Pacific Symposium on Biocomputing 2002, Altman, *et al.* eds.,
Jan. 2002, World Scientific, pp 223–234

PRE-mRNA SECONDARY STRUCTURE PREDICTION AIDS SPLICE SITE PREDICTION

DONALD J. PATTERSON, KEN YASUHARA, WALTER L. RUZZO
Computer Science and Engineering
University of Washington, Box 352350
Seattle, WA 98195, USA

Accurate splice site prediction is a critical component of any computational approach to gene prediction in higher organisms. Existing approaches generally use sequence-based models that capture local dependencies among nucleotides in a small window around the splice site. We present evidence that computationally predicted secondary structure of moderate length pre-mRNA subsequences contains information that can be exploited to improve acceptor splice site prediction beyond that possible with conventional sequence-based approaches. Both decision tree and support vector machine classifiers, using folding energy and structure metrics characterizing helix formation near the splice site, achieve a 5–10% reduction in error rate with a human data set. Based on our data, we hypothesize that acceptors preferentially exhibit short helices at the splice site.

1 Introduction

Whole-genome analysis of a single organism or comparison of organisms depends on correct gene annotation. Hence, accurate gene prediction from DNA sequence data is an important goal for bioinformatics, both for purposes of providing "high-throughput annotation" to match high-throughput sequencing, and for the insight it may provide into the underlying biology. Accurate splice site prediction is a critical component of eukaryotic gene prediction. Unfortunately, while current approaches achieve accuracies above 90% with acceptable false negative rates, compounded errors for multi-exonic genes contribute to a substantially higher error rate for full-length gene predictions.[1–3]

Splice site prediction initially depended on very simple models involving consensus sequences in narrow windows around the splice sites.[4] As more data became available, zero-th order Markov models ("Weight Matrix Models" or "Position Specific Scoring Matrices") became possible.[5] With still more data, researchers adopted higher order Markov models ("Weight Array Matrices" or WAMs)[6] and variants such as the "Windowed Weight Array Matrices" of Burge and Karlin,[7] and various kinds of decision trees, such as the "Maximal Dependence Decomposition" model.[7] Despite the increasing sophistication of these models as more training data becomes available, they all basically exploit observed dependencies among *nearby* nucleotides in the vicinity of the splice site.

Much is known about the mechanisms underlying processing of spliceosomal introns.[8,9] In particular, it is known that certain short RNAs (the U1, U2, U4, U5,

U6 snRNAs) hybridize with each other and with complementary segments of the pre-message at the donor, branch point, and acceptor sites. These segments are probably important determinants of the specificity of splicing. The sequence-based models mentioned above are appropriate for characterizing this sequence complementarity. However, it also appears that the information content of these short neighborhoods around the splice sites is not adequate to fully account for the observed high specificity of splicing *in vivo*.

There has long been speculation that secondary structure in pre-mRNAs also plays a role in splicing, and there have been a number of experimentally verified cases where splicing defects have been tied to mutations that alter secondary structure near splice sites.[10–16] However, no clear pattern emerges from these reports, so although secondary structure may play a role in splice site recognition, a single, strongly conserved structure (as found in tRNAs or other functional RNAs) is not expected. Rather, some looser structure or collection of structures might incrementally contribute to the observed specificity of splicing.

For example, it seems plausible that initial hybridization of the spliceosomal snRNAs to the pre-mRNA might be enhanced or inhibited by the presence of short helices in the vicinity of the splice sites, without requiring conservation of a precisely determined structure at an exact position relative to the splice site. This is consistent with observations of Mir and Southern, who examined hybridization of a tRNA to an oligo microarray and reported significant influence of the tRNA's structure on hybridization.[17] In particular, strong hybridization generally seemed to require that the oligo match the entire length of one strand of a helix in the tRNA, together with a few adjacent unpaired bases, and additionally was stabilized by coaxial stacking with another helix.

