficiency of U1 snRNA displacement and inhibits splicing at low temperature (Staley and Guthrie, 1999). It is possible that more stable interaction will prevent normal displacement of U1 snRNA and consequent binding of other snRNAs. The stability of the U1-snRNA-pre-mRNA duplex at the 5' splice site is probably insured by binding of some spliceosomal proteins (Libri et al., 2002).

Considering our results, the stability of interaction between the U1 snRNA and candidate 5' splice site would not be sufficient to discriminate between real and false sites, but might be used to filter some of the pseudo sites. This is in agreement with the findings in Roca et al. (2005), where only the 'strong' 5' splice sites (ones that give rise to more efficient splicing) had a strong correlation with the MFE of donor-U1 binding, while for the weaker sites other 5' splice site features were also important.

The global minimum present in Figure 6.15 is a result of fairly good binding between the U1 snRNA segment and the 11-nt donor sequence, as well as sequence conservation at that position. We repeated the experiment with 202 random sequences that have the same dinucleotide composition as the STRIN extended donor sequences and are centered around an arbitrarily chosen consensus sequence AAGTTC (the relative frequencies of bases are identical to those for STRIN donor sites (Figure 6.10 (a)) but the letters




(a)




(b)

Figure 6.16: Distribution histograms of folding MFE between **(a)** U1 snRNA and donor/non-donor sequence windows (window size = 11 nt) and **(b)** cumulative distributions of the same data.

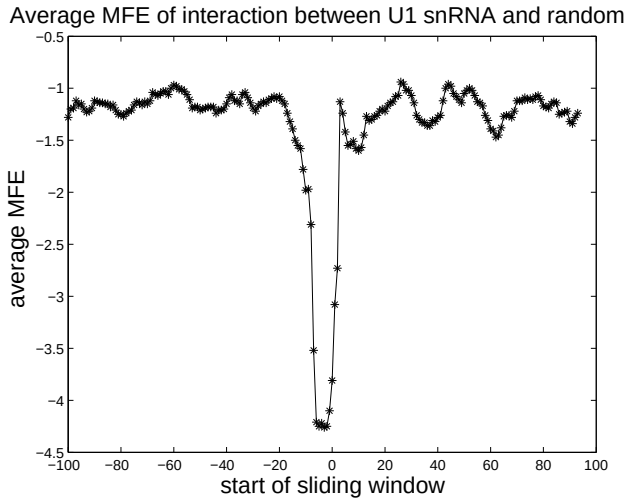



Figure 6.17: The average MFE of folding between U1 snRNA and sliding sequence window from a random sequence. For each sliding window position (window size = 11 nt, step size = 1 nt) the MFE value is averaged over all random sequences.

themselves are different) and obtained a similar downward peak as in Figure 6.15 (Figure 6.17).

We also repeated the same analysis with U6-snRNA-donor-site folding, U5-snRNA-donor-site folding and U2-snRNA-branchpoint-sequence folding and obtained similar results (least convincing for U5-snRNA-donor-site folding, which is expected from the nature of that interaction, which is not well characterized).

### 6.3.3 PairFold experiments using more specific pseudo sites

To further investigate the potential of the folding-energy signal to discriminate between real and pseudo splice sites, we used a slightly different approach, where the pseudo splice site sequences were selected more carefully to resemble the real splice signals. That approach is similar to the study done by Garland and Aalberts (2004), who used a modified version of mfold to calculate minimum free folding energies of U1-snRNA and 5' splice site inter-

action for real and pseudo sites from the Kulp/Reese human dataset (`http://www.fruitfly.org/sequence/human-datasets.html`). Since mfold predicts the MFE structure of a single RNA molecule, Garland and Aalberts concatenated 11-nt U1 snRNA (the same sequence we used in our previous analysis except that the 11th nucleotide in higher eukaryotes is G instead of U) and 10-nt sequence from the 5' splice site. The pseudo donor sequences were all GU dinucleotide extracted with a 10-nt surrounding window (as for real sites). In order to allow interaction between any two bases of the U1 snRNA and 5' splice site sequences, they also added 5-nt 'linker' sequences between the U1 snRNA and 5' splice site sequences. This linker sequence is not allowed to basepair, thus it forms a loop of a hairpin structure. To calculate the free energy of a two-sequence interaction, the authors modified mfold's energy parameters to exclude loop penalties.

Garland and Aalberts (2004) showed that there is an apparent difference between the free energies of real and pseudo sites, which is further emphasized if a simple filter is applied to the pseudo folding results (accepting only those sites that have the correct pairing between the GU at the potential donor site and the corresponding AC in the U1 snRNA sequence).

For a similar study on the STRIN dataset, we extracted real and pseudo donor sites using all STRIN non-5'UTR-containing genes with their 500-nt flanking regions. The sequence window that we used was different than in Garland and Aalberts (2004), and was chosen based on the results of our analysis in Section 6.3.2, where the global minimum of the average U1-snRNA-donor folding MFE was achieved for the sequence window NNN|GUAUGUNN. Thus, the dataset of real donor sequences contained all 202 sequences of the form NNN|GUAUGUNN (the GUAUGU sequence inside the window can vary since not all real donor sites have this consensus GUAUGA sequence – see Figure 6.10 (a)). Pseudo donor sites were constructed by taking every GU dinucleotide that appears in the dataset and extracting its surrounding window NNN|GUNNNNNN. The real donor sequences were excluded from the dataset of pseudo donor sites. There were 21712 pseudo donor sequences found. Pseudo donor sites selected this way resemble real donor sites more than do random sequence windows drawn

from the extended donor sequences, as in our analysis in the previous section. It is possible to be even more specific by using weight matrices as screening tools, which we did for the branchpoint sequences. For the analysis of donor sites, however, we chose to require only the presence of the canonical GU dinucleotide, allowing us to compare our results with findings in Garland and Aalberts (2004).

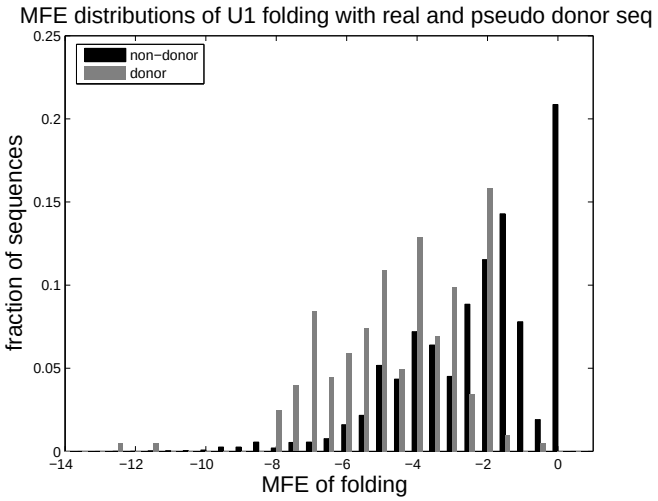
The real and pseudo donor sites were then folded with the 11-nt sequence from U1 snRNA (AUACUUACCUU) using the PairFold program. The minimum free energies of folding are plotted in Figure 6.18. The KS test for the real and pseudo donor site datasets yielded $D = 0.46$, with a corresponding p-value of $2.09 \cdot 10^{-37}$. This result suggests that even though the real sites still have significantly lower MFE of folding interaction with U1 snRNA than with the pseudo sites, this difference is smaller than in our analysis with random pseudo sites (Section 6.3.2). This is expected, since for the newly selected pseudo donor sites the presence of the GU dinucleotide increases the probability of pairing with the dinucleotide AC from the U1 snRNA sequence (AUACUUACCUU), resulting in a lower free energy of the interaction.

One way to use this thermodynamic approach for 5' splice site identification is to select a minimum free energy threshold for discriminating between the real and pseudo sites. To assess the accuracy of this approach for various MFE threshold values, we calculated the true positive and false positive rates. Let $TP$ be the number of real sites that were predicted as real and $FP$ be the number of pseudo sites predicted as real. If $Rcount$ and $Pcount$ are total numbers of real and pseudo sites, respectively, then we can define true positive (also known as sensitivity) and false positive rate in the following way:

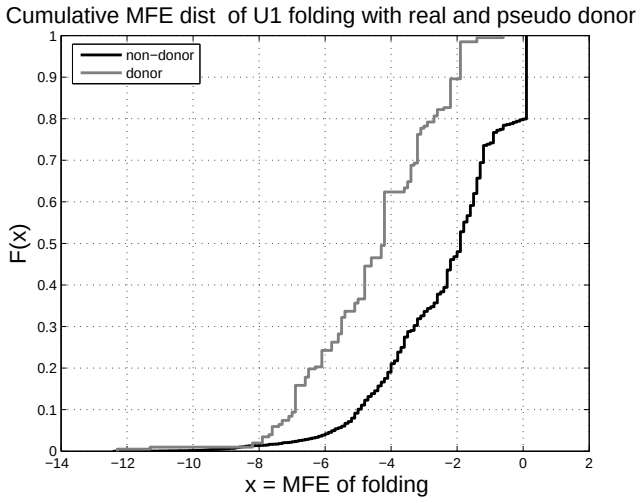
$$true\ positive\ rate = \frac{TP}{Rcount} \tag{6.2}$$

$$false\ positive\ rate = \frac{FP}{Pcount} \tag{6.3}$$

The typical way of displaying true positive and false positive rates for var-




(a)




(b)

Figure 6.18: **(a)** Distribution histograms of folding MFE between U1 snRNA and donor/non-donor sequence windows, where pseudo donor sites are required to have the consensus GU dinucleotide and **(b)** cumulative distributions of the same data.

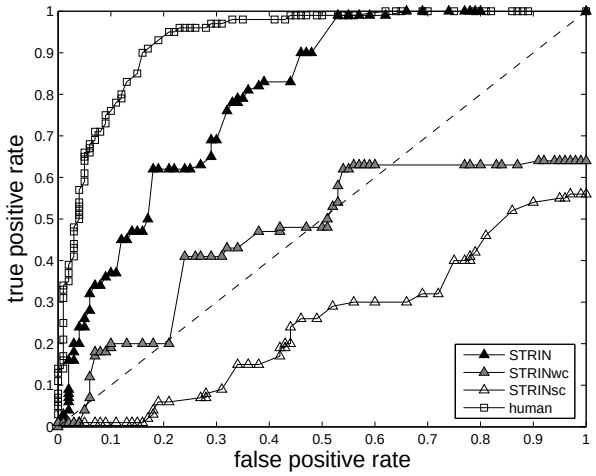



Figure 6.19: The Receiver Operating Characteristics (ROC) curves for the thermodynamic donor-identification approach. Different plots correspond to different selection of real and donor sites: all real donor sites and all pseudo sites containing a GU dinucleotide (STRIN); real and pseudo sites selected using the weak constraint (STRINwc); real and pseudo sites selected using the strong constraint (STRINsc); results for human data from Garland and Aalberts (2004). The dashed straight line at a 45-degree angle is known as the 'no-discrimination' line and it indicates no predictive ability.

ious threshold values is to use the Receiver Operating Characteristics (ROC) curve (Deleo, 1993). The ROC curve for our results is shown in Figure 6.19. We also repeated Garland and Aalberts' analysis on the Kulp/Reese human dataset and included the results, consistent with the ones in their study, for comparison.

To investigate if further filtering of candidate donor sites can improve the predictive power of the thermodynamic method, we used additional constraints. First, similar to the approach in Garland and Aalberts (2004), we rejected all candidate sites that did not have the expected pairing between the donor dinucleotide GU and the U1 dinucleotide AC. We call this the 'weak constraint'. We also used a more stringent constraint (the 'strong constraint'), rejecting all candidate sites that do not have the complete known

pairing between the donor sequence and the 5' end sequence of the U1 snRNA, as shown in Figure 6.11. We applied these filters on both real and pseudo donor sites, since if used for prediction, the method would not be able to differentiate between them *a priori*. This was not done by Garland and Aalberts (2004), and by applying their filter only on pseudo sites there were able to obtain better prediction accuracy (none of the real sites were excluded). Our results are shown in Figure 6.19. Comparing the results on human and STRIN data without any constraints, it is obvious that the thermodynamic approach in yeast is not as successful as in humans. One reason for this is certainly the nature of the secondary structure interaction between the U1 snRNA and the 5' splice site, which is more stable in humans due to the conserved A at the fourth intron position (GUAAGU) which forms one additional basepair with the U1 snRNA sequence (see Figure 6.12). In addition, higher eukaryotes have a slightly different sequence at the 5' end of the U1 snRNA (G instead of U – AUACUUACCUG) that can allow for basepairing with the exonic portion of the 5' splice site (Figure 6.12). It is also possible that the extended donor site consensus in humans is more optimal for U1-donor interactions than in yeast.

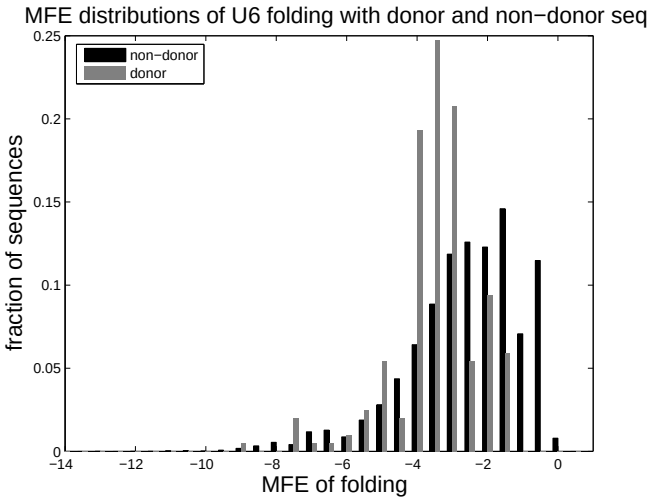
The reason why the ROC curves for filtered data are lower than for unfiltered data is that many real donor sequences are not predicted to form the required basepairs. From the plot, it can be seen that the true positive rate for the weak constraint goes only up to about 0.65 and for the strong constraint up to about 0.55. There are 122 real donor sites that satisfy the weak constraint and 114 that satisfy the strong constraint. One possibility, which we observed when inspecting our data more closely, is that the predicted structure has the same MFE as the required one and thus, even though the particular donor site has the required structure as its optimal, it would not be considered to satisfy the structural constraint. We found that 49 real donor sites belong to this category when the weak constraint is applied. Another possibility is that even though the basepairing between the 5' splice site and the U1 snRNA is not the most optimal one, there are some protein factors that stabilize it. Finally, as discussed earlier, the accuracy of computational secondary structure prediction is limited, and the structures

that we are predicting using PairFold might not be those present in nature.

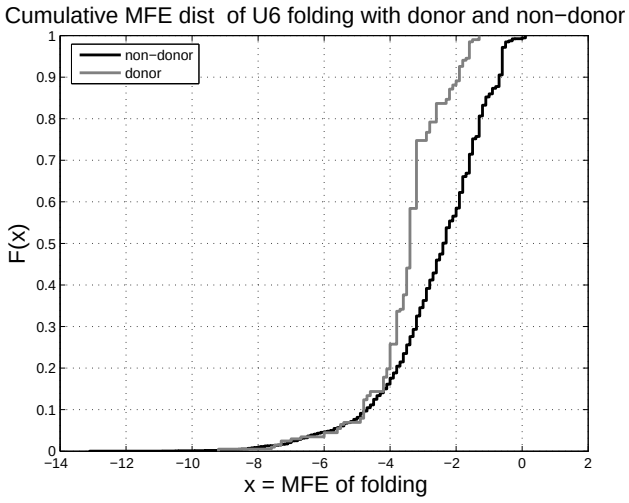
We repeated the same analysis for U6-snRNA-donor-site interaction and U2-snRNA-branchpoint-sequence interaction. For the former, the format of the donor sequence was chosen to be NNNN|GUAUGUNN, based on the position of the global minimum of the average U6-snRNA-donor folding MFE (same study as for the U1 snRNA and the 5' splice site interaction in Section 6.3.2, data not shown). There were 21697 pseudo donor sites selected. The U6 snRNA segment used for folding was AACAAUACAGAG. For branchpoint sequence analysis we extracted all real branchpoint sites of the form NNNUACUAACNNN. To extract the pseudo sequences, we used a positional weight matrix derived from real branchpoint sequences and extracted all 13-nt windows whose middle 7-nt, branchpoint-like sequence has a weight matrix score above the threshold. The threshold was selected as the minimum weight matrix score for all real branchpoint sequences. Due to the stringency of this approach, there were only 837 pseudo branchpoint sequences selected, not including the real sites. The U2 snRNA segment used for folding was AGUGUAGUAUCU. As for the U1-donor stability analysis, we used PairFold to fold extracted real and pseudo sites with their respective snRNAs. The results are shown in Figures 6.20 and 6.21.

Applying the KS test, $D$ statistic values of 0.42 and 0.43 and p-values of $5.46 \cdot 10^{-32}$ and $3.87 \cdot 10^{-27}$ were obtained for U6-donor and U2-branchpoint stability, respectively. This indicates a statistically significant difference between the real and pseudo sites with respect to the MFE of folding with their corresponding snRNAs, but overlaps between the distributions are significant and it is impossible to achieve good separation using any threshold value. This is confirmed by the ROC curves (Figure 6.22), which are close to the no-discrimination line (dashed line).

To summarize, the results of this section show that the real 5' splice sites and branchpoint sequences have lower minimum free energy of folding with snRNAs than do the pseudo sequences. However, these differences are not discriminative enough to use the thermodynamics splice-signal-identification approach in isolation. If combined with other identification methods, how-

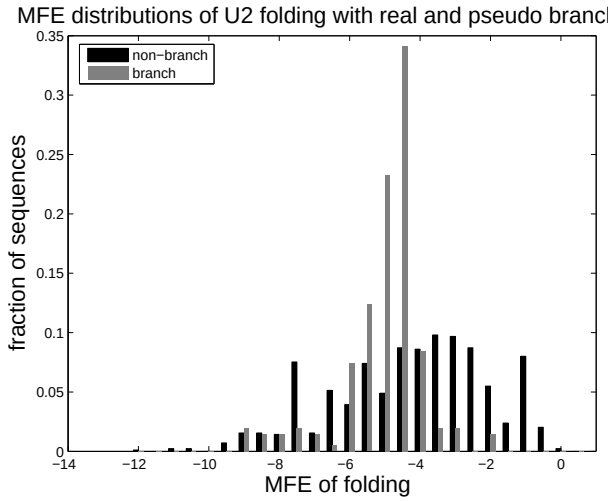



(a)

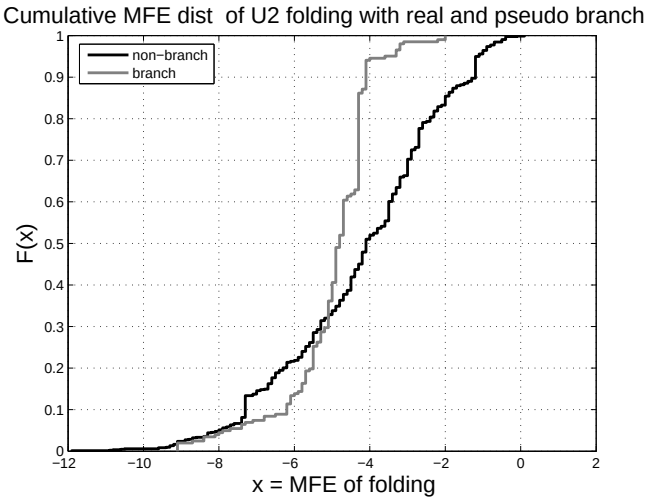



(b)

Figure 6.20: **(a)** Distribution histograms of folding MFE between U6 sn-RNA and donor/non-donor sequence windows where pseudo donor sites are required to contain the consensus GU and **(b)** cumulative distributions of the same data.




(a)




(b)

Figure 6.21: **(a)** Distribution histograms of folding MFE between U2 snRNA and branch/non-branch sequence windows where pseudo branchpoint sites were found using positional weight matrix and **(b)** cumulative distributions of the same data.

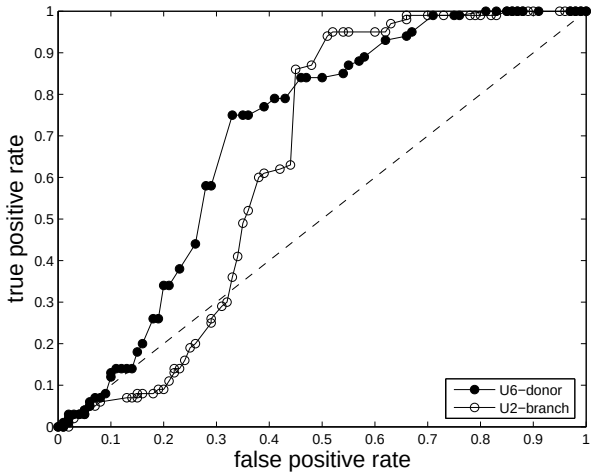



Figure 6.22: ROC curves for the thermodynamic donor- and branchpoint-identification approach.

ever, these weak statistical signals could potentially improve the accuracy of prediction.

## 6.4 Phylogenetic analysis of sequences around splice signals

In Section 6.1.2, we investigated whether there are any stable local structures in the vicinity of splice signals. The approach considered only the minimum free energy of observed structures and was dependent on their exact location with respect to the splicing signals. In Section 6.2, we conducted a simplified analysis of structural configurations of splice signals, considering only the number of unbound bases at the signals.

These studies, although providing insights into general 'foldability' of extended donor, acceptor and branchpoint sequences, did not attempt to identify any specific structural signals that might be present in the considered sequences. As discussed in Chapter 2, the spliceosome is a large

ribonucleo-protein complex that contains a large number of protein-splicing factors. Many of these proteins bind to pre-mRNA during the process of splicing. The protein binding to the RNA molecule can be either sequence specific, structure specific or a combination of both. There are several different classes of RNA-binding proteins that recognize specific structural motifs. Proteins containing an arginine-rich motif can recognize stem-loops, internal loops and bulges; the structure of these RNA motifs, rather than sequence, is the major binding determinant (Burd and Dreyfuss, 1994). There are also double-stranded RNA-binding proteins that, as their name indicates, recognize double-stranded RNA regions but this binding is also sequence dependent. Another class of RNA-binding proteins are those that contain RNA-recognition motifs. An example is the A protein from the U1 snRNP (which is therefore directly involved in splicing) that is known to recognize a specific stem-loop in the U1 snRNA (Zamore et al., 1990). Proteins with RNA-recognition motifs can recognize wide varieties of RNA structures.

We are not aware of any studies that discuss the binding of protein splicing factors to structural RNA motifs in the vicinity of splice signals. However, given the facts that a large number of proteins interact with the pre-mRNA molecule during splicing, and the ability of certain classes of proteins to recognize binding sites based on their structural conformation, we believe that there is a reasonable chance that these kinds of RNA-protein interactions occur during the splicing process. If this is the case, we would expect to find specific structural motifs in these regions.

In order to investigate this hypothesis, we employed two different approaches: the first attempts to identify conserved structural motifs in the *Saccharomyces sensu stricto* species, and the second searches for common motifs in the neighbourhood of splice signals of all STRIN introns.

### 6.4.1 Phylogenetic analysis of *sensu stricto* species

We use comparative structure analysis to search for the conserved structural motifs in the four *sensu stricto* species, discussed in Section 3.3. We focused on the regions around splice site signals that are binding targets for the

splicing protein factors. Based on the finding that splicing enhancers are usually found within 100 nt from the splice sites (Hertel et al., 1997), we chose to extract 200-nt windows centered at the splice signals.

The phylogenetic dataset described in Section 3.3 contains only alignments of intronic sequences, thus we needed to obtain the 100-nt regions upstream of the donor site and downstream of the acceptor site. To do so we downloaded aligned *sensu stricto* genes (ORFs) with the flanking regions. The alignments were found at `http://www.broad.mit.edu/ftp/pub/annotation/fungi/comp_yeast/` (June 2006). These alignments are different from those used in Section 3.3: availability of an intron alignment for a certain gene does not guarantee the availability of the alignment for that gene and vice versa. We searched for genes that have intron alignments for all four species (explained in Section 3.3). Alignments for two of these genes (YBR090C and YAL030W) were not found, which left 95 alignments of extended gene regions. Next, we extracted the 200-nt long alignments centered at donor, acceptor and branchpoint sequences. There were 43 alignments where the *S. cerevisiae* sequence was different than in the SGD database, and we therefore could not identify the correct position of the splice signals. These sequences were excluded from the analysis (52 alignments remained in the dataset). The last filtering step was to exclude the alignments that did not have all four *sensu stricto* sequences aligned. The final dataset consisted of 4-specie multiple sequence alignments of extended donor, acceptor and branchpoint sequence for 29 introns.

Based on the results of our analysis in Section 4.4, we chose the program Pfrali (Hofacker and Stadler, 1999) to search for conserved local structures. Like in Section 4.4.2, we post-processed Pfrali's output to keep only base-pairs that were inconsistent with at most one sequence. The output of this post-processing step is a secondary structure in dot-bracket notation that contains all locally conserved substructures.

We analyzed Pfrali's results in two ways: we looked at the structural context of the splice signals and we also enumerated all found structural motifs. When inspecting the structural context of the donor, acceptor and branchpoint sequences, we mostly focused on the presence of hairpin loops

at or adjacent to the splice signals. For 29 Pfrali predictions of conserved sequences in the neighbourhood of donor sites, we identified 15 cases where the 6-nt donor sequence (consensus sequence GUAUGU) was either adjacent to or in a stem region. This included three cases where the donor sequence was in a hairpin loop. For acceptor sequences, we identified 11 cases where the 3' splice site was either adjacent to or in a stem region. In 14 other cases, there were no structural elements present in the vicinity of the 3' splice site. Finally, for the Pfrali predictions of conserved structures in the neighbourhood of branchpoint sequences, in 26 cases the branchpoint sequence had one or fewer paired bases, and in two cases it had two. There is only one case where the sequence was enclosed in a stem region.

Overall, this analysis failed to offer any conclusive evidence for a specific conserved structural conformation at the donor or acceptor site. However, we find the results for the branchpoint sequences interesting and supportive of the analysis in Section 6.2 where we found evidence that the real branchpoint sequences tend to have more free bases than do random sequences.

Next, we catalogued all stem-loop motifs in Pfrali's predictions. We looked for stem-loop structures that have at least two adjacent basepairs in the stem (to avoid isolated basepairs) and have sequence length $\leq$ 20 nt. For each type of stem-loop structure, we counted the number of occurrences in 29 Pfrali predictions.

In Pfrali's predictions for aligned extended donor sequences, we found 24 stem-loop motifs that satisfy the requirements. The most common motifs were:

```
((......)) 4
((.......)) 4
(((......))) 4
```

The number to the right indicates the number of donor sequences containing these conserved motifs. For aligned extended acceptor site sequences, the following motifs were the most frequent:

```
(((......))) 8
```

```
((......)) 6
((.....)) 5
```

The most common motifs for aligned branchpoint sequences were as follows:

```
(((......))) 7
((......)) 5
(((.......))) 4
```

Clearly, the number of instances of each of the motifs is not significant, which leads us to conclude that there are no conserved secondary structure elements in the vicinity of splice signals among *sensu stricto* species. However, we need to be aware that this approach is limited: it assumes that not only the secondary structure but also the approximate location (including the alignment gaps) is conserved. This does not have to be the case – many functional structural and sequence motifs can be found at variable positions and still perform their functions. Therefore, it might be preferable to look for motifs that have similar structure and sequence characteristics but can be located anywhere in a given set of related sequences. One way to do this is using covariance models for motif description, which we explore in the next section.

### 6.4.2 Searching for common motifs in STRIN introns

In the following study, we decided to look beyond conservation among closely related species and to search for common structural motifs found in the neighbourhoods of splice signals of STRIN introns. The rationale is that if certain protein factors bind to specific structural signals, those signals should be present in the majority of introns with their flanking regions. We perform this analysis using covariance models, which do not require high sequence similarity of input sequences, but instead automatically build probabilistic models that flexibly describe the secondary structure and primary sequence consensus motifs in unaligned input RNA sequences.

For our analysis, we used the program CMfinder developed by Yao et al. (2006). CMfinder uses expectation maximization to search for high-scoring motifs that are described using covariance models. It employs a Bayesian framework for integration of information-based and folding-energy-based approaches to predicting structure. CMfinder can be applied to unaligned input of unrelated sequences with low sequence similarity, which is suitable for the dataset that we have. It is also capable of identifying motifs that are present only in a subset of input sequences.

We downloaded the CMfinder software package, version 0.2, from `http://bio.cs.washington.edu/yzizhen/CMfinder/` (April 2006). This program has a number of input parameters, the following being of consequence for our analysis:

- Number of stem-loops in a motif, for which we used the default value of 1.
- Number of motifs to be outputted, for which we used the default value of 3. The authors claim that this number is sufficient when looking for single stem-loop motifs.
- Minimum length of a motif, which has to be in the range of 15-250 nt. We chose the lower bound, since we are looking for simple stem-loop motifs.
- Maximum length of a motif, for which we chose 30. The actual output motifs can be slightly outside the specified range.
- Fraction of sequences containing the motif, for which we selected a relatively low value of 0.3. This parameter does not specify how many sequences actually contain the motif instances in the final output, but simply serves as a preliminary guess.

For each intron in the STRIN dataset, excluding 5' UTR introns (see discussion in Section 5.5), we extracted 200-nt windows centered at the donor, acceptor and branchpoint sequences. The extended donor (acceptor, branchpoint) sequences were considered as one input dataset for CMfinder.

The program was run on each of three datasets. The output of the program consists of motifs in Stockholm format, a system for marking up features in a multiple alignment (`http://www.cgb.ki.se/cgb/groups/sonnhammer/Stockholm.html`), as well as a covariance model for each motif. For each motif instance found, the following information is given: the name of the sequence where it is found, start and end location in the sequence, a motif weight that is proportional to a candidate's probability to be a motif instance, and the alignment score (alignment between a covariance model and a sequence).

For each of the donor, acceptor and branchpoint datasets CMfinder outputted three motifs. For each motif found, it is possible to obtain a summary of features using the subroutine *summarize* from the CMfinder software package. Output values from the procedure include the number of sequences in the dataset that contain motif instances, the sum of weights for all of the instances found, the weighted average of alignment scores and the average folding energies of motif instances. The predicted motifs and motif features for extended donor, acceptor and branchpoint sequences are shown in Table 6.3.

The given consensus structure and sequence describe all instances of the motifs found in the STRIN sequences. A matching pair of '<' and '>' represents a basepair, '-' represents a conserved base in a loop and '.' are insertions relative to the consensus. All of the motifs found are hairpin loops and the majority of motif instances have uninterrupted stems with contiguous basepairs. Only a smaller portion of motif instances have internal bulges and loops.

The results shown in the table indicate that each motif was found in almost all input sequences. The values for the sum of weights over all motif instances are quite variable, while there is no larger variation for the weighted averages of alignment scores. The average folding energies are relatively high, mainly due to the small size of the motifs found.

It is hard to say if these results are significant or not, since we have no reference values for either weights or alignment scores. Yao et al. did not provide any guidelines on how to interpret motif scores and weights when

| motif name | structure & sequence consensus | # of motif instances | sum of weights weights | avg. align. score | avg. folding energy | avg. cmsearch score |
|---|---|---|---|---|---|---|
| donor motif1 | <<.<.<.<<..-..-.-..-....>..>.>.>.>><br>cg.c.c.cc..c..G.u..A....g..g.g.g.cg | 196 | 1.57 | 7.89 | −1.60 | 4.11 |
| donor motif2 | .<<.<<..-...-.-.-.-..-..>>>>.<br>ugg.Gc..A...A.u.U.A..a..gCccu | 200 | 0.39 | 8.12 | −0.87 | 3.65 |
| donor motif3 | ..<<<.<<----->>.>>>..<br>uuccU.cGCcgGCCg.GggCU | 181 | 28.48 | 14.02 | −0.34 | 5.55 |
| acceptor motif1 | ..<<<<---->>>>..<br>ggaCaauuUUuuGuuU | 196 | 50.98 | 18.70 | −2.96 | 6.63 |
| acceptor motif2 | .<<<<<<-........----....->>..>>>>.<br>UuaGuGaG........AUUu....CuC..aCuaA | 165 | 10.54 | 18.71 | −2.96 | 3.85 |
| acceptor motif3 | .<<<<..----..->>>>.<br>ucguC..CAAA..AGacgc | 200 | 58.51 | 10.43 | 0.49 | 7.09 |
| branch motif1 | .<<<..<<..-..---...>>>.>>.<br>gCgc..cc..A..AGU...ggg.cGu | 201 | 11.32 | 10.94 | −3.14 | 5.43 |
| branch motif2 | .<<<..<<..<<...---...-...>>>.>...>..>>.<br>ucgG..cg..aG...AUu...u...Cuu.g...C..cga | 182 | 5.72 | 18.59 | −6.38 | 4.26 |
| branch motif3 | .<<<<<...---..--..>>>>>.<br>GGgGaC...AGc..CA..GuCgCg | 196 | 15.63 | 14.89 | −1.67 | 5.29 |

Table 6.3: Motifs predicted for extended donor, acceptor and branchpoint sequences and their features: number of sequences from STRIN dataset in which motif instances are found, sum of weights for all instances found, average weighted alignment score (alignment between covariance model and sequence), and average folding energy of the motif. The last column of the table is the average score from the program *cmsearch* (see text).

searching for unknown motifs. In order to access the significance of our findings, we once again resorted to using random sequences.

Similar to the analysis in Section 6.1.1, we used the Altschul-Erikson dinucleotide shuffling algorithm (Altschul and Erickson, 1985; Clote et al., 2005) from `http://clavius.bc.edu/~clotelab`, which is guaranteed to preserve the dinucleotide content of a native sequence. For each extended donor (acceptor, branchpoint) sequence, we generated 10 randomized sequences using the Altschul-Erikson algorithm. We used these shuffled sequences to assemble 10 datasets of random sequences, each having the same number of sequences as the original extended donor dataset. The first sequence in each of the random datasets was a shuffled first donor sequence, and so on.

Next, we searched for occurrences of the donor motif $i$ ($i = \{1, 2, 3\}$) in the datasets of random sequences. As mentioned before, CMfinder outputs a file that describes the learned covariance model for each motif found. This model can be used to search other sequences and databases for instances of the model. The CMfinder software package provides a tool to perform this kind of search – *cmsearch*. This program was originally developed by Sean Eddy's lab and is part of the Infernal package (`http://www.genetics.wustl.edu/eddy/infernal/`) (Eddy, 2002).

The cmsearch program takes as an input a file with a covariance model and a FASTA file of sequences. It searches both strands of each sequence in the sequence database, and returns alignments for high-scoring hits. Cmsearch calculates the alignment scores in the same way as CMfinder, thus the scores are comparable.

We ran cmsearch for each of the three covariance models (for the three motifs found) on 10 shuffled donor (acceptor,branchpoint) datasets. For each sequence we selected only the highest scoring hit on the forward strand and calculated the average scores for each of the 10 random datasets. We ran cmsearch on the real donor (acceptor, branchpoint) sequences in order to be consistent (also, the weighted average scores given by CMfinder's summarize procedure use weights calculated by CMfinder, and cmsearch does not calculate these). The average cmsearch alignment scores for each found motif and each dataset of real splice signals are given in the last column of

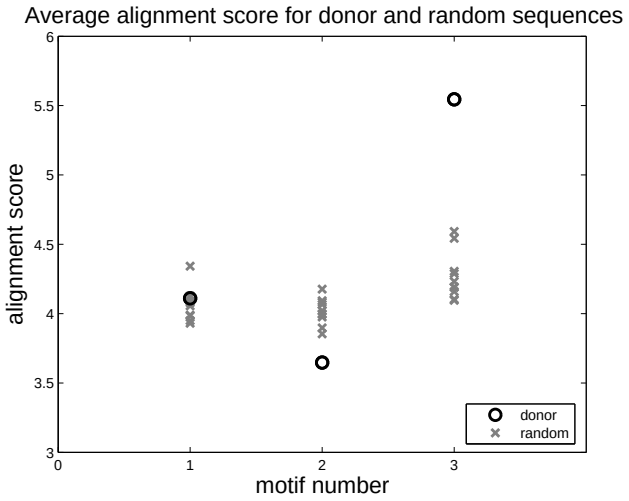



Figure 6.23: Average alignment scores between motif covariance model and the instances of the motifs in real extended donor sequences from the STRIN dataset and 10 datasets of shuffled donor sequences. Data points correspond to the average score values for each dataset. The motif numbers as they have been identified by CMfinder are shown on the x-axis. They correspond to the motif names in Table 6.3.

Table 6.3. These scores along with the scores for random sequence datasets are plotted in Figures 6.23, 6.24 and 6.25.

If the identified structural motifs are conserved in the vicinity of real splice signals, we would expect that their scores would be higher than if they are found by chance, which is the case in random sequences. The figures for the average alignment scores show that this is the case for some of the identified motifs: motif 3 found in extended donor sequences, all of the motifs found in extended acceptor sequences and motifs 1 and 3 found in extended branchpoint sequences have a higher average score for real sequences than for random ones.

These results suggest that instances of motif 3 found in extended donor sequences, all three motifs found in acceptor sequences and motifs 1 and 3 found in extended branchpoint sequences have a degree of conservation that is not expected for random occurrences of the motifs. In order to analyze if

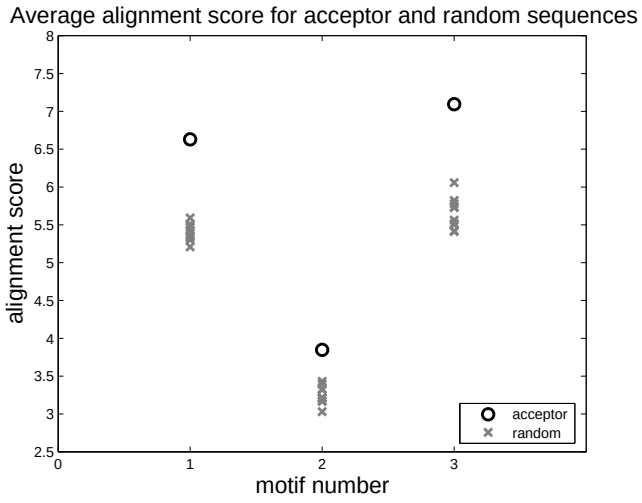



Figure 6.24: Average alignment scores between motif covariance model and the instances of the motifs in real extended acceptor sequences from the STRIN dataset and 10 datasets of shuffled acceptor sequences.

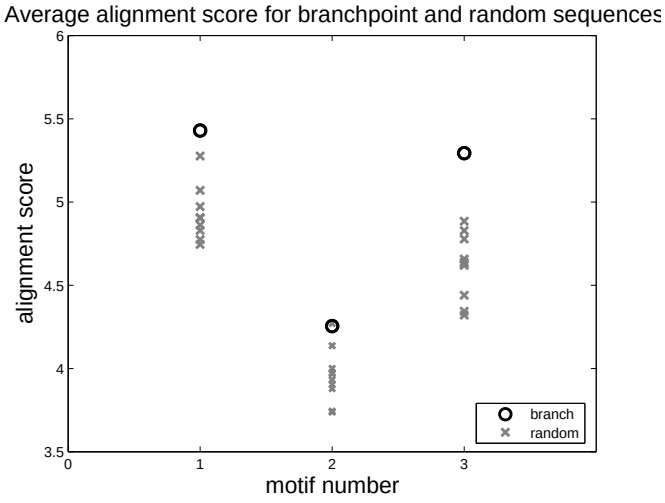



Figure 6.25: Average alignment scores between motif covariance model and the instances of the motifs in real extended branchpoint sequences from the STRIN dataset and 10 datasets of shuffled branchpoint sequences.

these motifs have biological significance, further tests need to be carried out. One approach would be to search for the occurrences of these motifs in the splice site neighbourhoods of a related yeast species. Ideally, the selected species should not have high sequence similarity in these regions, since this could bias the results. Finding the same structural motifs in spite of low sequence conservation would provide strong evidence that these motifs are associated with splice sites and that their structure is probably functionally important. Another way to test the hypothesis that the found structural motifs are functional and important for splicing would be to perform a series of mutational analyses in the laboratory and to seek for protein factors that bind to these motifs.

## 6.5 Conclusions

In this chapter we conducted a series of analyses on the STRIN dataset, with the goal of identifying any specific structural characteristics that might be important for pre-mRNA splicing. Our comparison of the global secondary structure stability of yeast introns versus random sequences hinted at weak biases in favor of native sequences. This is mainly due to several 5'L introns, which have minimum free energies significantly lower than random sequences with the same dinucleotide composition.

We found that donor sites exhibit a statistically significant bias against stable secondary structures when folded locally, which is in agreement with the molecular biology of the splicing reaction: a donor site is recognized multiple times during the splicing process and thus having the region around the donor site relatively free of secondary structure will allow these interactions to happen more easily. Similarly, we found that the branchpoint sequences tend to be unbound in global intron folds, which would allow for easier recognition by the U2 snRNA.

We also studied the stability of binding interactions between splice signals and small nuclear RNAs that bind to them. These interactions are the main reason for sequence conservation of the splice sites, but it is possible that the thermodynamic aspect of interactions plays an important role in

splice site identification. We found that both donor sites and branchpoint sequences have lower minimum free energy of folding with snRNAs than do pseudo sequences.

Finally, we identified short structural motifs in the vicinity of donor, acceptor and branchpoint sequences that have a degree of conservation among STRIN introns that is not expected for random occurrences of the motifs.

# Chapter 7

# Using structural information for intron identification

In Chapter 5, we proposed a model of the role of pre-mRNA secondary structure in the splicing of long (5'L) introns in *S. cerevisiae*. The existence of highly probable secondary structures (whose free energy is within 5% from the minimum free energy) that have short branchpoint distance (calculated by our implementation of Dijkstra's algorithm, Section 5.3.2) appears to be required for efficient splicing of a yeast intron. We tested our splicing model experimentally, and the results were mostly consistent with our splicing efficiency predictions based on the model.

In Chapter 6, we analyzed the stability of global intron folding, the stability of local structures in the vicinity of splice signals, the structural accessibility of splice sites in global intron folds, the stability of snRNA-splice-signal binding interactions and the existence of conserved structural motifs in the vicinity of splice signals. In most cases, we showed that splicing signals have slightly different structural characteristics than do random sequences. None of these signals, however, seems strong enough to discriminate between real and pseudo sites.

In this chapter, we test if the structural characteristics of long introns can be used to computationally predict splicing efficiency. Using machine learning techniques, we test the predictive power of shortened branchpoint distances and secondary structure probabilities obtained by the procedure StructureAnalyze (Section 5.3.3), and of summary statistics obtained by post-processing of these metrics (same section). Finally, we investigate whether the weak structural context signals identified in Chapter 6 can be

used to improve the accuracy of splice site and intron recognition in yeast.

## 7.1 Machine learning and classification

In machine learning, the task of supervised learning is to create a rule or a function from training data. The training data consist of pairs of input objects, typically vectors, and output values. If $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is an input vector, its features or attributes, $x_i$, can be real-valued numbers, discrete-valued numbers, or categorical values. The output value, $y$, can be categorical, in which case the process of learning from training examples is called classification, or it can be a real value, in which case the process is called regression. An important special case of categorical output is Boolean output, where training examples that have value 1 are called positive instances and training examples with value 0 are called negative instances. Once the supervised learner is created, its task is to predict a value of the function for any valid input object.

Formally, the classification problem can be stated as follows: Let $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ be a set of input vectors and $Y = \{y_1, \ldots, y_n\}$ the corresponding set of output values. Given training data $\{(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)\}$ produce a classifier $h : \mathbf{X} \rightarrow Y$ that maps an object $\mathbf{x_i} \in \mathbf{X}$ to its classification label $y_i \in Y$.

There are a number of methods for supervised learning, but for our purposes we will focus on neural networks and support vector machines. These algorithms provide state-of-the-art performance in a variety of application domains and have been widely used in Bioinformatics for various pattern recognition and classification problems.

### 7.1.1 Neural networks

An artificial neural network, also commonly called a neural network, is an information-processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. It is composed of a number of computational units, connected together such that the output of

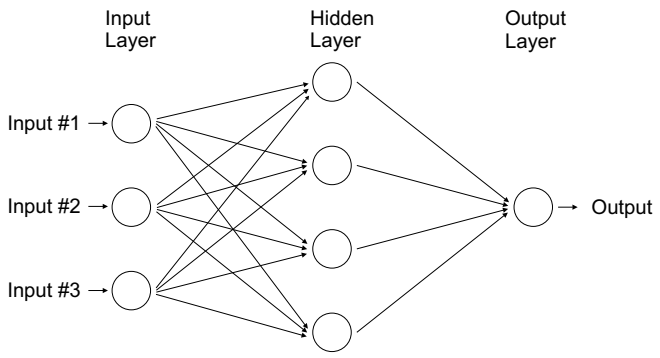



Figure 7.1: An example of a feed-forward three-layer neural network architecture.

a unit is a function of its inputs and some adjustable parameters (see, e.g., Poole et al. (1998)). Learning is accomplished by adjusting the parameters to fit the training data.

### Network layers

The most common type of neural network consists of three groups, or layers, of units: input units, hidden units and output units. An example of such a network architecture is shown in Figure 7.1.

The input units receive the raw information (feature vectors $\mathbf{x_i}$) that is fed into the network. The activity of each hidden unit is determined by the activities of the input units and the weights on the connections between the input and the hidden units. The behaviour of the output units depends on the activity of the hidden units and the weights between the hidden and output units. This type of neural network, where the information flows in one direction only (no cycles), is called a feed-forward neural network, and is the most commonly used.

### Activation functions

The behaviour of a neural network depends on both the weights and the input-output function, also called the transfer or activation function, that is specified for the units. This function typically falls into one of three categories: linear, threshold or sigmoid. For linear units, the output activity is proportional to the total weighted input. For threshold units, the output is set at one of two levels, depending on whether the total weighted input is greater than or less than some threshold value. For sigmoid units, the output varies continuously but not linearly as the input changes. The sigmoid function, a special case of the logistic function, is applied to the weighted input to obtain the output value:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{7.1}$$

### Computing the outputs

The analytic function corresponding to a three-layer, feed-forward neural network can be derived as follows (Bishop, 1996). Let us assume that there are $n$ input units, $l$ hidden units and $m$ output units. The output of the $j$-th hidden unit is obtained by first computing a weighted sum of the $n$ input values and adding a bias:

$$a_j = \sum_{i=1}^{n} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \tag{7.2}$$

Here $w_{ji}^{(1)}$ denotes a weight from the first layer of weights, going from input unit $i$ to hidden unit $j$, and $w_{j0}^{(1)}$ is the bias for hidden unit $j$. The activation of hidden unit $j$ is then obtained by transforming the sum in Equation 7.2 using an activation function $g$:

$$z_j = g(a_j) \tag{7.3}$$

Next, the outputs of the network are obtained by using the weighted sum of inputs $z_j$, where $j = 1, \ldots, l$ and $z_j$ are outputs of hidden units. For

each output unit $k$, we first compute a weighted sum of the $l$ hidden unit values and add a bias:

$$a_k = \sum_{i=1}^{l} w_{kj}^{(2)} z_j + w_{k0}^{(2)} \tag{7.4}$$

The activation of the $k$-th output unit is obtained by transforming the linear sum in Equation 7.4 using an activation function:

$$y_k = \tilde{g}(a_k) \tag{7.5}$$

As can be seen from the above formulae, the activation functions for the hidden and output layers do not have to be the same.

We can absorb the bias terms into Equations 7.2 and 7.4 by including an extra input feature $x_0$ whose value is permanently clamped to $x_0 = 1$. If we combine Equations 7.2, 7.3, 7.4, and 7.5, we get an explicit expression of the function representing the described neural network:

$$y_k = \tilde{g}\left(\sum_{i=0}^{l} w_{kj}^{(2)} g\left(\sum_{i=0}^{n} w_{ji}^{(1)} x_i\right)\right) \tag{7.6}$$

**Network training and testing**

The aim of neural network training is, given a set of training examples, to find parameter settings that minimize the error between the desired output and the actual output. If the network has $M$ parameters in total, finding the optimal parameter setting involves searching through an $M$-dimensional Euclidean parameter space. One method that is commonly used for this purpose is back-propagation, which is a gradient descent search through the parameter space to minimize the error function (usually sum-of-squares error function).