In this paper, we report positive results from a series of computational experiments designed to discover such correlations between splicing and computationally predicted secondary structure of pre-mRNAs for a sample of human genes. We identified several structure metrics showing subtle but statistically significant correlation to acceptor splice sites (i.e., 3' ends of introns), beyond that already accounted for by a good sequence-based model. Comparable results were obtained with two very different classification methods and hence are unlikely to be simply an artifact of either classifier. Although the net improvement in classifier accuracy was small, approximately a 5–10% reduction in misclassification rate, this could translate into a substantial improvement in the accuracy of full-length gene predictions for genes with 10, 20, or more exons. However, we feel that our most important contribution is not this direct application but rather the evidence that structure does play a role in splicing, and that current structure prediction tools are accurate enough to exploit it. Additionally, structure might play a role in other processes, e.g., ribozyme binding[18] and perhaps mRNA stabilization and degradation. Because computational tools for discovering

informative structural features are much less well-developed than tools for features based on primary sequence, we expect the methods outlined here to be of value in other contexts.

In outline, the methodology we employed is as follows. From a test set of annotated, multi-exon human genes, we extracted acceptor splice sites and a representative sample of nearby non-sites matching the acceptor AG dinucleotide consensus. We used Zuker's MFOLD[19,20] to predict foldings for a 100-base window centered on each site/nonsite. Various sequence and structure features, such as per-position dinucleotide frequencies and pairing probabilities, were extracted for use in our classifiers. Each test, using 10-fold cross-validation, examined the change in accuracy between a baseline, sequence-based model and the same model plus one or more of the structure metrics. Tests were performed with two standard machine learning approaches—C4.5 decision trees,[21,22] and support vector machines.[23–26] Details of our methodology are presented in Section 2.

Our results are described in Section 3. To briefly summarize, we obtained statistically significant accuracy improvements with various combinations of three structure metrics. The first was the simplest: energy of the predicted optimal folding. Sequences containing acceptors on average had slightly more stable structures than nonsites. Second, for each position i in a sequence, we computed a "Max Helix" score, roughly an estimate of the probability of a helix within 5 bases of position i. We observed Max Helix scores to be relatively independent of i for nonsites, whereas acceptors showed a dip in Max Helix score roughly 10 bases upstream of the splice site. For our third and most detailed metric, we determined whether each position of a folded sequence was paired and stacked onto the nucleotide preceding it, paired but unstacked, or unpaired, then built a second order Markov model of the resulting ternary sequences. Again, the profiles of acceptors and non-acceptors tended to differ. For example, it appears that acceptors more often have a short helix at the splice site. In all performance comparisons we included the score from a first order Markov model (which we refer to as a weight array model or WAM) trained on the primary sequence near the acceptor site. Some portion of the structural consensus noted above is probably just a reflection of the acceptor sequence consensus. Nevertheless, in our tests, classifiers using one or more of these structural features in addition to WAM score consistently outperformed classifiers using WAM score alone.

2 Methods

2.1 Data Set

For training and testing, we started with 462 unrelated, annotated, multi-exon genes with standard splicing (i.e., excluding cases of alternative or self-splicing) from a data

set proposed by Reese *et al.* as a benchmark set for evaluating gene-finding software.[27] Using exon annotations, we extracted a 100-base window centered on each acceptor splice site having sufficient flanking sequence. This formed our collection of positive samples. The non-acceptor, negative sample set was a random sampling of 100-base subsequences centered on an AG dinucleotide that were not annotated as acceptors, but were within 100 bases of an actual acceptor. We imposed these criteria to evaluate how our structure-based methods might enhance gene prediction methods, which must discriminate among several putative acceptor splice sites occurring close to each other. We formed a negative sample set of the same size as our positive set, each with 1,980 subsequences, so that the machine learners gave equal weight to false negative and false positive errors.

We randomly partitioned the 3,960 subsequences into 10 equal-sized groups for cross-validated training and testing, with each group containing an equal number of positive and negative samples. The same groups were used for all tests, allowing comparison of results on a per-group basis, as well as averaged over the 10 groups.