The evaluation of a neural network's performance can be done using a test set that does not overlap with the dataset used for the training of the network. Another approach is to perform both training and testing of the network on one dataset using the cross-validation technique. In cross-

validation, a portion of the data is set aside as training data leaving the remainder as testing data. There are several types of cross-validation, the most common being the $K$-fold cross-validation, where the original dataset is partitioned into $K$ subsets. $K-1$ of these subsets are used together for training, and the last subset is used for testing. This process is repeated $K$ times, and the average error across all $K$ trials is computed.

### 7.1.2 Support vector machines

Support vector machines (SVMs) are a set of supervised learning methods used for classification and regression. Their common characteristic is that they use a technique known as 'kernel trick' to apply linear classification techniques to non-linear classification problems.

Suppose we are given $l$ training examples, each one consisting of an input vector $\mathbf{x_i} \in R^n$ and the output $y_i \in \{-1, +1\}$. The learning task is to estimate a decision function $f : \mathbf{X} \rightarrow \{-1, +1\}$ that predicts the label of any $\mathbf{x} \in R^n$, i.e. the function separates the input space $\mathbf{X}$ in two classes, $-1$ and $+1$. Support vector machines construct such a decision function in the form of a linear separating hyperplane:

$$f(\mathbf{x}) = sign(\mathbf{w} \cdot \mathbf{x} + b) \tag{7.7}$$

The goal of SVM training is to find the optimal hyperplane, defined by the weights $\mathbf{w} \in R^n$ and the bias $b \in R$ ($\mathbf{w} \cdot \mathbf{x} + b = 0$), such that the margin of separation between the two classes is maximized. The hyperplane with this property can be uniquely constructed solving a constrained quadratic optimization problem whose solution has the following expansion (Cristianini and Shawe-Taylor, 2000):

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i \tag{7.8}$$

The training examples that lie on the margin are called 'support vectors', and they carry all the relevant information about the classification problem. The decision function can be rewritten in dual coordinates using Equation

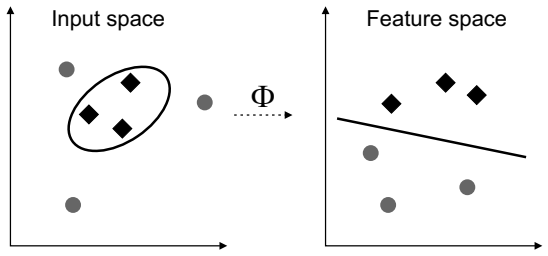



Figure 7.2: Non-linear mapping $\Phi$ of the input space into the feature space: training data that was not linearly separable in the input space becomes so in the feature space.

7.8:

$$f(\mathbf{x}) = sign\left(\sum_{i=1}^{l} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + b\right) \tag{7.9}$$

The described approach constructs a separating hyperplane with maximum margin in the input space. However, this linear classification might not be sufficient for more complex problems where linear separation is not possible. This is where the main idea of SVMs comes into play: to transform the training data non-linearly into a feature space using the mapping $\Phi : R^n \rightarrow F$, and to construct a separating hyperplane with maximum margin in $F$ (see Figure 7.2). This yields a non-linear decision boundary in the input space.

After mapping the input vectors into a feature space and using the dual representation of the decision function (Equation 7.9), we obtain (Cristianini and Shawe-Taylor, 2000):

$$f(\mathbf{x}) = sign\left(\sum_{i=1}^{l} \alpha_i y_i \left(\mathbf{\Phi}(\mathbf{x}_i) \cdot \mathbf{\Phi}(\mathbf{x})\right) + b\right) \tag{7.10}$$

This means that the decision function can be expressed as a linear combination of the training points, so that the decision rule can be evaluated

using the dot product between the test point and the training points. By using a function $K$ (*kernel* function), such that for all $\mathbf{x}, \mathbf{z} \in R^n$

$$K(\mathbf{x}, \mathbf{z}) = (\mathbf{\Phi}(\mathbf{x}) \cdot \mathbf{\Phi}(\mathbf{z})) \tag{7.11}$$

$\mathbf{\Phi}(\mathbf{x})$ is never explicitly computed; results are computed directly in the input space. The decision function becomes:

$$f(\mathbf{x}) = sign\left(\sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \tag{7.12}$$

This is the essence of the previously mentioned kernel trick. What remains is to find a kernel function that can be evaluated efficiently. Some examples of commonly used kernel functions are:

1. The linear kernel

$$K(\mathbf{x}, \mathbf{z}) = (\mathbf{x} \cdot \mathbf{z}) = \sum_{i=0}^{n-1} x_i z_i \tag{7.13}$$

2. The polynomial kernel

$$K(\mathbf{x}, \mathbf{z}) = (\mathbf{x} \cdot \mathbf{z})^d = \left(\sum_{i=0}^{n-1} x_i z_i\right)^d \tag{7.14}$$

3. The Gaussian radial basis function (RBF) kernel

$$K(\mathbf{x}, \mathbf{z}) = \mathbf{e}^{-\left(\frac{\|\mathbf{x}-\mathbf{z}\|^2}{\sigma^2}\right)} \tag{7.15}$$

## 7.2 Learning the splicing efficiency function

For all of our machine learning experiments, we used the software package WEKA, which is a comprehensive toolbench for machine learning and data mining (Witten and Frank, 2005). It contains Java implementations of many machine learning algorithms and data transformation tools for data

preprocessing. We downloaded the WEKA software from `http://www.cs.waikato.ac.nz/ml/weka/` (September 2005).

Data used for either training or testing of machine learning algorithms in WEKA needs to be in a specific ARFF format, where attributes and output value are defined as real-valued, discrete-valued or categorical. Once the training data is uploaded into the system, the user can choose which machine learning classifier to train. For each classifier there are a number of options and parameters that can be modified. The user can also choose what data to use to test the classifier once the training phase is complete. There is an option of using a separate test set or performing cross-validation experiments.

In the following section, we describe machine learning experiments that are based on the findings in Chapter 5, where we analyzed how the splicing efficiency of introns is related to their ability to shorten the branchpoint distance by folding into favorable secondary structure.

### 7.2.1 Training on shortened branchpoint distance and structure probability data

For the first phase of our computational experiments, we tested to see if machine learners can learn the 'splicing efficiency function' based on raw data only – shortened branchpoint distances and probabilities for each predicted secondary structure within a certain percentage from the minimum free energy. These values were computed using the procedure StructureAnalyze, described in Section 5.3.3.

For training data, we chose the wild type intron and 15 intron mutants of the gene RP51B, described by Libri et al. (1995) and by Charpentier and Rosbash (1996) and used for our analysis in Chapter 5: 3mUB1, 5mUB1, 8mUB1, 3mDB1, 5mDB1, 3mUB1/3mDB1, 5mUB1/5mDB1, 6mUB1, UB1i, DB1i, UB1iDB1i, mut5, mut12, and mut18. Mutant 4mUB1 was excluded as a borderline case based on the copper resistance results shown in Figure 5.2. The splicing efficiency of some of the mutants from the work of Libri et al. (1995) was also confirmed by our own laboratory experiments (see

Section 5.6). The dataset for testing the performance of the machine learners trained on the previously described training dataset contained the newly designed RP51B intron mutants, for which we experimentally tested splicing efficiency levels (Section 5.6.2). We only considered the mutants that were successfully inserted into the genome and expressed at the detectable levels, with the exception of mutant bad4, which was a borderline case with respect to its protein expression level and mutant good5, for which splicing efficiency was decreased, probably due to a very stable stem (our model does not capture this kind of structural information).

We used the procedure StructureAnalyze to compute shortened branchpoint distances and probabilities of all suboptimal structures within a certain percentage from the MFE. We ran the procedure for 5, 10 and 20 percent of suboptimality (parameter $P$ in mfold). Since the feature vectors of all data instances in training and test sets have to be of the same length, we also had to specify the number of suboptimal structures computed by mfold (parameter $M$). In some cases, some of the mutants had less than the required number of suboptimal structure predictions and were excluded from the respective dataset.

A dataset of instances whose feature vectors are computed based on the five most probable suboptimal structures within 5% from MFE ($P = 5$ and $M = 5$) would have 10 attributes, $\bar{d}_1, p_1, \bar{d}_2, p_2, \ldots, \bar{d}_5, p_5$, where $\bar{d}_i$ are the shortened branchpoint distances for the five computed suboptimal structures and $p_i$ are the relative probabilities of those structures (the relative probabilities of the structures are computed based on all suboptimal structures within 5% from the MFE – see Equation 5.2 on p. 111). The output value is defined as a Boolean value, where 1 represents efficient splicing and 0 represents decreased or inhibited splicing. The output values are extracted from Tables 5.1 and 5.4, assigning value 1 to all the mutants that are spliced at the normal, slightly reduced, slightly improved or improved levels, and value 0 to all the mutants with reduced or inhibited splicing.

We also considered the mutant sequences with their 50-nt flanking regions, even though in the analysis in Chapter 5 they did not have as good results as did the intron-only sequences. For these extended sequences, we

| dataset | P=5 M=5 | | P=10 M=10 | | P=10 M=5 | |
|---|---|---|---|---|---|---|
| | NN | SVM | NN | SVM | NN | SVM |
| **intron only** | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| **intron +/- 50 nt** | 0.8 | 1.0 | 0.8 | 1.0 | 1.0 | 1.0 |

Table 7.1: Fraction of correctly classified instances for various mfold parameter settings and for two types of sequences: intron-only and introns with 50-nt flanking regions.

computed the branchpoint distances and structure probabilities for the previously mentioned values of the $P$ and $M$ parameters.

Finally, for each pair of parameter values $P = 5$, $M = 5$ and $P = 10$, $M = 10$ and $P = 10$, $M = 5$, and for a training set computed on intron-only sequences and intron sequences with 50-nt flanking regions, we trained a neural network (NN) and a support vector machine (SVM) algorithm using the WEKA software package. We then tested their performance on the appropriate version of the test set. In WEKA, the neural network classifier is called a 'multi-layer perceptron' (MultilayerPerceptron) and the SVM classifier is called 'SMO', based on the implementation of the sequential minimal optimization algorithm used for SVM training. For NN, the following WEKA default parameter settings were used: one hidden layer with (attributes+classes)/2 hidden units, learning rate $L = 0.3$, momentum $M = 0.2$, and number of training epochs $N = 500$. For SVM classification we used a linear kernel. The results are shown in Table 7.1.

When intron-only sequences are considered, both NN and SVM correctly predict the splicing efficiency level of all the mutants in the test dataset, regardless of the $P$ and $M$ parameter values used. For the datasets derived from the extended intron sequences, the accuracy results are slightly weaker. We also performed the same experiment with $P = 20$ and $M = 20$; however, the results are significantly weaker (data not shown).

It is important to note that the results of the StructureAnalyze procedure for $P = 5$, $M = 5$ and $P = 10$, $M = 5$ can be similar but are not necessarily identical because the computed relative probabilities can be different, and some of the mutants that did not have 5 structures predicted for $P = 5$ did

| dataset | P=5 M=5 | | P=10 M=10 | | P=10 M=5 | |
|---|---|---|---|---|---|---|
| | NN | SVM | NN | SVM | NN | SVM |
| **intron only** | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| **intron +/- 50 nt** | 0.8 | 1.0 | 0.8 | 1.0 | 0.8 | 1.0 |

Table 7.2: Fraction of correctly classified instances for various mfold parameter settings and for two types of sequences: intron-only and introns with 50-nt flanking regions. The branchpoint distance attributes, $\bar{d}_i$, are sorted in ascending order.

for $P = 10$.

The order of the attribute values is not random, but it is based on mfold output, which outputs suboptimal structures in descending order with respect to their predicted free energies. These free energies are proportional to the relative probabilities of the structures, therefore the values of attributes $p_i$ in our feature vectors are in descending order from $i = 1, \ldots, 5$. The values of the attributes $\bar{d}_i$ are not in any particular order, but are always paired with the relative probability of the structure from which they are derived. Considering this, the first attribute in a feature vector, $\bar{d}_1$, will always be the shortened branchpoint distance of the minimum free energy (optimal) structure, the second attribute $p_1$ will be its probability, and so on. This order will put an emphasis on the probabilities of the structures, which we believe are not very reliable and are not essential for our splicing model. If there is a structure that has a short branchpoint distance, this value can be any of the $\bar{d}_i$ attributes, depending on the probability of the structure, and thus the importance of its presence cannot be captured by a classifier. We can change this if we sort $\bar{d}_i$ attributes in either ascending or descending order. This ordering of attributes is better suited to our splicing model. The results for this approach are shown in Table 7.2 and are almost identical to the results in Table 7.1, indicating that for the test set that we are using the ordering does not make much difference. According to the results from Tables 7.1 and 7.2, it seems that the NN and SVM classifiers are able to learn the 'splicing efficiency' function based only on structural branchpoint distances and the probability of structures.

| **sorting** | **P=5 M=5** | | **P=10 M=10** | | **P=10 M=5** | |
|---|---|---|---|---|---|---|
| | **NN** | **SVM** | **NN** | **SVM** | **NN** | **SVM** |
| $p_1 > p_2 > \ldots > p_5$ | 0.55 | 0.57 | 0.48 | 0.47 | 0.65 | 0.59 |
| $\bar{d}_1 > \bar{d}_2 > \ldots > \bar{d}_5$ | 0.56 | 0.58 | 0.43 | 0.48 | **0.71** | 0.68 |

Table 7.3: Fraction of correctly classified instances, which in this case is equal to the true positive rate (sensitivity), for various mfold parameters and for two different sortings of attributes.

We also tested the performance of the trained classifiers on more challenging test datasets, where feature vectors were computed using STRIN 5'L introns (98 sequences). However, it is necessary to point out that we do not have the splicing efficiency data for these introns, which is what the classifiers are trying to predict. We cannot assume that all yeast introns are spliced out with the same efficiency. For some genes, low splicing efficiency might be desired, such as in the case of the *S. cerevisiae* gene Yra1p, whose increased splicing efficiency results in reduced cell growth (Preker and Guthrie, 2006). Hence, it is very likely that if real splicing efficiency results were available for the STRIN 5'L introns there would be both positive and negative instances with respect to splicing efficiency. However, we would still expect to see more positive instances than negative ones. Therefore, for the purposes of this test, we assume that all STRIN 5'L introns are efficiently spliced (output value = 1). The results for $p_1 > p_2 > \ldots > p_5$ and for $\bar{d}_1 > \bar{d}_2 > \ldots > \bar{d}_5$ cases are shown in Table 7.3.

The prediction accuracy values in this table are much lower than for the initial test set. It seems that the $\bar{d}_i$ ordering yields better results since the highest percentage of correctly classified instances is 0.71 (for NN prediction and $P = 10$, $M = 5$), corresponding to 70 true positives and 28 false negatives, compared to 0.65 (for NN prediction and $P = 10$, $M = 5$) for the $p_i$ ordering.

Upon closer inspection of these NN predictions, we observed that 60 of the 68 instances that have a minimum shortened branchpoint distance of less than 25 were predicted correctly with label 1 (efficient splicing). These results are in agreement with our splicing model and the training data that

we have available (none of the efficiently spliced mutants in the training set has a minimum shortened branchpoint distance greater than 20).

### 7.2.2 Training on structural summary statistics

In this section, we use the structural summary statistics described in Section 5.3.3 to distinguish between efficiently and poorly spliced introns. The training and test datasets are the same as described in the previous section. We used the procedure StructureAnalyze to compute certain structural properties (see Appendix C) for each predicted suboptimal structure that is within a certain percentage from the MFE. These values are then used to compute summary statistics for each sequence in the given dataset. The following values are used as attributes for machine learning classification:

- the minimum shortened branchpoint distance among all suboptimal predictions within 5% from the MFE (**MIN**)
- the average shortened branchpoint distances for all suboptimal predictions within 5% from the MFE (**AVG**)
- the **r_weight** summary statistics calculated as defined in Equation 5.5 (p. 112) (**R_WEIGHT**)
- secondary structure of the branchpoint sequence (is it in a stem or a loop?) (**BP_STRUCT**={stem, loop})
- probability of small distance: sum of relative probabilities of all suboptimal structures that have branchpoint distance shorter than some threshold (**PROB**) (we tried threshold values of 10, 20, 25 and 30)

We also calculated the base-pairing probability of the contact conformation from the dotplot matrix (**BP_PROB**) (see Section 5.4). As in the previous section, we used the WEKA software package to train and evaluate neural network and support vector machine classifiers using various combinations of previously described attributes (see Table 7.4) and two different percent suboptimality values ($P = 5$ and $P = 10$). The results are shown

| % subopt | Attributes | NN | SVM |
|---|---|---|---|
| **5 %** | MIN, AVG, BP_STRUCT, BP_PROB | 1.0 | 1.0 |
| | MIN, AVG, R_WEIGHT, BP_STRUCT, BP_PROB | 1.0 | 1.0 |
| | MIN, AVG, BP_STRUCT, PROB (th=10,20,25) | 1.0 | 1.0 |
| | MIN, AVG, BP_STRUCT, PROB (th=30) | 1.0 | 1.0 |
| **10 %** | MIN, AVG, BP_STRUCT, BP_PROB | 0.8 | 0.8 |
| | MIN, AVG, R_WEIGHT, BP_STRUCT, BP_PROB | 1.0 | 0.8 |
| | MIN, AVG, BP_STRUCT, PROB (th=10,20,25,30) | 1.0 | 1.0 |

Table 7.4: The accuracy results for machine learning classification of 5 RP51B mutants (bad1, bad3, good2, good3, good4). The attributes of feature vectors for both the training and test sets are listed in the column **Attributes**; the percent of suboptimality used to calculate the attribute values is given in the column **% subopt**. The accuracy of the prediction, given separately for **NN** and **SVM**, is the fraction of correctly classified instances (both positive and negative) achieved by the respective classifier.

in Table 7.4. As in the previous section, the accuracy results are very good: both NN and SVM correctly predicted the splicing efficiency level of all of the mutants in the test dataset when $P = 5$. For $P = 10$, the results are slightly weaker.

The accuracy results of the machine learning classifiers trained on the same training set and tested on the STRIN 5'L introns are given in Table 7.5. It is apparent that the classification accuracy for the STRIN dataset is much poorer than for the five designed mutants. The best prediction accuracy, 0.64, is achieved by the SVM classifier for $P = 10$ using attributes MIN, AVG, BP_STRUCT and PROB (th=20). This value is also the true positive rate (see Equation 6.2, p. 176) corresponding to 63 true positive predictions and 35 false negative predictions.

In summary, the machine learning approach yielded very good results when applied on the datasets that have experimentally determined splicing efficiency levels, indicating that the classifiers were able to learn splicing efficiency as a function of the provided attributes. On the other hand, the classification accuracy results on the STRIN 5'L introns were much lower, usually around 50%, with some exceptions. As explained in the previous section, this is not an entirely adequate test, since we do not have experi-

| % subopt | Attributes | NN | SVM |
|---|---|---|---|
| **5 %** | MIN, AVG, BP_STRUCT, BP_PROB | 0.51 | 0.32 |
| | MIN, AVG, R_WEIGHT, BP_STRUCT, BP_PROB | 0.51 | 0.31 |
| | MIN, AVG, BP_STRUCT, PROB (th=10) | 0.53 | 0.56 |
| | MIN, AVG, BP_STRUCT, PROB (th=20) | 0.47 | 0.56 |
| | MIN, AVG, BP_STRUCT, PROB (th=25) | 0.46 | 0.55 |
| | MIN, AVG, BP_STRUCT, PROB (th=30) | 0.52 | 0.56 |
| **10 %** | MIN, AVG, BP_STRUCT, BP_PROB | 0.12 | 0.11 |
| | MIN, AVG, R_WEIGHT, BP_STRUCT, BP_PROB | 0.54 | 0.11 |
| | MIN, AVG, BP_STRUCT, PROB (th=10) | 0.50 | 0.24 |
| | MIN, AVG, BP_STRUCT, PROB (th=20) | 0.62 | **0.64** |
| | MIN, AVG, BP_STRUCT, PROB (th=25) | 0.56 | 0.56 |
| | MIN, AVG, BP_STRUCT, PROB (th=30) | 0.52 | 0.60 |

Table 7.5: The accuracy results for machine learning classification of STRIN 5'L introns. The attributes of feature vectors for the both training and testing sets are listed in the column **Attributes**; the percent of suboptimality used to calculate the attribute values is given in the column **% subopt**. The accuracy of the prediction, given separately for **NN** and **SVM**, is the fraction of correctly classified instances (in this case also the true positive rate) achieved by the respective classifier.

mentally measured splicing efficiency levels, but assume that all yeast long introns are efficiently spliced, which is probably not the case.

## 7.3 Using weak structural signals to improve accuracy of computational splice site and intron prediction

To explore the capability of secondary structure information to aid splice site or intron prediction, we used the weak structural signals identified in Chapter 6 to filter out false positive predictions given by a splice-site prediction tool.

Since we based our study on yeast sequences, we attempted to find a prediction tool that is trained on yeast sequences. Unfortunately, because *S. cerevisiae* was sequenced in 1996 and has been thoroughly annotated since then, it seems that annotation tools are no longer needed, and there are very few that can be found. The only gene-finding program that we found

that is intended to work on yeast sequences is GeneMark.hmm (Lukashin and Borodovsky, 1998). The Web-based version of the program for yeast, or low eukaryotes, can be found at `http://opal.biology.gatech.edu/GeneMark/gmhmm2_loweuk.cgi` (last accessed in May 2006). However, after some testing and correspondence with the authors, we learned that this version is is a modification of the prokaryotic version of GeneMark.hmm and that it does not model exon/intron structure.

The only other prediction tool that we found is SPL, which is software for predicting splice sites in the following organisms: *H. sapience*, *D. melanogaster*, *C. elegans*, *A. thaliana* and *S. cerevisiae*. The tool is a commercial product distributed by the SoftBerry company (`http://softberry.com`), but they also offer limited on-line access (last accessed in July 2006). The SPL algorithm is based on linear discriminant analysis (see Section 2.1).

The Web-based tool takes a genomic sequence as an input and outputs a list of predicted donor and acceptor splice sites. The donor and acceptor splice sites are not coupled to indicate intron locations. A short description on the SoftBerry Web site offers some of the program's prediction accuracy statistics: the accuracy of donor site recognition is 97% and for acceptor splice sites it is 96% (in humans). The false positive rate is high – about one false positive per true site, for 97% accuracy of true sites prediction. It is important to note that the terminology used on this Web site is different from the one used here and in most other studies: the accuracy they are referring to is sensitivity of the prediction (i.e., the true positive rate, as given in Equation 6.2, p. 176). Also, the meaning of false positive rate on the SoftBerry Web site is different from the one in Equation 6.3. Their estimate of false positive rate would correspond to $1 - PPV$, where $PPV$ is positive predictive value, defined as:

$$PPV = \frac{TP}{TP + FP} \tag{7.16}$$

Thus, they claim that the $PPV$ of SPL's splice site prediction is 0.5 (one true positive per one true negative prediction).

For this analysis, we used 98 yeast genes that contain STRIN long introns

(not including 5' UTR introns), 104 genes that contain STRIN short introns and 100 genes that do not contain any introns (302 genes in total). For each gene, we also included 200-nt flanking regions on both sides. This is needed because some of the sequence content information measures used in the linear discriminant function are evaluated on a larger sequence window around a candidate splice site (up to $+/-$ 100 nt) and many yeast introns are located just downstream from the gene start site.

We carefully examined all of the genes in the dataset for any overlap with other genes in order to avoid any unaccounted splice sites, on a chromosome-by-chromosome basis. We selected the intron-less genes based on the following criteria:

- A gene does not overlap with any other gene and does not contain 5' UTR introns.
- A gene has to be 'verified' according to the Saccharomyces Genome Database, which indicates that experimental evidence exists that a gene product is produced in *S. cerevisiae*.

From 3181 candidate genes that satisfy these criteria, we randomly selected 100. We then stitched these 302 extended gene sequences together into longer sequences whose length is less than 10000 nt (assumed length limit for SPL Web tool) and submitted them to the SPL Web server, one by one, for prediction of splice sites. Outputs were gathered from the Web site and formated for further analysis.

There were 383 donor sites and 519 acceptor sites predicted for the gene sequences containing long introns (95 of these are real donor sites and 79 are real acceptor sites), 407 donor and 507 acceptor sites predicted for the gene sequences containing short introns (103 of these are real donor sites and 56 are real acceptor sites), and 422 donor and 738 acceptor sites predicted for the intron-less gene sequences. Altogether, 1212 donor splice sites (198 real ones and 1014 pseudo sites) and 1764 acceptor splice sites (135 real ones and 1629 pseudo sites) were predicted for our dataset. According to these results, the positive predictive values are 0.16 and 0.08 for donor and acceptor sites,

respectively. This indicates that the rate of false positive prediction (or $1 - PPV$) indicated on the SoftBerry Web site is grossly underestimated, at least for our dataset. The sensitivity of donor site prediction is 0.98, which corresponds well to SoftBerry's claim, however, the sensitivity of acceptor site prediction, 0.67, is much lower than claimed.

### 7.3.1 Improving the accuracy of splice site prediction

Trying to improve upon these results, we first explored if the weak structural signals discussed in Chapter 6 can be used to train machine learning classifiers to distinguish between real and pseudo splice sites. We used structural measures that were shown to be significantly different between real splice sites from the STRIN dataset and pseudo splice sites found in randomly generated sequences. For donor sites, these measures are:

- **(ATTD1)** MFE of folding of the 50-nt window located from position $-40$ to $+10$ w.r.t the candidate donor site (see Section 6.1.2)
- **(ATTD2)** MFE of folding of 50-nt window located from position $-20$ to $+30$ w.r.t the candidate donor site (same section)
- **(ATTD3)** MFE of folding between the U1 snRNA and 11-nt candidate donor sequence (3 nt from the 5' exon and first 8 intronic nt – NNN|GUAUGUNN) (see Section 6.3.2 and Figure 6.18)
- **(ATTD4)** MFE of folding between the U6 snRNA and 12-nt candidate donor sequence (4 nt from the 5' exon and first 8 intronic nt – NNNN|GUAUGUNN) (see Section 6.3.2 and Figure 6.20)
- **(ATTD5)** the highest instance score of structural motif 3 found in a 200-nt window centered at the donor splice site (see Section 6.4.2 and Figure 6.23)

The dataset of donor sequences contained 1212 sequences each of length 200-nt and centered at the donor sequences predicted by SPL. For each of these sequences, we calculated the above five measures and used them as

attribute values to assemble feature vectors for machine learning purposes. Real donor sites had a class label 1 (positive instances) and pseudo donors had a class label 0 (negative instances). We used this dataset to both train and test the neural network and SVM classifiers, using 10-fold cross validation experiments, as follows. The data was randomly partitioned in 10 equal subsets and a classifier was trained on 9/10 of the data and tested on 1/10 of the data. This was repeated 10 times, each time using a different subset for testing, and the final accuracy results were obtained by averaging the results of the 10 cycles.

We trained the classifiers on several different combinations of attributes. The attributes were grouped in three groups – the first of these relating to the MFE of folding of the 50-nt window (ATTD1 and ATTD2), the second group relating to the stability of interaction between snRNAs and donor sequences (ATTD3 and ATTD4) and the third group containing only ATTD5. We then combined these three groups of attributes in all possible ways.

The accuracy results of machine learning experiments on the donor dataset are given in Table 7.6. The SVM classifier was not able to learn a function that would differentiate between positive and negative instances for any combination of attributes and predicted all instances as negative. Therefore, only the accuracy results for neural network classification are given. The percentage of correctly classified instances (**Accuracy**) is fairly high for all the experiments, but this is due to the fact that a majority of the candidate donor sites are predicted as negative instances. Although this is what we wanted to accomplish for pseudo donor sites, most real donor sites are also predicted to be negative instances. There are two exceptions: using all five attributes ($TP = 55$), and using all the attributes except ATTD5. The latter scenario has the highest values for accuracy (0.85) and sensitivity (0.30) and the highest number of true positive predictions (60). Although the sensitivity is significantly lower than SPL's sensitivity, $PPV$ has improved considerably (0.55) compared to the original SPL prediction on our dataset.

| Attributes | Neural Network | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Accuracy** | **Sn** | **PPV** | **TN** | **FP** | **TP** | **FN** |
| ATTD1, ATTD2 | 0.84 | 0 | — | 1014 | 0 | 0 | 198 |
| ATTD3, ATTD4 | 0.83 | 0 | 0 | 1011 | 3 | 0 | 198 |
| ATTD1, ATTD2, ATTD3, ATTD4 | **0.85** | **0.30** | 0.55 | 965 | 49 | **60** | 138 |
| ATTD1, ATTD2, ATTD5 | 0.84 | 0.04 | 0.54 | 1008 | 6 | 7 | 191 |
| ATTD3, ATTD4, ATTD5 | 0.84 | 0.01 | 1.00 | 1014 | 0 | 1 | 197 |
| ATTD1, ATTD2, ATTD3, ATTD4, ATTD5 | 0.84 | 0.28 | 0.53 | 966 | 48 | 55 | 143 |

Table 7.6: The accuracy results for neural network classification of 1212 candidate donor sites. The attributes of feature vectors for both training and testing set are listed in the column **Attributes**. **Accuracy** is a fraction of correctly classified instances (both positive and negative) given by the classifier; **Sn** is the sensitivity of prediction defined as $\frac{TP}{TP+FN}$; **PPV** is the positive predictive value; **TN** – number of correctly predicted negative instances, **FP** – number of negative instances predicted as positives, **TP** – number of correctly predicted positive instances, **FN** – number of positive instances predicted as negatives.

Based on these results we can infer the following: The fact that there are still 138 real donor sites that are misclassified indicates that the structural signals used are not strong enough in all real introns. However, they seem to be present in a much higher proportion of real sequences than in pseudo sequences (30% of real donor sites are predicted as positives, while this is the case for only 5% of pseudo sites). It is also encouraging to see that when more attributes are used the prediction accuracy improves, indicating that the individual signals have very little predictive power but that they can still contribute when combined with other structural measures. Thus, identifying additional structural signals may further improve donor site prediction accuracy.

The analysis of acceptor site prediction accuracy based on structural signals is, in essence, the same as for the donor sites. The difference is that very few structural features were shown to be significantly different between real and pseudo acceptor sites sites. For acceptor sites, we used the following measures:

- **(ATTA1)** the highest instance score of motif 1 found in the 200-nt window centered at the acceptor splice site (see Section 6.4.2 and Figure 6.24)
- **(ATTA2)** the highest instance score of motif 2 found in the 200-nt window centered at the acceptor splice site
- **(ATTA3)** the highest instance score of motif 3 found in the 200-nt window centered at the acceptor splice site

The dataset of candidate acceptor sequences contained 1764 sequences each of length 200-nt and centered at the acceptor sites predicted by SPL. For each of these sequences, we calculated the three measures and assembled them in feature vectors. We then used the data to train and test neural network and SVM classifiers using 10-fold cross-validation. As was the case for our machine learning experiments for donor sites, the SVM classifier predicted all instances in the dataset as negative. The results of neural network classification were: $accuracy = 0.93$, $Sn = 0.04$, $TN = 1627$,

$FP = 2$, $TP = 6$ and $FN = 129$. Again, the high percentage of correctly classified instances is the result of classifying the majority of pseudo, as well as real sites, as negative instances.

It may be noted that our results from this section cannot be directly compared to the results of Patterson et al. (2002) and Marashi et al. (2006a,b), who also used some structural information to improve the accuracy of splice site prediction. A significant difference between our approaches precludes comparing them: these authors used a combination of structure- and sequence-based measurements for prediction of splice sites, while we used only structural signals for post-processing of SPL's splice site predictions.

### 7.3.2 Improving the accuracy of intron prediction

Based on the findings in the previous section that indicate that combining more structural signals yields better classification accuracy, in this section, we focus on computational prediction of introns and the capability of secondary structure information to reduce the number of false positives based on basic architectural characteristics of yeast introns and structural signals for donor, acceptor and branchpoint sites.

We used the SoftBerry program for splice site prediction, SPL, to predict donor and acceptor sites obtained from our dataset of 98 yeast 5'L introns, 104 5'S introns and 100 randomly chosen yeast genes that do not contain any introns. The details about the dataset and SPL's prediction are given the Section 7.3.1.

The candidate introns were assembled using the predicted splice sites. With no restrictions, except that the identified introns fall within the boundaries of one gene, it is possible to assemble 8625 introns. Since only 133 of these are real introns ($Sn = 0.66$), the positive predictive value (Equation 7.16) of SPL's intron prediction is $PPV = 0.02$, which is extremely low. However, we need to emphasize that the SPL program was not designed for intron prediction, which is typically performed by gene-finding programs using additional sequence content measures (see Section 2.1). As explained in the previous section we were not able to obtain a gene-finding program

trained on yeast sequences, thus we were constrained to use SPL for intron prediction.

We applied several filtering steps based on the basic knowledge of yeast intron architecture to filter out implausible intron candidates.

**Step 1** We selected introns such that they fall entirely within one gene (they cannot be located in two different genes) and their length is within the range of STRIN introns (55-1005 nt). This filtering step resulted in 2652 candidate introns. The $PPV$ after this step is 0.05.

**Step 2** Using a positional weight matrix (PWM) derived from 214 STRIN branchpoint sequences (length of the sequences 7 nt), we scanned for all candidate branchpoint sequences that exceed a threshold value. The threshold is usually chosen to be the minimum value scored by the sequences from which the profile has been derived. In the case of STRIN branchpoint sequences, the minimum weight matrix score is 0.96; however, this is relatively low compared to other scores, and choosing it for the threshold value results in many false positive predictions. Instead, we set the threshold value to 4.47 (the second lowest PWM score for STRIN branchpoint sequences).

The candidate intron sequences from the previous step are further filtered by excluding those that do not contain a branchpoint candidate sequence with a PWM score above the threshold. There are 1274 intron candidates after this step. If there is more than one branchpoint candidate sequence found in an intron, the intron is passed to the next step together with the locations of all candidate branchpoint sequences (1274 is the number of candidate introns, including all possible branchpoint locations; there are 1126 unique intron locations, including 131 real introns). The $PPV$ value after this step is 0.12 (considering only unique intron locations) and the sensitivity is 0.65.

**Step 3** In this step, we check to see if the candidate branchpoint sequence is positioned correctly within a candidate intron (the 5' splice site - branchpoint distance is usually much longer than the 3' splice site - branchpoint

distance in introns). We first verify if the 3' splice site - branchpoint distance is within the range 16-161 nt, as for the real STRIN introns. Unlike the 3' splice site - branchpoint distance, the 5' splice site - branchpoint distance is highly variable and depends on the total length of the intron (see Figure 3.5), therefore we calculate the ratio of these two distances $dist\_ratio = \frac{3'-bp\ dist}{5'-bp\ dist}$ and compare it with the ratio for real introns ($0.035 < dist\_ratio < 2$). There are three introns that have $dist\_ratio > 2$, but trying to include them would also retain many false positive introns. There are 531 intron candidates (542 if we consider all possible branchpoint locations) that satisfy these requirements, 129 of these being real introns ($Sn = 0.64$, $PPV = 0.24$). Using donor and acceptor site locations that define intron boundaries, we assembled intron sequences including 100-nt flanking regions.

In the last phase of our analysis, we attempted to discriminate between the real and pseudo introns by applying machine learning techniques. We used the same attributes as in the previous section, but also included structural signals relating to branchpoint sequences. The attributes are:

- **(ATTD1)** MFE of folding of the 50-nt window located from position $-40$ to $+10$ w.r.t the candidate donor site
- **(ATTD2)** MFE of folding of the 50-nt window located from position $-20$ to $+30$ w.r.t the candidate donor site
- **(ATTB1)** number of unpaired bases of the branchpoint sequences in a candidate intron secondary structure
- **(ATTD3)** MFE of folding between U1 snRNA and 11-nt candidate donor sequence (3 nt from the 5' exon and first 8 intronic nt – NNN|GUAUGUNN)
- **(ATTD4)** MFE of folding between U6 snRNA and 12-nt candidate donor sequence (4 nt from the 5' exon and first 8 intronic nt – NNNN|GUAUGUNN)
- **(ATTB2)** MFE of folding between U2 snRNA and 13-nt candidate branchpoint sequence (three flanking nucleotides on both sides of the

branchpoint sequence – NNNUACUAACNNN) (see Section 6.3.2 and Figure 6.21)

- **(ATTD5)** the highest instance score of motif 3 found in the 200-nt window centered at the donor splice site (see Section 6.4.2 and Figure 6.23)
- **(ATTA1)** the highest instance score of motif 1 found in the 200-nt window centered at the acceptor splice site
- **(ATTA2)** the highest instance score of motif 2 found in the 200-nt window centered at the acceptor splice site
- **(ATTA3)** the highest instance score of motif 3 found in the 200-nt window centered at the acceptor splice site
- **(ATTB3)** the highest instance score of motif 1 found in the 200-nt window centered at the branchpoint sequence (see Section 6.4.2 and Figure 6.25)
- **(ATTB4)** the highest instance score of motif 3 found in the 200-nt window centered at the branchpoint sequence

As in the previous sections, we grouped the attributes by the type of structural signal they measure (group 1 – ATTD1 and ATTD2; group 2 – ATTB1; group 3 – ATTD3, ATTD4 and ATTB2; group 4 – ATTD5, ATTA1, ATTA2, ATTA3, ATTB3 and ATTB4) and used different combinations of these groups for machine learning experiments. The results are given in Table 7.7 and correspond to the average accuracy values from 10-fold cross-validation experiments.

The results in the table are for neural network classification only, since the SVM classifier was predicting all candidate introns as negative instances. Keeping in mind that the goal of this analysis was to reduce the number of false positive predictions, while ensuring that the number of true positives remains as high as possible, we can conclude from the table that the best

| **Attributes** | **Neural Network** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Accuracy** | **Sn** | **PPV** | **TN** | **FP** | **TP** | **FN** |
| group 1 | 0.76 | 0.00 | — | 410 | 0 | 0 | 129 |
| group 1,2 | 0.76 | 0.00 | — | 410 | 0 | 0 | 129 |
| group 3 | 0.74 | 0.12 | 0.34 | 381 | 29 | 15 | 114 |
| group 2,3 | 0.74 | 0.19 | 0.39 | 371 | 39 | 25 | 104 |
| group 1,2,3 | 0.74 | **0.42** | 0.45 | 344 | 66 | **54** | 75 |
| group 4 | 0.74 | 0.09 | 0.36 | 389 | 21 | 12 | 117 |
| group 1,4 | 0.69 | 0.19 | 0.28 | 347 | 63 | 25 | 104 |
| group 1,2,4 | 0.70 | 0.20 | 0.31 | 351 | 59 | 26 | 103 |
| group 3,4 | 0.76 | 0.40 | 0.49 | 355 | 55 | 52 | 77 |
| group 1,2,3,4 | 0.72 | 0.38 | 0.40 | 337 | 73 | 49 | 80 |

Table 7.7: Accuracy results for neural network classification of 542 candidate introns. The attributes of feature vectors for both the training and testing sets are listed in the column **Attributes**. **Accuracy** is the fraction of correctly classified instances (both positive and negative) given by the classifier; **Sn** is the sensitivity of prediction; **PPV** is the positive predictive value; **TN** – number of correctly predicted negative instances, **FP** – number of negative instances predicted as positives, **TP** – number of correctly predicted positive instances, **FN** – number of positive instances predicted as negatives.

result occurs when the ATT1, ATT2, ATT3, ATT4, ATT5, and ATT6 attributes are used. The percentage of correctly classified instances (0.74) is not the highest one in the table, but the sensitivity (0.42) and the number of true positives (54) are. The positive predictive value is 0.45. Compared to the results in Step 3, the sensitivity value decreased from 0.64 to 0.42 and the $PPV$ increased from 0.24 to 0.45. There is an obvious trade-off between $Sn$ and $PPV$, but considering the significant increase in $PPV$ value we can conclude that structural signals found in intronic sequences have the potential to filter out false positive predictions, at the expense of missing some true introns.

To summarize, we started with 2652 intron candidates assembled from the SPL-predicted splice sites (with the length constraint), and using the presence and location of a branchpoint sequence as an additional constraint filtered a significant proportion of false introns (namely, 2121). For the remaining 531 intron candidates, we computed feature vectors containing

values for 12 attributes that were selected based on the results from Chapter 6. This classification step filtered an additional 344 false introns at the expense of misclassifying 75 true introns, leading to a final classification performance characterized by a sensitivity of 0.42 and PPV of 0.45.

## 7.4 Conclusions

In this chapter, we used machine learning techniques to test the predictive power of structural characteristics of the yeast introns discussed in the previous two chapters. First, we tested if the splicing efficency levels can be computed as a function of shortened branchpoint distances and secondary structure probabilities of long introns, obtained by the procedure StructureAnalyze (Figure 5.5, p. 110). Our results show that the classification accuracy is excellent on the small dataset for which we have experimentally determined splicing efficiency levels. When tested on the entire dataset of 5'L STRIN introns, we obtained much lower accuracy results, which is, at least partially, due to the unavailability of splicing efficiency measures for STRIN introns. Similar results are obtained for machine learning experiments using structural summary statistics, such as the minimum and average branchpoint distances and base-pairing probability of contact conformation.

The analyses discussed in Chapter 6 identified certain structural biases in STRIN introns that are statistically significant, but in isolation are not strong enough to unambiguously discriminate between real and pseudo splicing signals. Therefore, in the second part of this chapter, we used combinations of structural signals, with the goal of achieving improved predictive power. The neural network classifier trained to distinguish between real and pseudo donor sites based on four structural features reduced the number of false positive predictions by 95%, improving the initial positive predictive value (PPV) from 0.16 (prediction by SPL, a splice prediction program) to 0.55. However, this approach also misclassified 70% of real donors. Similarly, the neural network classification of the dataset of candidate introns predicted by SPL based on 12 structural features of donor, acceptor and branchpoint

regions reduced the number of false positive predictions by 84%, improving the PPV from 0.24 to 0.42. The PPV for the initial SPL prediction was 0.02, which we improved to 0.24 by using known yeast intron architecture characteristics to filter out unacceptable candidates. The neural network classification misclassified 58% of real introns.

The misclassification of so many real introns is probably due to the fact that our classification approach relies solely upon structural signals, which, as discussed before, are not strong enough to clearly differentiate between real and pseudo sites. This also indicates that sequence-based signals are irreplaceable when it comes to splice site prediction. However, our analysis still shows that the structural characteristics of yeast long introns can be used to significantly reduce the number of false positive predictions, since the identified structural signals seem to be present in a much higher proportion of real than pseudo sequences (30% vs 5%). A better way to use these structural signals would be in combination with sequence-based signals for initial splice site prediction.

# Chapter 8

# Conclusions and future work

Pre-mRNA splicing is one of the essential cellular processes in the pathway leading from DNA to protein. Even though the process has been thoroughly studied since its discovery three decades ago (Chow et al., 1977; Berget et al., 1977), there are still unanswered questions. One question, which we addressed in this thesis, is: how are the splice sites accurately identified and correctly paired across the intron? Part of the answer to this question has been known for quite some time, since the discovery of base-pairing interactions between spliceosomal snRNAs and conserved sequences at the boundaries of and within introns (Mount et al., 1983; Black et al., 1985; Zhuang and Weiner, 1986; Parker and Patterson, 1987). Subsequent studies of a number of spliceosomal protein factors and their complex interactions with pre-mRNA helped elucidate the phenomenon further. But the full answer to the previously posed question is still not clear: scientist are still unable to unambiguously identify functional splice sites among the vast number of pseudo sites that are not used for splicing.

One hypothesis, proposed by many authors, is that not only primary but also secondary structure of pre-mRNA plays a role in splicing. There is a significant body of biological literature that discusses various examples of how secondary structure at intron boundaries or within introns affects splicing. There are also a few computational studies showing that including some structural information improves splice site prediction accuracy (Patterson et al., 2002; Marashi et al., 2006a,b). However, the biological studies are usually limited to one or a small number of genes; thus, the conclusions are gene-specific and do not have broad implications. On the other hand, the computational studies are primarily applications of sophisticated machine learning techniques and only superficially consider the existing biological

evidence.

In this thesis, we attempted to remedy the weaknesses of previous approaches by performing a comprehensive computational study of the structural characteristics of *Saccharomyces cerevisiae's* introns and their possible roles in pre-mRNA splicing. We carefully considered available biological evidence and formulated our hypothesis to be consistent with the current model of splicing in *S. cerevisiae*. We also performed a number of computational and statistical analyses that were previously used for other RNA species to study structural characteristics of real versus artificial, randomly designed, RNAs. Finally, similar to previous work in splice site prediction, we used machine learning classifiers to distinguish between real and pseudo sites based on structural features found to be statistically different between yeast introns and random sequences.

Since a high-quality dataset is essential for any bioinformatics study, as discussed at the beginning of Chapter 3, we constructed the STRIN dataset, which contains all *Saccharomyces cerevisiae* introns whose annotation is consistent with at least two of three public yeast databases considered. The dataset was further filtered to exclude introns that were not supported by the latest comparative genomic study on yeast (Kellis et al., 2003). The availability of reliable datasets allowed us to reexamine basic features of yeast intron architecture that were reported before but on less credible datasets (Parker and Patterson, 1987; Spingola et al., 1999). We confirmed that both intron length and branchpoint distance distributions are bimodal, and that these two intron characteristics are strongly correlated (with a Pearson correlation coefficient $r$ of 0.99). We also used orthologous introns from three yeast species closely related to *S. cerevisiae* to examine the conservation of these architectural intron characteristics. The results from this novel analysis indicate that both intron length and branchpoint distance are very well conserved among these *Saccharomyces sensu stricto* species, despite relatively low conservation of intronic sequences (50-74%). Conservation of these intron features suggests their functional importance.

A number of biological studies conducted in several different eukaryotic species identified long-range basepairing interactions that bring the 5'

and 3' ends of long introns closer together (see Section 2.3). It was hypothesized in these studies that these secondary structure interactions are important for efficient splicing of these introns and that their role is to facilitate spliceosomal assembly. To complement these biological studies, we conducted a computational analysis of yeast introns, searching for stems that would shorten the distance between the splice sites and branchpoint sequences in long (5'L) STRIN introns. Zipper stems, as we named them, were found in all of the 5'L introns in the STRIN dataset, which is not surprising considering the strong tendency of RNA sequences to form basepairing interactions. However, the shortened branchpoint distances of zipped long introns are distributed similarly to the branchpoint distances of short (5'S) yeast introns (which are believed to have optimal branchpoint distances) and very differently from the corresponding distances of zipped random and exonic sequences. This finding suggests that the occurrence of zipper stems in introns is not random but calibrated to modify branchpoint distances in long introns in such a way that the resulting shortened distances resemble 5'S branchpoint distances not only by range but by distribution shape as well.

The results of our comparative structure analysis further support the importance of zipper stems: careful manual analysis on a sample intron dataset of 9 introns found zipper stems conserved between the four considered *sensu stricto* species in all of the introns. This is a significant result, considering that sequence conservation is not very high within the introns (50-74%). Similarly, more automated analysis on a larger set of STRIN introns identified conserved zipper stems in almost all of them. Most of the resulting shortened branchpoint distances fall within the optimal range observed in 5'S introns. The possible reasons why conserved stems were not identified in all of the introns include limited accuracy of the current multiple alignment and RNA secondary structure algorithms, which are essential for the comparative structure analysis approach that we used. Another possibility is that in some cases zipper stems are found at approximately the same locations in all four considered species but the exact base-pairing interactions are not conserved (our automated approach would not identify these stems).

These stems could still perform the same intron-zipping function.

In Chapter 5, we analyzed more specifically the effect of the zipper stems on splicing efficiency of yeast introns. Since splicing efficiency results are not available for all yeast introns but only for a few that have been more extensively studied (Newman, 1987; Goguel and Rosbash, 1993; Libri et al., 1995; Charpentier and Rosbash, 1996; Howe and Ares, 1997), we focused on the existing studies of the RP51B introns, with the intention of formulating a model of structural requirements for efficient splicing. Our initial zipper stem approach was not able to explain the differences in the splicing efficiency levels observed between various RP51B intron mutants, thus we refined our model to be more consistent with the nature of mRNA molecules. We did so by extending our initial zipper stem model to incorporate suboptimal structure predictions and modified calculation of shortened branchpoint distances. The refined model exhibited very good agreement with the results of splicing efficiency studies of RP51B intron mutants with modified secondary structures.

Using this new approach to analyze structural characteristics of all STRIN 5'L introns, we identified an important subgroup of yeast long introns ($\sim 1/3$ of 5'L STRIN introns) that have the same value of shortened branchpoint distance as the RP51B gene (namely, 5). This distance was achieved in very few random sequences with the same sequence characteristics as 5'L introns. Since we are not aware of any systematic splicing efficiency study of yeast introns, we were not able to correlate our findings with splicing efficiency measurements, which would test the significance of our finding.