2.2 Sequence-based Metric

Weight Array Model (WAM).[6] A first order WAM models a primary sequence pattern by storing, for each base offset, the probability of observing each base conditioned on the previous base. Given WAMs trained on positive and negative example subsequences and an unclassified subsequence, a log likelihood score can be computed that reflects the likelihood that the subsequence contains a splice site. As in Burge's work,[28] we scored sequences using positions -21 to +3 relative to the putative acceptor site. To ensure that overfitting did not occur, we trained each WAM on 9 groups and scored the remaining group with this model. Cross-validated testing with an optimal threshold classifier confirmed that widening this window by 5 positions on either side did not improve accuracy.

2.3 Pre-mRNA Structure Prediction and Structure Metrics

For each subsequence, we used MFOLD to produce a comprehensive set of foldings, typically hundreds in number, each annotated with a free energy. Low free energy is correlated with folding stability and likelihood. The equilibrium partition function can be used to calculate the probability that a folding will occur in nature, given its free energy and the total free energy of all possible foldings.[29] In computing structure metrics from a given subsequence's many predicted foldings, we used these probabilities to weight each folding's contribution to an aggregate score. More probable foldings (i.e., ones with lower free energy) are accordingly weighted.

For each subsequence, we computed the following structure metrics:

Optimal Folding Energy (OFE). Our simplest metric was the free energy of the optimal folding. This number roughly reflects the stability of the fold and typically is lower with more paired bases.

Max Helix (MH). For each position around the putative splice site, we calculated the probability, according to the equilibrium partition function, of a helix starting or ending at that position. To relax the positional specificity of this metric, for each position, we recorded the maximum probability of a helix start/end within a neighborhood of 5 positions up- and downstream.

Neighbor Pairing Correlation Model (NPCM). A folded structure can be summarized by a string over the three symbol alphabet $\{\mathsf{S}, \mathsf{P}, \mathsf{O}\}$, corresponding to whether each position is paired and stacked onto the nucleotide preceding it, paired but unstacked, or unpaired, respectively. The string's length is equal to the length of the original pre-mRNA sequence. For example, a three-base helix flanked by unpaired regions would be represented by . . . OPSSO . . .

Given a set of RNA sequences of equal length, we converted each predicted folding of each sequence into a structure string and a corresponding folding probability. This collection of strings was used to train a second order Markov model, forming an aggregate model of the structure of the collection of sequences.

Although a Markov model can not fully describe the set of structure strings, we believe it can approximate many local features reasonably well. (The extra descriptive power afforded by using a stochastic context-free grammar[30] did not seem warranted at this stage.)

We trained two Markov models as described above—one on acceptor splice site sequences and the other on non-acceptor sequences. We scored the structure string of an unclassified pre-mRNA sequence by computing the posterior probability that each model generated this structure string. We then computed the log of the ratio of the site model probability over the non-site model probability, i.e., the log likelihood ratio.

2.4 Machine Learning Methods

We evaluated our structure metrics by aggregating them into real-valued feature vectors and training two machine learning classifiers, support vector machines (SVMs) and decision trees, on them.

SVMs perform binary classification by partitioning the feature space with a surface implied by a subset of the training vectors near the separating surface called *support vectors*.[24] SVMs are efficient with multi-dimensional data, subsume many other learning methods, and are solidly grounded in statistical theory. (See Hearst *et al.*[26] for a gentle introduction and Burges' tutorial[25] for a more formal, extensive introduction with further references.) In this study, we used Noble's implementation, svm 1.1.[31]

Decision trees are another form of supervised machine learner that classify feature vectors hierarchically. When predicting the class of a vector, a decision tree passes a vector down the tree from the root to a leaf. At each node, the decision tree examines one feature of the vector to determine which branch the vector should recursively travel down. Every leaf on the tree has an associated prediction, which is the classification that is ultimately assigned to the vector.

Decision trees are generated ("trained") by examining a collection of labeled vectors and statistically determining which feature contains the most information relevant to the classification. A node is formed to partition the training vectors into subsets based on this feature. These subsets are independently used to train the next lower level of the tree. When a subset's elements all belong to the same class or the amount of information in the subset is statistically insignificant, a leaf is formed, whose classification is equal to the majority classification of the subset.

We evaluated our feature sets with the C4.5 decision tree software package.[21] C4.5 forms decisions based on axis-parallel hyperplanes, corresponding to threshold tests on one feature at each node of the tree.