An important part of our research is validation of our model of structural requirements for splicing by laboratory experiments. The results of our experiments were consistent with our predictions based on the proposed model, and prompted the formulation of further hypotheses regarding the effect of secondary structure on yeast pre-mRNA splicing. Based on the RP51B intron mutants discussed by Libri et al. (1995) and by Charpentier and Rosbash (1996), it seemed that the existence of basepairing interactions between the donor site and the branchpoint sequence in one of the suboptimal secondary structure predictions is a requirement for efficient splicing of

the RP51B intron. Our laboratory experiments confirmed that introns with shorter branchpoint distances ($< 10$) are more efficiently spliced than those with longer distances, but suggest that contact conformation between the donor site and the branchpoint sequence is not a requirement for efficient splicing. Another interesting result of our experimental testing is that the thermodynamic stability of a zipper stem can also have a significant effect on splicing efficiency.

Based on our computational and experimental results, we propose that there are few structural requirements for efficient splicing of yeast long introns. The most important of these is a high probability of a relatively short branchpoint distance (less than 10, but the exact threshold is unknown and should be further investigated) in the intron's secondary structure. This means that there is either one highly probable secondary structure that has a short branchpoint distance or there are many suboptimal structures close to the MFE that have short branchpoint distance. The latter case results in lower average branchpoint distance, which we found to correlate well with observed splicing efficiency levels. Another structural requirement for efficient splicing seems to be limited thermodynamic stability of the zipper stem, since our experiments showed that very stable stems can inhibit splicing. It is possible that very stable zipper stems may be difficult to disrupt, which would hinder binding of spliceosome components to the splice signals. Finally, since the branchpoint sequence is found to be unpaired in secondary structures of all efficiently spliced RP51B mutants, we believe that the structural context of the branchpoint sequence, i.e., its basepairing status in the secondary structure of an intron, also has an effect on splicing efficiency. However, this possibility needs to be further investigated.

In the second part of the thesis we considered a different role for pre-mRNA secondary structure in splicing: can a secondary structure context at the boundaries or within introns or secondary structure interactions with the components of the spliceosome be additional identifiers of intron locations? We used different approaches to look for any structural context that is specific for real yeast introns. We found that donor sites exhibit a statistically significant bias against stable secondary structures when folded locally

and that branchpoint sequences tend to be unbound in global intron folds. We also found that both, donor sites and branchpoint sequences have lower minimum free energy of folding with snRNAs than do pseudo sequences. Finally, we identified structural motifs in the vicinity of donor, acceptor and branchpoint sequences that have a degree of conservation among STRIN introns that is not expected for random occurrences of the motifs.

Even though the structural biases are statistically significant, the corresponding structural signals in isolation are not strong enough to unambiguously discriminate between real and pseudo splicing signals. Therefore, we used different combinations of structural signals to reduce the rate of false positive predictions produced by a sequence-based splice-site prediction program (SPL). The neural network classifier trained to distinguish between real and pseudo donor sites based on four structural features reduced the number of false positive predictions by 95%, improving the initial positive predictive value (PPV) from 0.16 (prediction by SPL) to 0.55. However, this approach also misclassified 70% of real donors (SPL missed only 4 out of 202). Similarly, the neural network classification of the dataset of candidate introns predicted by SPL based on 12 structural features of donor, acceptor and branchpoint regions reduced false positive predictions by 84%, improving the PPV from 0.24 to 0.42. The PPV for the initial SPL prediction was 0.02 which was improved (to 0.24) by using known yeast intron architectural characteristics to filter out unacceptable candidates. The neural network classification misclassified 58% of real introns.

Misclassification of this many real donor sites and introns is probably due to the fact that our prediction approach does not use any sequence signals that have higher information content than the structural signals and are the basis of any splice site prediction tool. The differences in approaches also preclude us from comparing our results with those from previous related machine learning studies (Patterson et al., 2002; Marashi et al., 2006a,b). However, our analysis still shows that the structural characteristics of yeast long introns can be used to significantly reduce the number of false positive predictions since the identified structural signals seem to be present in a much higher proportion of real than pseudo sequences (30% vs 5%).

We believe that combining the structural signals discussed in this thesis with sequence-based signals, would yield better accuracy than the current sequence-based splice site prediction approaches.

Some other contributions of this thesis include:

- An algorithm for identification of one or more stems in a given RNA secondary structure with specific thermodynamic and structural characteristics.
- An algorithm for calculation of distances between the 5' splice site and the branchpoint sequence in a secondary structure of RNA that can also be applied to any pair of nucleotides in an RNA sequence.
- An algorithm for the computation of shortened branchpoint distances and some other structural metrics of an intron secondary structure considering not only the MFE but also suboptimal structure predictions.

In summary, we performed a comprehensive computational study of the secondary structure characteristics of yeast introns and their relationship with pre-mRNA splicing. The computational approach is intended to complement current biological evidence, which suggests an important role for secondary structure elements in splicing of various eukaryotic species, by testing the universality of presumably splicing-related structural features in introns of *Saccharomyces cerevisiae*. Although the computational approach allows us to test hypotheses on a large number of introns, which is infeasible in laboratory experiments, there are a number of uncertainties embedded in the computational analysis that can influence the credibility of results. Some issues that we cannot be certain of include the following: the window of pre-mRNA sequence available for folding at the time of spliceosome assembly, the MFE secondary structure predicted by the currently available RNA prediction programs, which are known to have limited prediction accuracy, and the branchpoint distance, which had to be approximated from the secondary structure prediction. Although one needs to be aware of these limitations, we believe that our computational study offers further insights

into the existence of specific secondary structure features in yeast introns and their role in pre-mRNA splicing.

## Future work

Our work could be extended in several directions. An obvious and straightforward extension would be to obtain splicing efficiency levels for *S.cerevisiae* genes by laboratory experiments and to correlate them with the results of our structural studies; more specifically, the analysis performed in Section 5.5.

Further experimental work would also be beneficial. Additional laboratory experiments that would test the validity of our model of structural requirements for efficient splicing on some other yeast genes are needed to show that our new model is not gene-specific. Another approach would be to design an artificial intron that would contain essential splicing signals, while the intronic sequence itself would be designed as a random sequence. Experiments with the branchpoint distance and secondary structure of this artificial intron would test our model and its universality in yeast.

Laboratory experiments could also be performed to test some of the other hypotheses discussed and computationally analyzed in this thesis, such as the minimum shortened branchpoint distance that is needed for efficient splicing, the effect of branchpoint-sequence structural context on splicing efficiency and acceptable thermodynamic stability of zipper stems.

Another extension of our work would be to test the significance of the conserved structural motifs found in the vicinity of splice signals (Section 6.4.2). This could be done by searching for occurrences of these motifs in the splice site neighbourhoods of a related yeast species that do not have a very high sequence similarity in these regions. Finding the same structural motifs in spite of low sequence conservation would provide strong evidence that these motifs are associated with splice sites and that their structure is probably functionally important. Another way to test the hypothesis that the structural motifs found are functional and important for splicing would be to perform a series of mutational analyses in the laboratory and to search for protein factors that bind to these motifs.

Another important research direction based on the results of this thesis would be to test our findings in other eukaryotic organisms. Since several biological studies indicate that the shortening of the branchpoint distance, either by formation of zipper stems or protein interactions, is important for efficient splicing in *Drosophila melanogaster* and some mammalian species (Chen and Stephan, 2003; Martinez-Contreras et al., 2006), it is plausible that this mechanism is universal for all eukaryotes. The roundworm, *Caenorhabditis elegans*, and the fruit fly, *Drosophila melanogaster*, would be suitable organisms for these further studies, considering that both have been extensively studied as biological model organisms and have been fully sequenced and carefully annotated. The structural signals found in yeast introns (Chapter 6) are another phenomenon that would gain significance if found in other eukaryotic species.

Our machine learning experiments showed that the weak structural signals described in Chapter 6 can serve to filter out a certain portion of false positive splice site predictions. However, we used these signals only in the post-processing phase, where they were not strong enough tho clearly separate real and pseudo sites. Thus, an important extension of this thesis would be to integrate these weak structural signals with sequence-based signals for initial prediction of splice sites, and to test their contribution to prediction accuracy.

# Bibliography

J Abelson, C R Trotta, and H Li. tRNA splicing. *J Biol Chem*, 273(21): 12685–12688, May 1998.

S F Altschul and B W Erickson. Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Mol Biol Evol*, 2(6):526–538, 1985.

R K Alvi, M Lund, and R T Okeefe. ATP-dependent interaction of yeast U5 snRNA loop 1 with the 5' splice site. *RNA*, 7(7):1013–1023, Jul 2001.

M Andronescu. personal communication, 2006.

M Andronescu, V Bereg, H H Hoos, and A Condon. RNA SSTRAND - A new database for RNA secondary structure data and statistical analysis of RNA structural motifs. Submitted for publication.

M Andronescu, Z C Zhang, and A Condon. Secondary structure prediction of interacting RNA molecules. *J Mol Biol*, 345(5):987–1001, Feb 2005.

M S Andronescu. Algorithms for predicting the seconday structure of pairs and combinatorial sets of nucleic acid strands. Master's thesis, Computer Science Department, The University of British Columbia, 2003.

C B Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(96):223–230, Jul 1973.

M Ares, L Grate, and M H Pauling. A handful of intron-containing genes produces the lion's share of yeast mRNA. *RNA*, 5(9):1138–1139, Sep 1999.

M Arita, K Tsuda, and K Asai. Modeling splicing sites with pairwise correlations. *Bioinformatics*, 18 Suppl 2:27–34, 2002.

G Baurén and L Wieslander. Splicing of Balbiani ring 1 gene pre-mRNA occurs simultaneously with transcription. *Cell*, 76(1):183–192, Jan 1994.

D L Bentley. Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Curr Opin Cell Biol*, 17(3):251–256, Jun 2005.

S M Berget, C Moore, and P A Sharp. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A*, 74(8):3171–3175, Aug 1977.

L Betts and L L Spremulli. Analysis of the role of the Shine-Dalgarno sequence and mRNA secondary structure on the efficiency of translational initiation in the Euglena gracilis chloroplast atpH mRNA. *J Biol Chem*, 269(42):26456–26463, Oct 1994.

A Beyer, J Hollunder, H P Nasheuer, and T Wilhelm. Post-transcriptional expression regulation in the yeast Saccharomyces cerevisiae on a genomic scale. *Mol Cell Proteomics*, 3(11):1083–1092, Nov 2004.

A L Beyer and Y N Osheim. Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes Dev*, 2(6):754–765, Jun 1988.

C M Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1996.

D L Black, B Chabot, and J A Steitz. U2 as well as U1 small nuclear ribonucleoproteins are involved in premessenger RNA splicing. *Cell*, 42(3): 737–750, Oct 1985.

M Blanchette and B Chabot. A highly stable duplex structure sequesters the 5' splice site region of hnRNP A1 alternative exon 7B. *RNA*, 3(4): 405–419, Apr 1997.

G Blandin, P Durrens, F Tekaia, M Aigle, M Bolotin-Fukuhara, E Bon, S Casarégola, J de Montigny, C Gaillardin, A Lépingle, B Llorente, A Malpertuy, C Neuvéglise, O Ozier-Kalogeropoulos, A Perrin, S Potier,

J Souciet, E Talla, C Toffano-Nioche, M Wésolowski-Louvel, C Marck, and B Dujon. Genomic exploration of the hemiascomycetous yeasts: 4. The genome of Saccharomyces cerevisiae revisited. *FEBS Lett*, 487(1):31–36, Dec 2000.

P Brion and E Westhof. Hierarchy and dynamics of RNA folding. *Annu Rev Biophys Biomol Struct*, 26:113–137, 1997.

D A Brow. Allosteric cascade of spliceosome activation. *Annu Rev Genet*, 36:333–360, 2002.

M Brudno, C B Do, G M Cooper, M F Kim, E Davydov, E D Green, A Sidow, and S Batzoglou. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*, 13(4): 721–731, Apr 2003.

F Brulé, A Grégoire, V Ségault, A Mougin, and C Branlant. Secondary structure conservation of the U3 small nucleolar RNA introns in Saccharomyces. *C R Acad Sci III*, 318(12):1197–1206, Dec 1995.

S Brunak, J Engelbrecht, and S Knudsen. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J Mol Biol*, 220(1): 49–65, Jul 1991.

C G Burd and G Dreyfuss. Conserved structures and diversity of functions of RNA-binding proteins. *Science*, 265(5172):615–621, Jul 1994.

C Burge. *Identification of complete gene structure in human genomics DNA*. PhD thesis, Department of Mathematics, Stanford University, 1997.

L Cartegni and A R Krainer. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. *Nat Genet*, 30(4):377–384, Apr 2002.

J L Casey, M W Hentze, D M Koeller, S W Caughman, T A Rouault, R D Klausner, and J B Harford. Iron-responsive elements: regulatory RNA sequences that control mRNA levels and translation. *Science*, 240(4854): 924–928, May 1988.

B Charpentier and M Rosbash. Intramolecular structure in yeast introns aids the early steps of in vitro spliceosome assembly. *RNA*, 2(6):509–522, Jun 1996.

K Chebli, R Gattoni, P Schmitt, G Hildwein, and J Stevenin. The 216-nucleotide intron of the E1A pre-mRNA contains a hairpin structure that permits utilization of unusually distant branch acceptors. *Mol Cell Biol*, 9 (11):4852–4861, Nov 1989.

Y Chen and W Stephan. Compensatory evolution of a precursor messenger RNA secondary structure in the Drosophila melanogaster Adh gene. *Proc Natl Acad Sci U S A*, 100(20):11499–11504, Sep 2003.

L T Chow, R E Gelinas, T R Broker, and R J Roberts. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*, 12(1):1–8, Sep 1977.

R E Christoffersen and D Mcswiggen, J and. Application of computational technologies to ribozyme biotechnology products. *J Mol Structure*, 311: 273–284, Jul 1994.

A E Churbanov, I B Rogozin, J S Deogun, and H H Ali. Method of predicting Splice Sites based on signal interactions. *Biol Direct*, 1(1):10–10, Apr 2006.

T A Clark, C W Sugnet, and M Ares. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science*, 296(5569): 907–910, May 2002.

P Clote, F Ferré, E Kranakis, and D Krizanc. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, 11:578–591, 2005.

B Clouet d'Orval, Y d'Aubenton Carafa, P Sirand-Pugnet, M Gallego, E Brody, and J Marie. RNA secondary structure repression of a muscle-specific exon in HeLa cell nuclear extracts. *Science*, 252(5014):1823–1828, Jun 1991.

T P Coleman and J R Roesser. RNA secondary structure: an important cis-element in rat calcitonin/CGRP pre-messenger RNA splicing. *Biochemistry*, 37(45):15941–15950, Nov 1998.

N Cristianini and J Shawe-Taylor. *An Intorduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, 2000.

J M Deleo. Receiver operating characteristic laboratory (ROCLAB): software for developing decision strategies that account for uncertainty. In *Proceedings of the Second International Symposium on Uncertainty Modelling and Analysis*, pages 318–325. IEEE Computer Society Press, 1993.

J O Deshler, G P Larson, and J J Rossi. Kluyveromyces lactis maintains Saccharomyces cerevisiae intron-encoded splicing signals. *Mol Cell Biol*, 9 (5):2208–2213, May 1989.

E W Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959.

Y Ding, C Y Chan, and C E Lawrence. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, 11(8):1157–1166, Aug 2005.

Y Ding, C Y Chan, and C E Lawrence. Clustering of RNA secondary structures with application to messenger RNAs. *to appear in J Mol Biol*, 2006.

Y Ding and C E Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res*, 31(24):7280–7301, Dec 2003.

A Diwa, A L Bricker, C Jain, and J G Belasco. An evolutionarily conserved RNA stem-loop functions as a sensor that directs feedback regulation of RNase E gene expression. *Genes Dev*, 14(10):1249–1260, May 2000.

S R Eddy. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, 3:18–18, Jul 2002.

D J Elliott and M Rosbash. Yeast pre-mRNA is composed of two populations with distinct kinetic properties. *Exp Cell Res*, 229(2):181–188, Dec 1996.

F J Eng and J R Warner. Structural basis for the regulation of splicing of a yeast messenger RNA. *Cell*, 65(5):797–804, May 1991.

L P Eperon, I R Graham, A D Griffiths, and I C Eperon. Effects of RNA secondary structure on alternative splicing of pre-mRNA: is folding limited to a region behind the transcribing RNA polymerase? *Cell*, 54(3):393–401, Jul 1988.

N A Faustino and T A Cooper. Pre-mRNA splicing and human disease. *Genes Dev*, 17(4):419–437, Feb 2003.

J W Fickett. The gene identification problem: an overview for developers. *Computers Chem*, 20:103–118, 1996.

S M Freier, R Kierzek, J A Jaeger, N Sugimoto, M H Caruthers, T Neilson, and D H Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Proc Natl Acad Sci U S A*, 83(24):9373–9377, Dec 1986.

E Freyhult, P P Gardner, and V Moulton. A comparison of RNA folding measures. *BMC Bioinformatics*, 6:241–241, 2005.

J A Garland and D P Aalberts. Thermodynamic modeling of donor splice site recognition in pre-mRNA. *Phys Rev E Stat Nonlin Soft Matter Phys*, 69(4 Pt 1):041903–041903, Apr 2004.

R Gattoni, P Schmitt, and J Stevenin. In vitro splicing of adenovirus E1A transcripts: characterization of novel reactions and of multiple branch points abnormally far from the 3' splice site. *Nucleic Acids Res*, 16(6): 2389–2409, Mar 1988.

S Ghaemmaghami, W K Huh, K Bower, R W Howson, A Belle, N Dephoure, E K O'Shea, and J S Weissman. Global analysis of protein expression in yeast. *Nature*, 425(6959):737–741, Oct 2003.

A Goffeau, B G Barrell, H Bussey, R W Davis, B Dujon, H Feldmann, F Galibert, J D Hoheisel, C Jacq, M Johnston, E J Louis, H W Mewes, Y Murakami, P Philippsen, H Tettelin, and S G Oliver. Life with 6000 genes. *Science*, 274(5287):563–567, Oct 1996.

V Goguel and M Rosbash. Splice site choice and splicing efficiency are positively influenced by pre-mRNA intramolecular base pairing in yeast. *Cell*, 72(6):893–901, Mar 1993.

V Goguel, Y Wang, and M Rosbash. Short artificial hairpins sequester splicing signals and inhibit yeast pre-mRNA splicing. *Mol Cell Biol*, 13 (11):6841–6848, Nov 1993.

J Görnemann, K M Kotovic, K Hujer, and K M Neugebauer. Cotranscriptional spliceosome assembly occurs in a stepwise fashion and requires the cap binding complex. *Mol Cell*, 19(1):53–63, Jul 2005.

J Gorodkin, L J Heyer, and G D Stormo. Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res*, 25(18):3724–3732, Sep 1997.

L Grate and M Ares. Searching yeast intron data at Ares lab Web site. *Methods Enzymol*, 350:380–392, 2002.

D Greenbaum, C Colangelo, K Williams, and M Gerstein. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol*, 4(9):117–117, 2003.

A Grover, H Houlden, M Baker, J Adamson, J Lewis, G Prihar, S Pickering-Brown, K Duff, and M Hutton. 5' splice site mutations in tau associated with the inherited dementia FTDP-17 affect a stem-loop structure that regulates alternative splicing of exon 10. *J Biol Chem*, 274(21):15134–15143, May 1999.

A P Gultyaev, F H van Batenburg, and C W Pleij. The computer simulation of RNA folding pathways using a genetic algorithm. *J Mol Biol*, 250 (1):37–51, Jun 1995.

R R Gutell, J C Lee, and J J Cannone. The accuracy of ribosomal RNA comparative structure models. *Curr Opin Struct Biol*, 12(3):301–310, Jun 2002.

S P Gygi, Y Rochon, B R Franza, and R Aebersold. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol*, 19(3):1720–1730, Mar 1999.

K B Hall, M R Green, and A G Redfield. Structure of a pre-mRNA branch point/3' splice site region. *Proc Natl Acad Sci U S A*, 85(3):704–708, Feb 1988.

S M Hebsgaard, P G Korning, N Tolstrup, J Engelbrecht, P Rouzé, and S Brunak. Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. *Nucleic Acids Res*, 24 (17):3439–3452, Sep 1996.

M W Hentze, S W Caughman, J L Casey, D M Koeller, T A Rouault, J B Harford, and R D Klausner. A model for the structure and functions of iron-responsive elements. *Gene*, 72(1-2):201–208, Dec 1988.

K J Hertel, K W Lynch, and T Maniatis. Common themes in the function of transcription and splicing enhancers. *Curr Opin Cell Biol*, 9(3):350–357, Jun 1997.

H Hieronymus and P A Silver. A systems view of mRNP biology. *Genes Dev*, 18(23):2845–2860, Dec 2004.

M Höchsmann, T Töller, R Giegerich, and S Kurtz. Local similarity in RNA secondary structures. *Proc IEEE Comput Soc Bioinform Conf*, 2: 159–168, 2003.

I L Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res*, 31(13):3429–3431, Jul 2003.

I L Hofacker, M Fekete, C Flamm, M A Huynen, S Rauscher, P E Stolorz, and P F Stadler. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucleic Acids Res*, 26(16):3825–3836, Aug 1998.

I L Hofacker, M Fekete, and P F Stadler. Secondary structure prediction for aligned RNA sequences. *J Mol Biol*, 319(5):1059–1066, Jun 2002.

I L Hofacker, W Fontana, L P F Stadler, S Bonhoeffer, M Tacker, and P Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh Chem*, 125:167–188, 1994.

I L Hofacker, B Priwitzer, and P F Stadler. Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, 20 (2):186–190, Jan 2004.

I L Hofacker and P F Stadler. Automatic detection of conserved base pairing patterns in RNA virus genomes. *Comput Chem*, 23(3-4):401–414, Jun 1999.

M Hollander and D A Wolfe. *Nonparametric statistical methods*. John Wiley & Sons, 2 edition, 1999.

K J Howe and M Ares. Intron self-complementarity enforces exon inclusion in a yeast pre-mRNA. *Proc Natl Acad Sci U S A*, 94(23):12467–12472, Nov 1997.

E H Hurowitz and P O Brown. Genome-wide analysis of mRNA lengths in Saccharomyces cerevisiae. *Genome Biol*, 5(1), 2003.

M Hutton, C L Lendon, P Rizzu, M Baker, S Froelich, H Houlden, S Pickering-Brown, S Chakraverty, A Isaacs, A Grover, J Hackett, J Adamson, S Lincoln, D Dickson, P Davies, R C Petersen, M Stevens, E de Graaff, E Wauters, J van Baren, M Hillebrand, M Joosse, J M Kwon, P Nowotny, L K Che, J Norton, J C Morris, L A Reed, J Trojanowski, H Basun, L Lannfelt, M Neystat, S Fahn, F Dark, T Tannenberg, P R Dodd, N Hayward, J B Kwok, P R Schofield, A Andreadis, J Snowden, D Craufurd, D Neary,

F Owen, B A Oostra, J Hardy, A Goate, J van Swieten, D Mann, T Lynch, and P Heutink. Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature*, 393(6686):702–705, Jun 1998.

A H Igel and M Ares. Internal sequences that distinguish yeast from metazoan U2 snRNA are unnecessary for pre-mRNA splicing. *Nature*, 334 (6181):450–453, Aug 1988.

T H Jensen, K Dower, D Libri, and M Rosbash. Early formation of mRNP: license for export or quality control? *Mol Cell*, 11(5):1129–1138, May 2003.

T L Johnson and J Abelson. Characterization of U4 and U6 interactions with the 5' splice site using a S. cerevisiae in vitro trans-splicing system. *Genes Dev*, 15(15):1957–1970, Aug 2001.

S Kandels-Lewis and B Séraphin. Involvement of U6 snRNA in 5' splice site selection. *Science*, 262(5142):2035–2039, Dec 1993.

M Karin, R Najarian, A Haslinger, P Valenzuela, J Welch, and S Fogel. Primary structure and transcription of an amplified genetic locus: the CUP1 locus of yeast. *Proc Natl Acad Sci U S A*, 81(2):337–341, Jan 1984.

T Kashima and J L Manley. A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. *Nat Genet*, 34(4):460–463, Aug 2003.

L Katz and C B Burge. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res*, 13(9):2042–2051, Sep 2003.

M Kellis. *Computational Comparative Genomics: Genes, Regulation, Evolution*. PhD thesis, Computer Science Department, MIT, 2003.

M Kellis, N Patterson, M Endrizzi, B Birren, and ES Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254, May 2003.

F J Klinz and D Gallwitz. Size and position of intervening sequences are critical for the splicing efficiency of pre-mRNA in the yeast Saccharomyces cerevisiae. *Nucleic Acids Res*, 13(11):3791–3804, Jun 1985.

B Knudsen and J Hein. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15(6): 446–454, Jun 1999.

K Köhrer and H Domdey. Splicing and spliceosome formation of the yeast MATa1 transcript require a minimum distance from the 5' splice site to the internal branch acceptor site. *Nucleic Acids Res*, 16(20):9457–9475, Oct 1988.

K M Kotovic, D Lockshon, L Boric, and K M Neugebauer. Cotranscriptional recruitment of the U1 snRNP to intron-containing genes in yeast. *Mol Cell Biol*, 23(16):5768–5779, Aug 2003.

M Kowalczuk, P Mackiewicz, A Gierlik, M R Dudek, and S Cebrat. Total number of coding open reading frames in the yeast genome. *Yeast*, 15(11): 1031–1034, Aug 1999.

M Krawczak, J Reiss, and D N Cooper. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum Genet*, 90(1-2):41–54, Sep-Oct 1992.

J M Kreahling and B R Graveley. The iStem, a long-range RNA secondary structure element required for efficient exon inclusion in the Drosophila Dscam pre-mRNA. *Mol Cell Biol*, 25(23):10251–10260, Dec 2005.

L Kretzner, A Krol, and M Rosbash. Saccharomyces cerevisiae U1 small nuclear RNA secondary structure contains both universal and yeast-specific domains. *Proc Natl Acad Sci U S A*, 87(2):851–855, Jan 1990.

D Kulp, D Haussler, M G Reese, and F H Eeckman. A generalized hidden Markov model for the recognition of human genes in DNA. *Proc Int Conf Intell Syst Mol Biol*, 4:134–142, 1996.

D M Kupfer, S D Drabenstot, K L Buchanan, H Lai, H Zhu, D W Dyer, B A Roe, and J W Murphy. Introns and splicing elements of five diverse fungi. *Eukaryot Cell*, 3(5):1088–1100, Oct 2004.

C J Langford, F J Klinz, C Donath, and D Gallwitz. Point mutations identify the conserved, intron-contained TACTAAC box as an essential splicing signal sequence in yeast. *Cell*, 36(3):645–653, Mar 1984.

D M Layton and R Bundschuh. A statistical analysis of RNA folding algorithms through thermodynamic parameter perturbation. *Nucleic Acids Res*, 33(2):519–524, 2005.

J D Lewis, D Görlich, and I W Mattaj. A yeast cap binding protein complex (yCBC) acts at an early step in pre-mRNA splicing. *Nucleic Acids Res*, 24 (17):3332–3336, Sep 1996a.

J D Lewis, E Izaurralde, A Jarmolowski, C McGuigan, and I W Mattaj. A nuclear cap-binding complex facilitates association of U1 snRNP with the cap-proximal 5' splice site. *Genes Dev*, 10(13):1683–1698, Jul 1996b.

D Libri, F Duconge, L Levy, and M Vinauger. A role for the Psi-U mismatch in the recognition of the 5' splice site of yeast introns by the U1 small nuclear ribonucleoprotein particle. *J Biol Chem*, 277(20):18173–18181, May 2002.

D Libri, A Piseri, and M Y Fiszman. Tissue-specific splicing in vivo of the beta-tropomyosin gene: dependence on an RNA secondary structure. *Science*, 252(5014):1842–1845, Jun 1991.

D Libri, F Stutz, T McCarthy, and M Rosbash. RNA structural patterns and splicing: molecular basis for an RNA-based enhancer. *RNA*, 1(4): 425–436, Jun 1995.

J Lin and J J Rossi. Identification and characterization of yeast mutants that overcome an experimentally introduced block to splicing at the 3' splice site. *RNA*, 2(8):835–848, Aug 1996.

H X Liu, G J Goodall, R Kole, and W Filipowicz. Effects of secondary structure on pre-mRNA splicing: hairpins sequestering the 5' but not the 3' splice site inhibit intron processing in Nicotiana plumbaginifolia. *EMBO J*, 14(2):377–388, Jan 1995.

P J Lopez and B Séraphin. YIDB: the Yeast Intron DataBase. *Nucleic Acids Res*, 28(1):85–86, Jan 2000.

R Lück, S Gräf, and G Steger. ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucleic Acids Res*, 27(21):4208–4217, Nov 1999.

A V Lukashin and M Borodovsky. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res*, 26(4):1107–1115, Feb 1998.

B G Luukkonen and B Séraphin. The role of branchpoint-3' splice site spacing and interaction between intron terminal nucleotides in 3' splice site selection in Saccharomyces cerevisiae. *EMBO J*, 16(4):779–792, Feb 1997.

R B Lyngsø and C N Pedersen. RNA pseudoknot prediction in energy-based models. *J Comput Biol*, 7(3-4):409–427, 2000.

H D Madhani and C Guthrie. Dynamic RNA-RNA interaction in the spliceosome. *Annu Rev Genet*, 28:1–26, 1994.

S A Marashi, C Eslahchi, H Pezeshk, and M Sadeghi. Impact of RNA structure on the prediction of donor and acceptor splice sites. *BMC Bioinformatics*, 7(1):297–297, Jun 2006a.

S A Marashi, H Goodarzi, M Sadeghi, C Eslahchi, and H Pezeshk. Importance of RNA secondary structure information for yeast donor and acceptor splice site predictions by neural networks. *Comput Biol Chem*, 30(1):50–57, Feb 2006b.

R Martinez-Contreras, J F Fisette, F U Nasim, R Madden, M Cordeau, and B Chabot. Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. *PLoS Biol*, 4(2), Feb 2006.

D H Mathews, J Sabina, M Zuker, and D H Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, 288(5):911–940, May 1999.

D H Mathews and D H Turner. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J Mol Biol*, 317(2): 191–203, Mar 2002.

DH Mathews. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, 10(8):1178–1190, Aug 2004.

J S McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, May-Jun 1990.

T S McConnell and J A Steitz. Proximity of the invariant loop of U5 snRNA to the second intron residue during pre-mRNA splicing. *EMBO J*, 20(13):3577–3586, Jul 2001.

H W Mewes, D Frishman, U Güldener, G Mannhaupt, K Mayer, M Mokrejs, B Morgenstern, M Münsterkötter, S Rudd, and B Weil. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*, 30(1): 31–34, Jan 2002.

H Miyaso, M Okumura, S Kondo, S Higashide, H Miyajima, and K Imaizumi. An intronic splicing enhancer element in survival motor neuron (SMN) pre-mRNA. *J Biol Chem*, 278(18):15825–15831, May 2003.

M J Moore. Intron recognition comes of AGe. *Nat Struct Biol*, 7(1):14–16, Jan 2000.

M J Moore. From birth to death: the complex lives of eukaryotic mRNAs. *Science*, 309(5740):1514–1518, Sep 2005.

S R Morgan and P G Higgs. Evidence for kinetic effects in the folding of large RNA molecules. *J Chem Phys*, 105(16):7152–7157, 1996.

A Mougin, A Grégoire, J Banroques, V Ségault, R Fournier, F Brulé, M Chevrier-Miller, and C Branlant. Secondary structure of the yeast Saccharomyces cerevisiae pre-U3A snoRNA and its implication for splicing efficiency. *RNA*, 2(11):1079–1093, Nov 1996.

S M Mount, I Pettersson, M Hinterberger, A Karmas, and J A Steitz. The U1 small nuclear RNA-protein complex selectively binds a 5' splice site in vitro. *Cell*, 33(2):509–518, Jun 1983.

U Mückstein, H Tafer, J Hackermüller, S H Bernhart, P F Stadler, and I L Hofacker. Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22 (10):1177–1182, May 2006.

K M Neugebauer. On the importance of being co-transcriptional. *J Cell Sci*, 115(Pt 20):3865–3871, Oct 2002.

A Newman. Specific accessory sequences in Saccharomyces cerevisiae introns control assembly of pre-mRNAs into spliceosomes. *EMBO J*, 6(12): 3833–3839, Dec 1987.

A J Newman and C Norman. U5 snRNA interacts with exon sequences at 5' and 3' splice sites. *Cell*, 68(4):743–754, Feb 1992.

Y N Osheim, O L Miller, and A L Beyer. RNP particles at splice junction sequences on Drosophila chorion transcripts. *Cell*, 43(1):143–151, Nov 1985.

R A Padgett, P J Grabowski, M M Konarska, S Seiler, and P A Sharp. Splicing of messenger RNA precursors. *Annu Rev Biochem*, 55:1119–1150, 1986.

R Parker and B Patterson. Architecture of fungal introns: implications for spliceosome assembly. In M Inouye and B S Dudock, editors, *Molecular biology of RNA: new perspectives*, pages 133–149. Academic Press, Inc., San Diego, CA, USA, 1987.

R Parker, P G Siliciano, and C Guthrie. Recognition of the TACTAAC box during mRNA splicing in yeast involves base pairing to the U2-like snRNA. *Cell*, 49(2):229–239, Apr 1987.

D J Patterson, K Yasuhara, and W L Ruzzo. Pre-mRNA secondary structure prediction aids splice site prediction. In R B Altman, A K Dunker, L Hunter, and T E Klein, editors, *Pacific Symposium on Biocomputing*, pages 223–234. World Scientific, 2002.

M Pertea, X Lin, and S L Salzberg. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res*, 29(5):1185–1190, Mar 2001.

C W Pikielny and M Rosbash. mRNA splicing efficiency in yeast and the contribution of nonconserved sequences. *Cell*, 41(1):119–126, May 1985.

C W Pleij and L Bosch. RNA pseudoknots: structure, detection, and prediction. *Methods Enzymol*, 180:289–303, 1989.

D I Poole, A K Mackworth, and R G Goebel. *Computational Intelligence: A Logical Approach*. Oxford University Press, 1998.

P J Preker and C Guthrie. Autoregulation of the mRNA export factor Yra1p requires inefficient splicing of its pre-mRNA. *RNA*, 12(6):994–1006, Jun 2006.

C C Query, M J Moore, and P A Sharp. Branch nucleophile selection in pre-mRNA splicing: evidence for the bulged duplex model. *Genes Dev*, 8 (5):587–597, Mar 1994.

J Reeder and R Giegerich. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, 5:104–104, Aug 2004.

M Reese, D Kulp, A Gentles, and U Ohler. Representative benchmark data sets of human DNA sequences, 1999. http://www.fruitfly.org/sequence/human-datasets.html.

M Rehmsmeier, P Steffen, M Hochsmann, and R Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–1517, Oct 2004.

J Ren, B Rastegari, A Condon, and H H Hoos. HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA*, 11(10): 1494–1504, Oct 2005.

E Rivas and S R Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J Mol Biol*, 285(5):2053–2068, Feb 1999.

E Rivas and S R Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 16 (7):583–605, 2000.

X Roca, R Sachidanandam, and A R Krainer. Determinants of the inherent strength of human 5' splice sites. *RNA*, 11(5):683–698, May 2005.

S Rogic, A K Mackworth, and F B Ouellette. Evaluation of gene-finding programs on mammalian sequences. *Genome Res*, 11(5):817–832, May 2001.

B C Rymond and M Rosbash. Yeast pre-mRNA splicing. In *The molecular and cellular biology of the yeast Saccharomyces: gene expression*, volume 2, pages 143–192. Cold Spring Harbor Laboratory Press, 1992.

S L Salzberg. A method for identifying splice sites and translational start sites in eukaryotic mRNA. *Comput Appl Biosci*, 13(4):365–376, Aug 1997.

H Sawa and J Abelson. Evidence for a base-pairing interaction between U6 small nuclear RNA and 5' splice site during the splicing reaction in yeast. *Proc Natl Acad Sci U S A*, 89(23):11269–11273, Dec 1992.

W Seffens and D Digby. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res*, 27 (7):1578–1584, Apr 1999.

D Solnick. Alternative splicing caused by RNA secondary structure. *Cell*, 43(3 Pt 2):667–676, Dec 1985.

V V Solovyev, A A Salamov, and C B Lawrence. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res*, 22(24):5156–5163, Dec 1994.

E J Sontheimer and J A Steitz. The U5 and U6 small nuclear RNAs as active site components of the spliceosome. *Science*, 262(5142):1989–1996, Dec 1993.

M Spingola, L Grate, D Haussler, and M Ares. Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae. *RNA*, 5(2): 221–234, Feb 1999.

J P Staley and C Guthrie. Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell*, 92(3):315–326, Feb 1998.

J P Staley and C Guthrie. An RNA switch at the 5' splice site requires ATP and the DEAD box protein Prp28p. *Mol Cell*, 3(1):55–64, Jan 1999.

W Stephan and D A Kirby. RNA folding in Drosophila shows a distance effect for compensatory fitness interactions. *Genetics*, 135(1):97–103, Sep 1993.

F Stutz and M Rosbash. A functional interaction between Rev and yeast pre-mRNA is related to splicing complex formation. *EMBO J*, 13(17): 4096–4104, Sep 1994.

H Sun and L A Chasin. Multiple splicing defects in an intronic false exon. *Mol Cell Biol*, 20(17):6414–6425, Sep 2000.

C N Tennyson, H J Klamut, and R G Worton. The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. *Nat Genet*, 9(2):184–190, Feb 1995.

J D Thompson, D G Higgins, and T J Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–4680, Nov 1994.

S Thompson-Jäger and H Domdey. Yeast pre-mRNA splicing requires a minimum distance between the 5' splice site and the internal branch acceptor site. *Mol Cell Biol*, 7(11):4010–4016, Nov 1987.

I Tinoco and C Bustamante. How RNA folds. *J Mol Biol*, 293(2):271–281, Oct 1999.

H Touzet and O Perriquet. CARNAC: folding families of related RNAs. *Nucleic Acids Res*, 32(Web Server issue):142–145, Jul 2004.

D H Turner and N Sugimoto. RNA structure prediction. *Annu Rev Biophys Biophys Chem*, 17:167–192, 1988.

D H Turner, N Sugimoto, J A Jaeger, C E Longfellow, S M Freier, and R Kierzek. Improved parameters for prediction of RNA structure. *Cold Spring Harb Symp Quant Biol*, 52:123–133, 1987.

J G Umen and C Guthrie. The second catalytic step of pre-mRNA splicing. *RNA*, 1(9):869–885, Nov 1995.

V E Velculescu, L Zhang, W Zhou, J Vogelstein, M A Basrai, D E Bassett, P Hieter, B Vogelstein, and K W Kinzler. Characterization of the yeast transcriptome. *Cell*, 88(2):243–251, Jan 1997.

L Vignal, F Lisacek, J Quinqueton, Y d'Aubenton Carafa, and C Thermes. A multi-agent system simulating human splice site recognition. *Comput Chem*, 23(3-4):219–231, Jun 1999.

J D Wagner, E Jankowsky, M Company, A M Pyle, and J N Abelson. The DEAH-box protein PRP22 is an ATPase that mediates ATP-dependent mRNA release from the spliceosome and unwinds RNA duplexes. *EMBO J*, 17(10):2926–2937, May 1998.

Y Wang, J D Wagner, and C Guthrie. The DEAH-box splicing factor Prp16 unwinds RNA duplexes in vitro. *Curr Biol*, 8(8):441–451, Apr 1998.

I Wetterberg, G Baurén, and L Wieslander. The intranuclear site of excision of each intron in Balbiani ring 3 pre-mRNA is influenced by the time

remaining to transcription termination and different excision efficiencies for the various introns. *RNA*, 2(7):641–651, Jul 1996.

I H Witten and E Frank. *Data Mining: Practical machine learning tools and techniques"*. Morgan Kaufmann, San Francisco, 2 edition, 2005.

V Wood, K M Rutherford, A Ivens, M A Rajandream, and B Barrell. A re-annotation of the *Saccharomyces cerevisiae* genome. *Comp Funct Genome*, 2:143–154, 2001.

C Workman and A Krogh. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res*, 27(24):4816–4822, Dec 1999.

J Wu and J L Manley. Mammalian pre-mRNA branch site selection by U2 snRNP involves base pairing. *Genes Dev*, 3(10):1553–1561, Oct 1989.

J Wuarin and U Schibler. Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. *Mol Cell Biol*, 14(11):7219–7225, Nov 1994.

S Wuchty, W Fontana, I L Hofacker, and P Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49 (2):145–165, Feb 1999.

T Xia, J SantaLucia, M E Burkard, R Kierzek, S J Schroeder, X Jiao, C Cox, and D H Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37(42):14719–14735, Oct 1998.

Z Yao, Z Weinberg, and W L Ruzzo. Cmfinder–a covariance model based rna motif finding algorithm. *Bioinformatics*, 22(4):445–452, Feb 2006.

M C Yu, F Bachand, A E McBride, S Komili, J M Casolari, and P A Silver. Arginine methyltransferase affects interactions and recruitment of mRNA processing and export factors. *Genes Dev*, 18(16):2024–2035, Aug 2004.

P D Zamore, M L Zapp, and M R Green. Gene expression. RNA binding: beta s and basics. *Nature*, 348(6301):485–486, Dec 1990.

Z Zhang and F S Dietrich. Mapping of transcription start sites in Saccharomyces cerevisiae using 5' SAGE. *Nucleic Acids Res*, 33(9):2838–2851, 2005.

Y Zhuang and A M Weiner. A compensatory base change in U1 snRNA suppresses a 5' splice site mutation. *Cell*, 46(6):827–835, Sep 1986.

Y Zhuang and A M Weiner. A compensatory base change in human U2 snRNA can suppress a branch site mutation. *Genes Dev*, 3(10):1545–1552, Oct 1989.

M Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, 31(13):3406–3415, Jul 2003.

M Zuker and A B Jacobson. Using reliability information to annotate RNA secondary structures. *RNA*, 4(6):669–679, Jun 1998.

# Appendix A

# The STRIN dataset

The list of introns in the STRIN dataset is given below. The following information is given: the SGD name of a gene that contains the intron, intron length, the 7-nt branchpoint sequence and the branchpoint distance (distance from the 5' splice site to the branchpoint sequence). Introns with branchpoint distances less than 200 nt are 5'S introns and the rest are 5'L introns (see discussion in Section 3.2.2).

| | **gene name** | **intron length** | **branchpoint sequence** | **branchpoint distance** |
|---|---|---|---|---|
| 1 | YAL001C | 90 | UACUAAC | 71 |
| 2 | YAL003W | 366 | UACUAAC | 326 |
| 3 | YAL030W | 113 | UACUAAC | 45 |
| 4 | YBL018C | 75 | UACUAAC | 50 |
| 5 | YBL026W | 128 | UACUAAC | 86 |
| 6 | YBL027W | 384 | UACUAAC | 337 |
| 7 | YBL040C | 97 | UACUAAC | 57 |
| 8 | YBL050W | 116 | UACUAAC | 71 |
| 9 | YBL059C-A | 85 | UACUAAC | 51 |
| 10 | YBL059W | 69 | UACUAAC | 45 |
| 11 | YBL072C | 308 | UACUAAC | 275 |
| 12 | YBL087C | 504 | UACUAAC | 454 |
| 13 | YBL092W | 333 | UACUAAC | 289 |
| 14 | YBR048W | 511 | UACUAAC | 486 |
| 15 | YBR078W | 330 | UACUAAC | 300 |
| 16 | YBR082C | 95 | UACUAAC | 64 |
| 17 | YBR084C-A | 506 | AACUAAC | 486 |
| 18 | YBR090C | 357 | UACUAAC | 338 |
| 19 | YBR119W | 89 | UACUAAC | 44 |
| 20 | YBR181C | 352 | UACUAAC | 324 |
| 21 | YBR186W | 113 | CACUAAC | 71 |
| 22 | YBR189W | 413 | GACUAAC | 376 |
| 23 | YBR191W | 388 | UACUAAC | 334 |
| 24 | YBR230C | 97 | GACUAAC | 56 |
| 25 | YCR028C-A | 83 | UACUAAC | 52 |

| | gene name | intron length | branchpoint sequence | branchpoint distance |
|---|---|---|---|---|
| 26 | YCR031C | 307 | UACUAAC | 252 |
| 27 | YDL029W | 123 | UACUAAC | 103 |
| 28 | YDL061C | 409 | UACUAAC | 365 |
| 29 | YDL064W | 110 | UACUAAC | 66 |
| 30 | YDL075W | 421 | UACUAAC | 386 |
| 31 | YDL079C | 292 | UACUAAC | 275 |
| 32 | YDL082W | 365 | UACUAAC | 320 |
| 33 | YDL083C | 432 | UACUAAC | 385 |
| 34 | YDL108W | 81 | UACUAAC | 49 |
| 35 | YDL125C | 111 | UACUAAC | 83 |
| 36 | YDL130W | 301 | UACUAAC | 237 |
| 37 | YDL136W | 405 | UACUAAC | 362 |
| 38 | YDL191W | 491 | UACUAAC | 437 |
| 39 | YDL219W | 71 | UACUAAC | 45 |
| 40 | YDR005C | 80 | UACUAAC | 42 |
| 41 | YDR025W | 339 | UACUAAC | 310 |
| 42 | YDR059C | 90 | UACUAAC | 52 |
| 43 | YDR064W | 539 | UACUAAC | 480 |
| 44 | YDR092W | 268 | UACUAAC | 107 |
| 45 | YDR129C | 111 | UACUAAC | 86 |
| 46 | YDR139C | 73 | UACUAAC | 55 |
| 47 | YDR305C | 89 | UACUAAC | 47 |
| 48 | YDR367W | 101 | UACUAAC | 65 |
| 49 | YDR381C-A | 194 | AACUAAC | 43 |
| 50 | YDR381W | 766 | GACUAAC | 740 |
| 51 | YDR397C | 92 | UACUAAC | 59 |
| 52 | YDR447C | 314 | UACUAAC | 271 |
| 53 | YDR450W | 435 | UACUAAC | 384 |
| 54 | YDR471W | 384 | UACUAAC | 357 |
| 55 | YDR500C | 389 | UACUAAC | 349 |
| 56 | YEL012W | 123 | CACUAAC | 71 |
| 57 | YER003C | 93 | AACUAAC | 68 |
| 58 | YER007C-A | 103 | UACUAAC | 82 |
| 59 | YER044C-A | 88 | UACUAAC | 38 |
| 60 | YER056C-A | 397 | UACUAAC | 338 |
| 61 | YER074W | 466 | UACUAAC | 418 |
| 62 | YER093C-A | 75 | UACUAAC | 50 |
| 63 | YER102W | 360 | UACUAAC | 331 |
| 64 | YER117W | 471 | UACUAAC | 413 |
| 65 | YER131W | 361 | UACUAAC | 322 |

| | gene name | intron length | branchpoint sequence | branchpoint distance |
|---|---|---|---|---|
| 66 | YER133W | 525 | UACUAAC | 489 |
| 67 | YER179W | 92 | UACUAAC | 58 |
| 68 | YFL034C-A | 321 | UACUAAC | 299 |
| 69 | YFL034C-B | 114 | UACUAAC | 52 |
| 70 | YFL039C | 308 | UACUAAC | 259 |
| 71 | YFR024C-A | 118 | UACUAAC | 65 |
| 72 | YFR031C-A | 147 | UACUAAC | 106 |
| 73 | YFR032C-A | 331 | UACUAAC | 276 |
| 74 | YGL030W | 230 | UACUAAC | 184 |
| 75 | YGL031C | 456 | UACUAAC | 406 |
| 76 | YGL087C | 85 | UACUAAC | 37 |
| 77 | YGL103W | 511 | UACUAAC | 462 |
| 78 | YGL137W | 200 | UACUAAC | 123 |
| 79 | YGL178W | 640 | UACUAAC | 610 |
| 80 | YGL187C | 342 | UACUAAC | 294 |
| 81 | YGL189C | 368 | UACUAAC | 307 |
| 82 | YGL226C-A | 149 | UACUAAC | 97 |
| 83 | YGL232W | 58 | UACUAAC | 36 |
| 84 | YGR027C | 312 | UACUAAC | 274 |
| 85 | YGR029W | 83 | UAUUAAC | 60 |
| 86 | YGR118W | 320 | UACUAAC | 292 |
| 87 | YGR183C | 213 | UACUAAC | 186 |
| 88 | YGR214W | 455 | UACUAAC | 423 |
| 89 | YGR225W | 93 | UACUAAC | 71 |
| 90 | YGR296W | 148 | UACUAAC | 118 |
| 91 | YHL001W | 398 | UACUAAC | 345 |
| 92 | YHL050C | 772 | UACUAAC | 736 |
| 93 | YHR001W-A | 63 | UACUAAC | 43 |
| 94 | YHR010W | 561 | UACUAAC | 525 |
| 95 | YHR012W | 119 | UACUAAC | 83 |
| 96 | YHR016C | 168 | UACUAAC | 122 |
| 97 | YHR021C | 550 | UACUAAC | 489 |
| 98 | YHR039C-A | 162 | UACUAAC | 51 |
| 99 | YHR041C | 101 | UGCUAAC | 60 |
| 100 | YHR077C | 113 | UACUAAC | 58 |
| 101 | YHR097C | 124 | UACUAAC | 64 |
| 102 | YHR101C | 87 | UACUAAC | 46 |
| 103 | YHR123W | 91 | UACUAAC | 53 |
| 104 | YHR141C | 441 | UACUAAC | 389 |
| 105 | YHR203C | 269 | UACUAAC | 222 |

| | gene name | intron length | branchpoint sequence | branchpoint distance |
|---|---|---|---|---|
| 106 | YIL004C | 131 | UACUAAC | 106 |
| 107 | YIL018W | 400 | UACUAAC | 361 |
| 108 | YIL052C | 472 | UACUAAC | 431 |
| 109 | YIL069C | 409 | UACUAAC | 346 |
| 110 | YIL106W | 85 | UAUUAAC | 47 |
| 111 | YIL133C | 290 | UACUAAC | 227 |
| 112 | YIL148W | 434 | UACUAAC | 393 |
| 113 | YIL177C | 388 | UACUAAC | 264 |
| 114 | YJL001W | 116 | UACUAAC | 76 |
| 115 | YJL024C | 77 | UACUAAC | 47 |
| 116 | YJL041W | 118 | UACUAAC | 69 |
| 117 | YJL136C | 460 | UACUAAC | 428 |
| 118 | YJL177W | 317 | UACUAAC | 277 |
| 119 | YJL189W | 386 | UACUAAC | 325 |
| 120 | YJL191W | 408 | UACUAAC | 377 |
| 121 | YJL205C-A | 143 | UACUAAC | 55 |
| 122 | YJL225C | 388 | UACUAAC | 264 |
| 123 | YJR021C | 80 | UACUAAC | 60 |
| 124 | YJR079W | 704 | UACUAAC | 657 |
| 125 | YJR094W-A | 275 | UACUAAC | 236 |
| 126 | YJR145C | 256 | UACUAAC | 168 |
| 127 | YKL002W | 68 | UACUAAC | 40 |
| 128 | YKL006C-A | 141 | UACUAAC | 71 |
| 129 | YKL006W | 398 | UACUAAC | 344 |
| 130 | YKL081W | 326 | UACUAAC | 271 |
| 131 | YKL156W | 350 | UACUAAC | 311 |
| 132 | YKL157W | 383 | UACUAAC | 322 |
| 133 | YKL180W | 306 | UACUAAC | 277 |
| 134 | YKL190W | 76 | UACUAAC | 52 |
| 135 | YKR057W | 322 | AACUAAC | 295 |
| 136 | YKR094C | 368 | UACUAAC | 283 |
| 137 | YLL050C | 179 | UACUAAC | 41 |
| 138 | YLR048W | 359 | UACUAAC | 324 |
| 139 | YLR061W | 389 | UACUAAC | 328 |
| 140 | YLR078C | 89 | UACUAAC | 54 |
| 141 | YLR093C | 141 | GACUAAC | 100 |
| 142 | YLR128W | 94 | UACUAAC | 44 |
| 143 | YLR185W | 359 | UACUAAC | 307 |
| 144 | YLR211C | 59 | GACUAAC | 43 |
| 145 | YLR275W | 90 | UACUAAC | 54 |

| | gene name | intron length | branchpoint sequence | branchpoint distance |
|---|---|---|---|---|
| 146 | YLR287C-A | 430 | UACUAAC | 385 |
| 147 | YLR306W | 134 | UACUAAC | 58 |
| 148 | YLR333C | 423 | CACUAAC | 387 |
| 149 | YLR344W | 447 | UACUAAC | 421 |
| 150 | YLR388W | 488 | UACUAAC | 461 |
| 151 | YLR406C | 349 | UACUAAC | 313 |
| 152 | YLR426W | 71 | UACUAAC | 47 |
| 153 | YLR448W | 384 | UACUAAC | 340 |
| 154 | YLR464W | 279 | UACUAAC | 184 |
| 155 | YML024W | 398 | UACUAAC | 334 |
| 156 | YML025C | 99 | UACUAAC | 69 |
| 157 | YML026C | 401 | UACUAAC | 355 |
| 158 | YML056C | 408 | UACUAAC | 289 |
| 159 | YML067C | 93 | CACUAAC | 65 |
| 160 | YML073C | 415 | UACUAAC | 375 |
| 161 | YML085C | 116 | UACUAAC | 69 |
| 162 | YML094W | 83 | UACUAAC | 47 |
| 163 | YML124C | 298 | UACUAAC | 153 |
| 164 | YMR033W | 86 | UACUAAC | 64 |
| 165 | YMR079W | 156 | UACUAAC | 111 |
| 166 | YMR116C | 273 | UACUAAC | 247 |
| 167 | YMR125W | 322 | UACUAAC | 274 |
| 168 | YMR133W | 116 | UACUAAC | 74 |
| 169 | YMR142C | 402 | UACUAAC | 375 |
| 170 | YMR194W | 463 | UACUAAC | 428 |
| 171 | YMR201C | 84 | UACUAAC | 50 |
| 172 | YMR225C | 147 | UACUAAC | 108 |
| 173 | YMR230W | 410 | UACUAAC | 365 |
| 174 | YMR292W | 82 | UACUAAC | 58 |
| 175 | YNL012W | 84 | AACUAAC | 40 |
| 176 | YNL044W | 79 | UACUAAC | 43 |
| 177 | YNL050C | 91 | UACUAAC | 46 |
| 178 | YNL069C | 449 | UACUAAC | 418 |
| 179 | YNL096C | 345 | UACUAAC | 304 |
| 180 | YNL112W | 1002 | UACUAAC | 953 |
| 181 | YNL147W | 120 | UACUAAC | 56 |
| 182 | YNL162W | 512 | UACUAAC | 470 |
| 183 | YNL246W | 95 | UACUAAC | 55 |
| 184 | YNL265C | 105 | UACUAAC | 59 |
| 185 | YNL301C | 432 | UACUAAC | 386 |

| | gene name | intron length | branchpoint sequence | branchpoint distance |
|---|---|---|---|---|
| 186 | YNL302C | 551 | UACUAAC | 511 |
| 187 | YNL312W | 108 | UACUAAC | 79 |
| 188 | YNL339C | 148 | UACUAAC | 118 |
| 189 | YNR053C | 531 | UACUAAC | 487 |
| 190 | YOL047C | 63 | UACUAAC | 45 |
| 191 | YOL120C | 447 | UACUAAC | 417 |
| 192 | YOL121C | 390 | UACUAAC | 349 |
| 193 | YOL127W | 414 | UACUAAC | 367 |
| 194 | YOR096W | 401 | UACUAAC | 363 |
| 195 | YOR122C | 209 | UACUAAC | 131 |
| 196 | YOR182C | 411 | UACUAAC | 365 |
| 197 | YOR234C | 527 | UACUAAC | 472 |
| 198 | YOR293W | 437 | UACUAAC | 362 |
| 199 | YPL031C | 102 | UACUAAC | 55 |
| 200 | YPL079W | 421 | UACUAAC | 353 |
| 201 | YPL081W | 501 | CACUAAC | 467 |
| 202 | YPL090C | 394 | UACUAAC | 351 |
| 203 | YPL129W | 105 | UACUAAC | 63 |
| 204 | YPL143W | 525 | UACUAAC | 481 |
| 205 | YPL218W | 139 | UACUAAC | 115 |
| 206 | YPL241C | 80 | AAUUAAC | 39 |
| 207 | YPL249C-A | 238 | UACUAAC | 204 |
| 208 | YPL283C | 148 | UACUAAC | 118 |
| 209 | YPR028W | 133 | UACUAAC | 82 |
| 210 | YPR043W | 403 | UACUAAC | 364 |
| 211 | YPR063C | 86 | UACUAAC | 52 |
| 212 | YPR132W | 365 | UACUAAC | 299 |
| 213 | YPR187W | 76 | UACUAAC | 50 |
| 214 | YPR202W | 148 | UACUAAC | 118 |

# Appendix B

# Experimental procedure for testing the effects of intron mutations on splicing in RP51B and RPS6B genes

In order to test our hypothesis that secondary structure within the introns of *S. cerevisiae* genes can affect splicing and consequently the mRNA expression levels, we have conducted a series of laboratory experiments. The basic idea behind the experiments is to check for the difference in gene expression level between the wild type gene and the mutants carrying mutations within introns that, according to our hypothesis, would cause the intronic pre-mRNA to fold unfavorably for splicing. Since direct RNA manipulation is difficult since RNA molecules are very susceptible to degradation, we have decided to test the expression levels of corresponding proteins instead of mRNA expression levels.

Unlike the experiments described in Libri et al. (1995) and Charpentier and Rosbash (1996), where the RP51B intron was inserted into the CUP1 reporter gene, we performed all of our experiment within endogenous genes. This technique allows us to be more certain that the secondary structure formed within introns is equivalent to that formed in nature, and is not modified by interactions with flanking reporter gene sequences.

The yeast strain used for our experiments is a *trp1* auxotrophic mutant, which means that it has the TRP1 gene deleted, making the cells incapable of producing tryptophan, an essential amino acid. These yeast cells will

grow only in the presence ofa tryptophan-enriched medium. This will allow us to use TRP1 as a selectable marker for protein tagging. Protein tagging involves inserting an epitope tag, which is usually a peptide sequence that can be targeted by antibodies, at the 3' of our gene of interest (C-terminal fusion). In this way, the tagged protein will be recognized by labeled antibodies that attach themselves to the epitope tag, enabling its detection and measurement of expression level. The epitope tag is inserted into the genome along with a functional TRP1 gene, thus allowing for selection of cells carrying the construct using the tryptophan-depleted medium.

The gene we tested, RP51B, was tagged at its genomic locus with the 13Myc epitope to generate C-terminal fusion. Myc refers to the peptide sequence AEEQKLISEEDLL and can be targeted by the MYC antibody (commercially available from Covance Research Products). The epitope is called 13Myc because the tag is designed with 13 repeats of the mentioned peptide sequence, enabling easier detection by antibody molecules. The insertion of the 13Myc::TRP1 construct is accomplished using a transformation process, where the target yeast cells are specially treated to be able to take up foreign DNA. Heat shock or electroporation are then applied so that the exogenous DNA is absorbed by the cells, where it gets incorporated into the genome by recombination events.

Successful transformants are selected in the tryptophan-depleted medium (only the cells that have taken up the construct will have a functional TRP1) and further prepared for Western blotting. Western blot analysis enables detection of a specific protein and determination of its size and relative amount. The first step in the procedure is separation of cell proteins by size on a polyacrylamide gel using electrophoresis, which are then transfered to a nitrocellulose membrane that will be imprinted with the same protein bands as the gel. An antibody is then added to the membrane that is able to bind to its specific protein. The antibody has an enzyme or a dye attached to it that is used for visualization of the target protein.

The performed Western blot analysis confirmed the expression of the correct-size-protein product for the wild type RP51B gene. This result shows that tagging of the rp51b protein did not inhibit its expression. The

observed protein level will be compared with the levels of protein after intron mutations.

The next phase of the experimental study was to generate designed intron mutants and test the effect of these mutations on intron splicing as reflected in the final protein expression levels. Starting with the *trp1* and *ura3* auxotrophic mutant strain where the gene of interest was tagged with the 13Myc::TRP1 construct, we first deleted the intron through homologous recombination with the URA3 selectable marker. (The construct carrying the URA3 gene also has 40bp flanking sequences that are homologous to the 5' and 3' ends of sequence we wish to delete, which allows for recombination of the URA3 construct and genomic DNA.) Next, primers containing specific mutated intron sequences were stitched together with sequence segments homologous to the 5' and 3' ends of URA3 insertion using PCR. Transformation of these construct into the appropriate intron deletion strain resulted in recombination, leading to removal of the URA3 gene and insertion of the mutant intron sequence.

The URA3 gene product leads to cell death when placed on 5-floroorotic acid (5-FOA) due to the conversion of 5-FOA to fluorodeoxyuridine, which is toxic to cells. After transformation, cells that have taken up mutated intron inserts and thus have lost the URA3 gene can be selected on 5-FOA. Cells where transformation was not successful and that still contain the URA3 gene will die.

The selected cells were further subjected to PCR analysis to determine if 5-FOA resistance is a result of insertion of the mutated intron in place of the URA3 gene. URA3 is longer than the intron, therefore when PCR is done on the strain using primers that flank the intron/URA3 region of the DNA, the size of the PCR product will tell us which insertion is present. The larger product size indicates that the URA3 gene is still in, but the strain is 5-FOA resistant for another reason (e.g., mutation inside the URA3 coding region). If the PCR product is smaller, this is an indication that the strain contains the desired insertion. Correct size does not guarantee correct sequence, so to be sure that we need the desired mutations we have to sequence the intron sequence.

Strains containing the correct intron mutations were assayed for the effect these mutations have on pre-mRNA splicing through measuring expression levels of the resulting proteins. Western blotting was performed to quantitate the amount of rp51b-13MYC protein using a MYC antibody. The resulting Western blot signals, which were developed with ECL Plus Western Blotting Detection Reagent (Amersham Bioscience), were quantified using the Storm Imaging System (Amersham Bioscience). This imaging system employs fluorescence imaging technology to capture quantitative data from chemifluorescent blots and is characterized by accurate signal quantitation (10-100 times more sensitive than film) and significantly shorter exposure time (many times faster than film).

A flow chart of our experimental procedure is shown in Figure B.1.

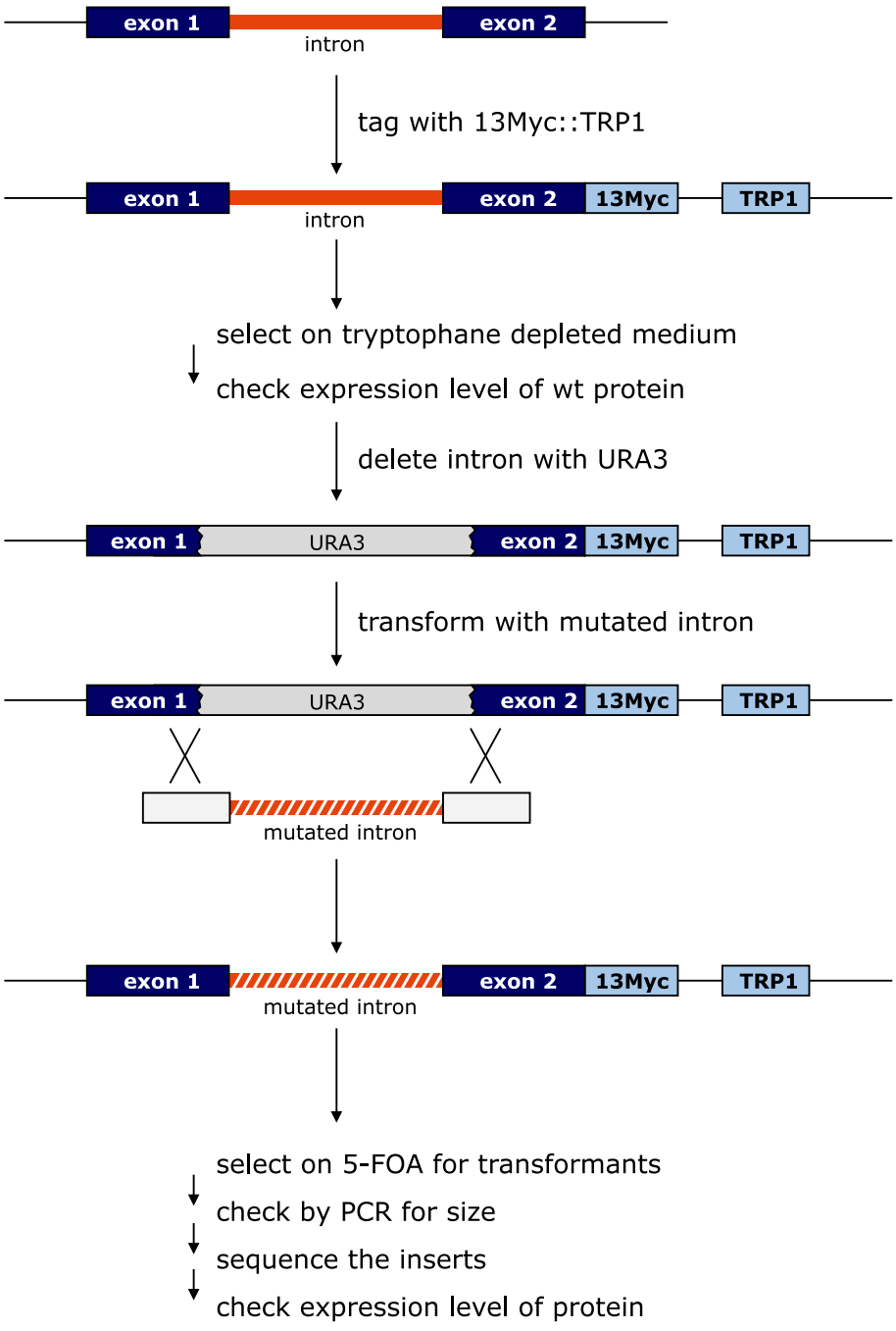



Figure B.1: Flow chart of the experimental procedure.

# Appendix C

# Outputs from StructureAnalyze for RP51B mutants

The post-processed output from the StructureAnalyze procedure is given in the form of a table where each row corresponds to one predicted structure and columns are different measures computed for this structure. The first column is shortened branchpoint distance ($\bar{d}_{ij}$), the second column is normalized (relative) probability ($P_{norm}(R_{ij})$), the third column is is the number of free bases in the branchpoint structure and the fourth column is $b\_weight_{ij}$ as defined in Equation 5.3 (p. 111). The output is divided in sections that correspond to individual input sequences. Each section starts with the name of the sequence and ends with the summary statistics.

## C.1 Output for the Libri's mutants (introns only)

```
wt
  5       0.63       4       0.63
 27       0.20       5       0.20
  5       0.07       4       0.07
 41       0.06       5       0.06
 42       0.04       5       0.04
  5       0.00       4       0.00
  5       0.00       4       0.00
 43       0.00       1       0.00
  5       0.00       4       0.00
-----------------------------------------------------------------
```

```
Average distance: 19.78   r_weight: 0.1494   r_b_weight: 0.1494
-------------------------------------------------------------
3mUB1
 41       0.85       5       0.85
 20       0.12       1       0.58
  5       0.02       4       0.02
 35       0.00       1       0.00
 35       0.00       1       0.00
 20       0.01       1       0.05
 42       0.00       1       0.00
-------------------------------------------------------------
Average distance: 28.29   r_weight: 0.0319   r_b_weight: 0.0573
-------------------------------------------------------------
5mUB1
 41       0.88       5       0.88
 20       0.10       1       0.51
 20       0.00       1       0.01
  5       0.02       4       0.02
 48       0.00       1       0.00
 41       0.00       5       0.00
-------------------------------------------------------------
Average distance: 29.17   r_weight: 0.0302   r_b_weight: 0.0509
-------------------------------------------------------------
8mUB1
 41       0.52       5       0.52
 41       0.44       5       0.44
  5       0.01       4       0.01
 41       0.01       5       0.01
  5       0.01       4       0.01
 44       0.01       5       0.01
  5       0.00       4       0.00
-------------------------------------------------------------
Average distance: 26.00   r_weight: 0.0283   r_b_weight: 0.0283
-------------------------------------------------------------
3mDB1
 41       1.00       5       1.00
 41       0.00       5       0.00
-------------------------------------------------------------
Average distance: 41.00   r_weight: 0.0244   r_b_weight: 0.0244
-------------------------------------------------------------
```

```
5mDB1
 43       0.02       1       0.12
 38       0.80       5       0.80
 43       0.02       1       0.10
 20       0.15       1       0.73
 48       0.00       1       0.01
 48       0.00       1       0.00
 48       0.01       5       0.01
-----------------------------------------------------------------
Average distance: 41.14   r_weight: 0.0296   r_b_weight: 0.0631
-----------------------------------------------------------------
3mUB1/3mDB1
 41       0.99       5       0.99
 48       0.01       5       0.01
 41       0.00       5       0.00
-----------------------------------------------------------------
Average distance: 43.33   r_weight: 0.0243   r_b_weight: 0.0243
-----------------------------------------------------------------
5mUB1/5mDB1
 38       0.52       5       0.52
 20       0.42       1       2.09
 44       0.03       5       0.03
 44       0.03       5       0.03
 20       0.01       1       0.03
 48       0.00       1       0.01
  5       0.00       4       0.00
 34       0.00       1       0.00
 35       0.00       1       0.00
 20       0.00       1       0.02
 48       0.00       1       0.00
 38       0.00       5       0.00
 47       0.00       5       0.00
-----------------------------------------------------------------
Average distance: 33.92   r_weight: 0.0364   r_b_weight: 0.1218
-----------------------------------------------------------------
6mUB1
 48       1.00       1       5.00
-----------------------------------------------------------------
Average distance: 48.00   r_weight: 0.0208   r_b_weight: 0.1042
-----------------------------------------------------------------
```

```
4mUB1
 20       0.89       1       4.43
 46       0.01       1       0.05
 20       0.09       1       0.46
 43       0.00       1       0.00
 48       0.00       1       0.00
 46       0.00       1       0.00
 41       0.01       5       0.01
---------------------------------------------------------------
Average distance: 37.71   r_weight: 0.0494   r_b_weight: 0.2458
---------------------------------------------------------------
```

## C.2 Output for the Libri's mutants (introns and 5' flanking region)

```
wt
 49       0.29       5       0.29
 49       0.29       5       0.29
 41       0.17       6       0.17
 40       0.11       5       0.11
 50       0.10       1       0.48
 34       0.00       1       0.00
 58       0.02       1       0.11
 62       0.02       1       0.08
 49       0.00       5       0.00
 61       0.01       5       0.01
 49       0.00       5       0.00
 49       0.00       5       0.00
 49       0.00       5       0.00
 49       0.00       5       0.00
---------------------------------------------------------------
Average distance: 49.21   r_weight: 0.0214   r_b_weight: 0.0317
---------------------------------------------------------------
3mUB1
 50       0.45       1       2.27
 41       0.20       1       1.01
```

```
 56       0.12       1       0.62
 58       0.11       1       0.53
 61       0.09       5       0.09
 50       0.02       1       0.10
 46       0.00       1       0.00
 39       0.00       1       0.00
 33       0.00       1       0.00
 65       0.00       1       0.00
-----------------------------------------------------------------
Average distance: 49.90   r_weight: 0.0200   r_b_weight: 0.0939
-----------------------------------------------------------------
5mUB1
 41       0.59       1       2.96
 61       0.20       5       0.20
 56       0.11       1       0.54
 54       0.08       1       0.42
 46       0.02       1       0.08
 85       0.00       5       0.00
 85       0.00       5       0.00
 33       0.00       1       0.00
 85       0.00       1       0.00
-----------------------------------------------------------------
Average distance: 60.67   r_weight: 0.0216   r_b_weight: 0.0949
-----------------------------------------------------------------
8mUB1
 61       0.33       5       0.33
 61       0.28       5       0.28
107       0.01       5       0.01
 50       0.25       1       1.25
 28       0.00       6       0.00
 61       0.06       5       0.06
 58       0.06       1       0.29
 61       0.00       5       0.00
-----------------------------------------------------------------
Average distance: 60.88   r_weight: 0.0173   r_b_weight: 0.0414
-----------------------------------------------------------------
3mDB1
 50       0.41       1       2.05
 41       0.18       1       0.91
 34       0.00       1       0.02
```

```
 83       0.14       5       0.14
 34       0.00       1       0.02
 90       0.10       5       0.10
 58       0.10       1       0.48
 61       0.04       5       0.04
108       0.00       5       0.00
 50       0.02       1       0.09
101       0.00       5       0.00
 61       0.01       5       0.01
106       0.00       5       0.00
 44       0.00       1       0.00
 84       0.00       5       0.00
 80       0.00       1       0.00
-------------------------------------------------------------
Average distance: 67.81   r_weight: 0.0184   r_b_weight: 0.0781
-------------------------------------------------------------
5mDB1
 50       0.52       1       2.59
 41       0.23       1       1.15
 34       0.01       1       0.03
 34       0.00       1       0.02
 58       0.12       1       0.60
 62       0.09       1       0.44
 50       0.02       1       0.12
 44       0.00       1       0.00
 83       0.01       5       0.01
 52       0.00       1       0.00
 56       0.00       1       0.01
108       0.00       5       0.00
 80       0.00       1       0.00
 61       0.00       5       0.00
-------------------------------------------------------------
Average distance: 58.07   r_weight: 0.0204   r_b_weight: 0.1014
-------------------------------------------------------------
3mUB1/3mDB1
 50       0.47       1       2.36
 41       0.21       1       1.05
 84       0.18       5       0.18
 58       0.11       1       0.55
 50       0.02       1       0.11
```

```
 58       0.00       5       0.00
 46       0.00       1       0.00
 37       0.00       1       0.00
108       0.00       5       0.00
 33       0.00       1       0.00
 65       0.00       1       0.00
 84       0.00       5       0.00
-------------------------------------------------------------
Average distance: 59.50   r_weight: 0.0191   r_b_weight: 0.0868
-------------------------------------------------------------
5mUB1/5mDB1
 41       0.72       1       3.58
 85       0.01       5       0.01
 85       0.01       5       0.01
 56       0.13       1       0.65
 54       0.10       1       0.51
 46       0.02       1       0.10
 33       0.00       1       0.00
 85       0.00       1       0.00
 84       0.01       5       0.01
 69       0.00       1       0.00
-------------------------------------------------------------
Average distance: 63.80   r_weight: 0.0225   r_b_weight: 0.1112
-------------------------------------------------------------
6mUB1
 50       0.68       1       3.40
 58       0.16       1       0.79
 41       0.11       1       0.57
 50       0.03       1       0.16
 37       0.00       1       0.00
 61       0.01       1       0.05
 49       0.00       1       0.00
 69       0.00       1       0.02
 42       0.00       1       0.00
 78       0.00       5       0.00
-------------------------------------------------------------
Average distance: 53.50   r_weight: 0.0200   r_b_weight: 0.0998
-------------------------------------------------------------
4mUB1
 69       0.40       1       1.98
```

```
 50       0.34       1       1.68
 78       0.11       5       0.11
 58       0.08       1       0.39
 55       0.00       1       0.01
 41       0.06       1       0.28
 50       0.02       1       0.08
 55       0.00       1       0.00
 56       0.00       1       0.01
 33       0.00       1       0.00
 65       0.00       1       0.00
-----------------------------------------------------------------
Average distance: 55.45   r_weight: 0.0170   r_b_weight: 0.0795
-----------------------------------------------------------------
```

## C.3 Output for the Libri's mutants (introns and both flanking regions)

```
wt
 79       0.05       1       0.23
 79       0.86       1       4.30
 84       0.01       1       0.03
 83       0.00       1       0.01
 79       0.00       1       0.01
 97       0.08       1       0.38
 81       0.00       1       0.00
 85       0.01       1       0.03
-----------------------------------------------------------------
Average distance: 83.38   r_weight: 0.0125   r_b_weight: 0.0623
-----------------------------------------------------------------
3mUB1
 85       0.03       1       0.15
 85       0.54       1       2.72
 99       0.39       1       1.96
 83       0.00       1       0.01
 85       0.00       1       0.01
 87       0.03       1       0.15
```

```
-----------------------------------------------------------------
Average distance: 87.33   r_weight: 0.0111   r_b_weight: 0.0555
-----------------------------------------------------------------
5mUB1
 84       0.04       1       0.19
 84       0.71       1       3.54
 98       0.23       1       1.14
 75       0.01       1       0.03
 84       0.00       1       0.01
 86       0.02       1       0.09
 60       0.00       2       0.00
 41       0.00       1       0.01
-----------------------------------------------------------------
Average distance: 76.50   r_weight: 0.0115   r_b_weight: 0.0577
-----------------------------------------------------------------
8mUB1
 84       0.04       1       0.21
 84       0.77       1       3.83
 61       0.02       5       0.02
 61       0.01       5       0.01
 84       0.01       1       0.05
 98       0.12       1       0.58
107       0.00       5       0.00
 84       0.00       1       0.01
 60       0.00       2       0.00
 82       0.00       4       0.00
 61       0.03       1       0.14
 28       0.00       6       0.00
 61       0.00       5       0.00
-----------------------------------------------------------------
Average distance: 73.46   r_weight: 0.0120   r_b_weight: 0.0576
-----------------------------------------------------------------
3mDB1
 79       0.04       1       0.22
 79       0.81       1       4.05
 84       0.01       1       0.03
 83       0.00       1       0.01
 97       0.07       1       0.36
 79       0.00       1       0.01
110       0.03       5       0.03
```

```
 78       0.02       5       0.02
 81       0.00       1       0.00
 85       0.01       1       0.03
 79       0.00       5       0.00
 90       0.00       5       0.00
 79       0.00       5       0.00
-------------------------------------------------------------
Average distance: 84.85   r_weight: 0.0124   r_b_weight: 0.0593
-------------------------------------------------------------
5mDB1
 79       0.05       1       0.23
 79       0.86       1       4.30
 84       0.01       1       0.03
 83       0.00       1       0.01
 79       0.00       1       0.01
 97       0.08       1       0.38
 81       0.00       1       0.00
 85       0.01       1       0.03
110       0.00       5       0.00
 79       0.00       5       0.00
-------------------------------------------------------------
Average distance: 85.60   r_weight: 0.0125   r_b_weight: 0.0623
-------------------------------------------------------------
3mUB1/3mDB1
 85       0.02       1       0.12
 85       0.46       1       2.29
 99       0.33       1       1.66
112       0.15       5       0.15
 83       0.00       1       0.01
 85       0.00       1       0.01
 87       0.03       1       0.13
 90       0.00       5       0.00
 84       0.00       5       0.00
 85       0.00       5       0.00
 62       0.00       5       0.00
-------------------------------------------------------------
Average distance: 87.00   r_weight: 0.0108   r_b_weight: 0.0483
-------------------------------------------------------------
5mUB1/5mDB1
 84       0.04       1       0.19
```

```
 84       0.70       1       3.52
 98       0.23       1       1.13
 75       0.01       1       0.03
 84       0.00       1       0.01
 86       0.02       1       0.09
111       0.00       5       0.00
 60       0.00       2       0.00
111       0.00       5       0.00
 41       0.00       1       0.01
 86       0.00       5       0.00
 84       0.00       5       0.00
-----------------------------------------------------------------
Average distance: 83.67   r_weight: 0.0115   r_b_weight: 0.0574
-----------------------------------------------------------------
6mUB1
 74       0.04       1       0.21
 74       0.79       1       3.96
 60       0.00       2       0.01
 61       0.08       1       0.42
 74       0.00       1       0.01
 50       0.04       1       0.18
 58       0.04       1       0.18
 74       0.00       1       0.00
 84       0.00       1       0.00
 88       0.00       1       0.01
 87       0.00       1       0.00
 49       0.00       1       0.00
 74       0.00       1       0.00
-----------------------------------------------------------------
Average distance: 69.77   r_weight: 0.0141   r_b_weight: 0.0707
-----------------------------------------------------------------
4mUB1
 82       0.05       1       0.26
 82       0.94       1       4.71
 82       0.00       1       0.01
 82       0.00       1       0.00
 84       0.00       1       0.00
 60       0.00       2       0.00
 61       0.00       1       0.01
 78       0.00       1       0.01
```

```
------------------------------------------------------------
Average distance: 76.38   r_weight: 0.0122   r_b_weight: 0.0610
------------------------------------------------------------
```

## C.4 Output for the Charpentier's mutants (introns only)

```
UB1i
 43       0.13       1       0.63
 20       0.83       1       4.14
 41       0.02       5       0.02
 48       0.01       1       0.06
 41       0.01       5       0.01
 33       0.00       1       0.02
------------------------------------------------------------
Average distance: 37.67   r_weight: 0.0454   r_b_weight: 0.2240
------------------------------------------------------------
DB1i
 39       0.01       5       0.01
 39       0.01       5       0.01
 20       0.23       1       1.14
 42       0.00       1       0.02
 25       0.56       5       0.56
 42       0.09       5       0.09
  5       0.01       4       0.01
 37       0.01       5       0.01
 17       0.01       1       0.05
 20       0.04       1       0.18
 48       0.01       7       0.01
 20       0.00       1       0.00
 39       0.00       5       0.00
------------------------------------------------------------
Average distance: 30.23   r_weight: 0.0424   r_b_weight: 0.0978
------------------------------------------------------------
UB1iDB1i
  5       0.67       4       0.67
 27       0.22       5       0.22
```

```
  5       0.07       4       0.07
 44       0.04       5       0.04
  5       0.00       4       0.00
 17       0.00       1       0.02
  5       0.00       4       0.00
  5       0.00       4       0.00
  5       0.00       4       0.00
-------------------------------------------------------------
Average distance: 13.11   r_weight: 0.1569   r_b_weight: 0.1578
-------------------------------------------------------------
mut5
 20       0.96       1       4.82
 42       0.02       1       0.09
 43       0.02       1       0.08
 20       0.00       1       0.01
 48       0.00       1       0.00
-------------------------------------------------------------
Average distance: 34.60   r_weight: 0.0491   r_b_weight: 0.2455
-------------------------------------------------------------
mut12
  5       0.67       4       0.67
 27       0.21       5       0.21
  5       0.07       4       0.07
 42       0.04       5       0.04
  5       0.00       4       0.00
  5       0.00       4       0.00
  5       0.00       4       0.00
 17       0.00       1       0.02
  5       0.00       4       0.00
-------------------------------------------------------------
Average distance: 12.89   r_weight: 0.1569   r_b_weight: 0.1578
-------------------------------------------------------------
mut18
  5       0.67       4       0.67
 27       0.22       5       0.22
  5       0.07       4       0.07
 44       0.04       5       0.04
  5       0.00       4       0.00
 17       0.00       1       0.02
  5       0.00       4       0.00
```

```
  5       0.00       4       0.00
  5       0.00       4       0.00
-------------------------------------------------------------
Average distance: 13.11   r_weight: 0.1569   r_b_weight: 0.1578
-------------------------------------------------------------
```

## C.5 Output for the Charpentier's mutants (introns and 5' flanking region)

```
UB1i
 56       0.55       1       2.75
 50       0.21       1       1.04
 74       0.16       5       0.16
 58       0.05       1       0.24
 46       0.00       1       0.01
 41       0.03       1       0.13
 50       0.01       1       0.05
 46       0.00       1       0.00
 50       0.00       1       0.01
-------------------------------------------------------------
Average distance: 52.33   r_weight: 0.0178   r_b_weight: 0.0805
-------------------------------------------------------------
DB1i
 41       0.61       6       0.61
 41       0.09       6       0.09
 50       0.08       1       0.42
 47       0.06       6       0.06
 41       0.04       1       0.19
 57       0.11       5       0.11
 61       0.00       5       0.00
 61       0.00       5       0.00
 70       0.00       5       0.00
 80       0.00       7       0.00
 64       0.00       6       0.00
 82       0.00       5       0.00
 57       0.00       7       0.00
```

```
 28       0.00       6       0.00
 50       0.00       1       0.00
 65       0.00       7       0.00
109       0.00       5       0.00
 43       0.00       6       0.00
 61       0.00       5       0.00
-----------------------------------------------------------------
Average distance: 58.32   r_weight: 0.0230   r_b_weight: 0.0334
-----------------------------------------------------------------
UB1iDB1i
 49       0.33       5       0.33
 49       0.33       5       0.33
 41       0.19       6       0.19
 40       0.12       5       0.12
 37       0.02       5       0.02
 49       0.00       5       0.00
 49       0.00       5       0.00
 49       0.00       5       0.00
 49       0.00       5       0.00
  4       0.00       1       0.00
 49       0.00       5       0.00
-----------------------------------------------------------------
Average distance: 42.27   r_weight: 0.0220   r_b_weight: 0.0222
-----------------------------------------------------------------
mut5
 50       0.41       1       2.07
 62       0.30       1       1.50
 41       0.18       1       0.92
 58       0.10       1       0.48
 46       0.00       1       0.01
 46       0.00       1       0.00
 52       0.00       1       0.00
 80       0.00       1       0.00
 84       0.00       1       0.00
 65       0.00       1       0.00
 56       0.00       1       0.01
 63       0.00       1       0.00
 52       0.00       1       0.00
-----------------------------------------------------------------
Average distance: 58.08   r_weight: 0.0194   r_b_weight: 0.0969
```

```
------------------------------------------------------------
mut12
 49       0.32       5       0.32
 49       0.32       5       0.32
 41       0.19       6       0.19
 40       0.12       5       0.12
 49       0.03       5       0.03
 37       0.02       5       0.02
 49       0.00       5       0.00
 49       0.00       5       0.00
 49       0.00       5       0.00
 49       0.00       5       0.00
 49       0.00       5       0.00
------------------------------------------------------------
Average distance: 46.36   r_weight: 0.0218   r_b_weight: 0.0218
------------------------------------------------------------
mut18
 49       0.33       5       0.33
 49       0.33       5       0.33
 41       0.19       6       0.19
 40       0.12       5       0.12
 37       0.02       5       0.02
 49       0.00       5       0.00
 49       0.00       5       0.00
 49       0.00       5       0.00
 49       0.00       5       0.00
  4       0.00       1       0.00
 49       0.00       5       0.00
------------------------------------------------------------
Average distance: 42.27   r_weight: 0.0220   r_b_weight: 0.0222
------------------------------------------------------------
```

## C.6 Output for the Charpentier's mutants (introns and both flanking regions)

```
UB1i
 75       0.05       1       0.25
```

```
 83       0.03       1       0.13
 75       0.91       1       4.55
 75       0.00       1       0.01
 56       0.00       1       0.02
 60       0.00       2       0.00
 61       0.00       1       0.02
 89       0.00       1       0.01
 74       0.00       5       0.00
-----------------------------------------------------------------
Average distance: 72.00   r_weight: 0.0133   r_b_weight: 0.0666
-----------------------------------------------------------------
DB1i
 79       0.19       7       0.19
 79       0.69       4       0.69
 79       0.10       4       0.10
 79       0.00       1       0.01
 83       0.01       7       0.01
 79       0.00       1       0.00
 97       0.00       1       0.01
 79       0.00       5       0.00
 79       0.00       7       0.00
 85       0.00       1       0.00
-----------------------------------------------------------------
Average distance: 81.80   r_weight: 0.0126   r_b_weight: 0.0128
-----------------------------------------------------------------
UB1iDB1i
 49       0.60       5       0.60
 50       0.37       4       0.37
 49       0.00       5       0.00
 49       0.03       5       0.03
 41       0.01       6       0.01
 49       0.00       5       0.00
 49       0.00       5       0.00
  4       0.00       1       0.00
-----------------------------------------------------------------
Average distance: 42.50   r_weight: 0.0204   r_b_weight: 0.0207
-----------------------------------------------------------------
mut5
 79       0.05       1       0.23
 79       0.87       1       4.33
```

```
 83       0.00       1       0.01
 79       0.00       1       0.01
 97       0.08       1       0.38
 85       0.01       1       0.03
 79       0.00       1       0.00
-------------------------------------------------------------
Average distance: 83.00   r_weight: 0.0125   r_b_weight: 0.0624
-------------------------------------------------------------
mut12
 49       0.42       5       0.42
 83       0.01       1       0.04
 50       0.26       3       1.29
 83       0.16       1       0.81
 97       0.12       1       0.58
 49       0.00       5       0.00
 49       0.02       5       0.02
 83       0.00       1       0.00
 49       0.00       5       0.00
 85       0.01       1       0.05
 41       0.01       6       0.01
 60       0.00       2       0.00
-------------------------------------------------------------
Average distance: 64.83   r_weight: 0.0176   r_b_weight: 0.0518
-------------------------------------------------------------
mut18
 49       0.60       5       0.60
 50       0.37       4       0.37
 49       0.00       5       0.00
 49       0.03       5       0.03
 41       0.01       6       0.01
 49       0.00       5       0.00
 49       0.00       5       0.00
  4       0.00       1       0.00
-------------------------------------------------------------
Average distance: 42.50   r_weight: 0.0204   r_b_weight: 0.0207
-------------------------------------------------------------
```

# Appendix D

# Sequences of new RP51B mutants

The following is the FASTA format of the mutated RP51B intron sequences that we designed in order to test the accuracy of the hypothesis that short branchpoint distances are required for efficient splicing of yeast intron. The mutated sequences are represented by lower case letters.

>bad1
GUACGUACCACGAGAUGUUGAUGAAGCCGGAUAUGAUGGACUGGGCGCUGAACACAUGAA
AUGAGGGCAAGGUUUGCAGAGAGAUUGAAAGCGUUAUGGGAACGAGGGGACCAGCAGGGC
AUUCUUAUUUAUGAGCAGAUUAGAAAACUCCAUUACUGAUUAGUUUAGAAGAGCGCUCAA
UGAAGUAGUAGAUAUUUAAAAGAUCACCAAAUAACCAAUUGCUUUCGAAUGGCACAUUCU
AUCUUAUCCAAUGGUCUaugagacaacuAUUUACUAACUUAAGUUGUCUCAUUUGAUUAU
UGCUAUUUUUAUAG
>bad2
GUACGUACCACGAGAUGUUGAUGAAGCCGGAUAUGAUGGACUGGGCGCUGAACACAUGAA
AUGAGGGCAAGGUUUGCAGAGAGAUUGAAAGCGUUAUGGGAACGAGGGGACCAGCAGGGC
AUUCUUAUUUAUGAGCAGAUUAGAAAACUCCAUUACUGAUUAGUUUAGAAGAGCGCUCAA
UGAAGUAGUAGAUAUUUAAAAGAUCACCAAAUAACCAAUuaaucaaugagacaacuuaaU
AUCUUAUCCAAUGGUCUUGAAGAGAGGUAUUUACUAACUUAAGUUGUCUCAUUUGAUUAU
UGCUAUUUUUAUAG
>bad3
GUACGUACCACGAGAUGUUGAUGAAGCCGGAUAUGAUGGACUGGGCGCUGAACACAUGAA
AUGAGGGCAAGGUUUGCAGAGAGAUUGAAAGCGUUAUGGGAACGAGGGGACCAGCAGGGC
AUUCUUAUUUAUGAGCAcuuaaguuaguaaaauaccucGAUUAGUUUAGAAGAGCGCUCAA

UGAAGUAGUAGAUAUUUAAAAGAUCACCAAAUAACCAAUUGCUUUCGAAUGGCACAUUCU
AUCUUAUCCAAUGGUCUUGAAGAGAGGUAUUUACUAACUUAAGUUGUCUCAUUUGAUUAU
UGCUAUUUUUAUAG
>bad4
GUACGUACCACGAGAUGUUGAuuaAGCCGcuacuacuUGGACUGucgGCUGAACACAUGA
AAUGAGGGCAAGGUUUGCAGAGAGAUUGAAAGCGUUAUGGGAACGAGGGGACCAGCAGGG
CAUUCUUAUUUAUGAGCAGAUUAGAAAACUCCAUUACUGAUUAGUUUAGAAGAGCGCUCA
AUGAAGUAGUAGAUAUUUAAAAGAUCACCAAAUAACCAAUUGCUUUCGAAUGGCACAUUC
UAUCUUAUCCAAUGGUCUUGAAGAGAGGUAUUUACUAACUUAAGUUGUCUCAUUUGAUUA
UUGCUAUUUUUAUAG
>bad5
GUACGUACCACGAGAUGUUGAUGAAGCCGGAUAUGAUGGACUGGGCGCUGAACACAUGAA
AUGAGGGCAAGGUUUGCAGAGAGAUUGAAAGCGUUAUGGGAACGAGGGGACCAGCAGGGC
AUUCUUAUUUAUGAGCAGAUUAGaacaucucgugguaGAUUAGUUUAGAAGAGCGCUCAA
UGAAGUAGUAGAUAUUUAAAAGAUCACCAAAUAACCAAUUGCUUUCGAAUGGCACAUUCU
AUCUUAUCCAAUGGUCUUGAAGAGAGGUAUUUACUAACUUAAGUUGUCUCAUUUGAUUAU
UGCUAUUUUUAUAG
>good1
GUACGUACCucucuucaagGAUGAAGCCGGAUAUGAUGGACUGGGCGCUGAACACAUGAA
AUGAGGGCAAGGUUUGCAGAGAGAUUGAAAGCGUUAUGGGAACGAGGGGACCAGCAGGGC
AUUCUUAUUUAUGAGCAGAUUAGAAAACUCCAUUACUGAUUAGUUUAGAAGAGCGCUCAA
UGAAGUAGUAGAUAUUUAAAAGAUCACCAAAUAACCAAUUGCUUUCGAAUGGCACAUUCU
AUCUUAUCCAAUGGUCUUGAAGAGAGGUAUUUACUAACUUAAGUUGUCUCAUUUGAUUAU
UGCUAUUUUUAUAG
>good2
GUACGUACCACGAGAUGUUGAUGAAGCCGGAUAUGAUGGACUGGGCGCUGAACACAUGAA
AUGAGGGCAAGGUUUGCAGAGAGAUUGAAAGCGUUAUGGGAACGAGGGGACCAGCAGGGC
AUUCUUAUUUAUGAGCAGAUUAGAAAACUCCAUUACUGAUUAGUUUAGAAGAGCGCUCAA
UGAAGUAGUAGAUAUUUAAAAGAUCACCAAAUAACCuuucauguguucagcGCACAUUCU
AUCUUAUCCAAUGGUCUUGAAGAGAGGUAUUUACUAACUUAAGUUGUCUCAUUUGAUUAU
UGCUAUUUUUAUAG
>good3
GUACGUACCACGAGAUGUUGAUGAAGCCGGAUAUGAUGGACUGGGCGCUGAACACAUGAA

AUGAGGGCAAGGUUUGCAGAGAGAUUGAAAGCGUUAUGGGAACGAGGGGACCAGCAGGGC
AUUCUUAUUUAUGAGCAGAUUAGAAAACUCCAUUACUGAUUAGUUUAGAAGAGCGCUCAA
UGAAGUAGUAGAUAUUUAAAAGAUCACCAAAUAACCAAUUGCUUUCGAAUGGCACAUUCU
AUCUUAUCCAAUGGUCUUGAAGAGAGGUAUUUACUAACUuacguacUCUCAUUUGAUUAU
UGCUAUUUUUAUAG

>good4

GUACGUACCACGAGAUGUUGAUGAAGCCGGAUAUGAUGGACUGGGCGCUGAACACAUGAA
AUGAGGGCAAGGUUUGCAGAGAGAUUGAAAGCGUUAUGGGAACGAGGGGACCAGCAGGGC
AUUCUUAUUUAUGAGCAGAUUAGAAAACUCCAUUACUGAUUAGUUUAGAAGAGCGCUCAA
UGAAGUAGUAGAUAUUUAAAAGAUCACCAAAUAACCAAUUGCUUUCGAAUGGCACAUUCU
AUCUUuagCggcuGUCUUGAAGAGAGGUAUUUACUAACUUAAGUUGUCUCAUUUGAUUAU
UGCUAUUUUUAUAG

>good5

GUACGUACCACGAGAUGUUGAUGAAGCCGGAUAUGAUGGACUGGGCGCUGAACACAUGAA
AUGAGGGCAAGGUUUGCAGAGAGAUUGAAAGCGUUAUGGGAACGAGGGGACCAGCAGGGC
AUUCUUAUUUAUGAGCAGAUUAGAAAACUCCAUUACUGAUUAGUUUAGAAGAGCGCUCAA
UGAAGUAGUAGAUAUUUAAAAGAUCACCAAAUAACCAAUUGCUUUCGAAUGGCACAUUCU
AUCcuucaucaacGUCUUGAAGAGAGGUAUUUACUAACUUAAGUUGUCUCAUUUGAUUAU
UGCUAUUUUUAUAG