2.5 Testing Methodology

In all tests, we used accuracy (fraction of samples classified correctly) as our performance metric; observed false negative and false positive rates were roughly equal. We employed slightly different testing methodologies for decision trees and SVMs.

The decision tree methodology began with cross-validated tests of trees trained on WAM score alone, resulting in 10 per-group accuracies, our baseline. For each set of structure metrics, we then repeated the cross-validated tests, allowing the decision tree to train on WAM score in conjunction with combinations of structure metrics. If we observed a significant increase in accuracy relative to the baseline tests, we concluded that the structure metrics contained useful information that the WAM could not capture.

To avoid potential problems comparing performance of SVMs with different dimensionality, we used a *mixed/matched* methodology that only involved comparing results for models trained on data of the same dimensionality. For each of the 10 cross-validations, we trained two models. The *matched* model was trained on feature vectors that contained WAM score and the structure metrics. The *mixed* model was trained on the same data modified by randomly permuting ("mixing") the structure metrics. That is, for each training vector consisting of a WAM score and at least one structure metric, the structure metric components were replaced with those of another training vector, randomly selected (without replacement) independently of the vector's class. The mixed and matched models were then tested on the reserved test set, and significantly lower accuracy with the mixed model evinced useful information in

Table 1: Results of 10-fold cross-validated decision tree testing with Weight Array Model (WAM) and various structure metric combinations: Optimal Folding Energy (OFE), Neighbor Pairing Correlation Model (NPCM), and Max Helix scores (MH). NPCM was scored on positions -50 through +3, and MH scores are taken for positions -10 and +3 only. Mean accuracy (fraction of samples classified correctly), improvement over baseline WAM accuracy ($\Delta \pm$ one standard deviation) and paired Wilcoxon test p-values are shown.

features	*mean acc. (%)*	Δ	p
WAM (baseline)	92.73	-	-
WAM, OFE	93.13	$+0.40 \pm 0.87$	0.066
WAM, OFE, NPCM	93.16	$+0.43 \pm 0.80$	0.022
WAM, OFE, MH	93.21	$+0.48 \pm 0.90$	0.009
WAM, OFE, NPCM, MH	93.13	$+0.40 \pm 0.84$	0.016

the structure metrics not captured in the WAM score.

With no reason to assume the observed accuracy distributions were Gaussian, we conservatively tested statistical significance of accuracy differences using the paired Wilcoxon signed rank test, a nonparametric analogue of the paired t-test. For the paired Wilcoxon test, the p-value is the probability of obtaining test results as extreme as those we observed, assuming the null hypothesis—that the differences between paired accuracies have median zero.

3 Results

We trained decision trees and radial basis kernel SVMs with many combinations of the structure metrics we formulated. Testing with cross-validation as described in Section 2.5, we identified optimal folding energy (OFE), Max Helix, and Neighbor Pairing Correlation Model (NPCM) scores as useful metrics for acceptor recognition.

Decision tree test results in Table 1 show that training on WAM and OFE with each of the remaining structure metrics yielded statistically significant accuracy improvement, relative to training on WAM score alone. Because overfitting causes decision tree performance to degrade with the addition of features with redundant information, we chose only two Max Helix positions (-10 and +3). Adding the combination of OFE and these Max Helix scores yielded a 7% reduction in classification error rate.

We also saw statistically significant accuracy improvements in mixed/matched SVM testing when Max Helix scores for each position from -20 to +3 were combined with OFE. This result independently supports the decision tree results with Max Helix above. To examine the degree of variability of these results due to the randomized mixing step, we repeated the 10-fold cross-validation runs ten times. For each of these ten runs, Table 2 shows the mean accuracy with the mixed models and by how

Table 2: Results of radial kernel SVM mixed/matched testing with Weight Array Model (WAM) score, optimal folding energy (OFE) and Max Helix (MH) scores for each position from -20 to +3. 10-fold cross-validation runs were repeated ten times. For each run, mean mixed model accuracy, improvement with matched model ($\Delta \pm$ one standard deviation) and paired Wilcoxon test p-values are shown. Accuracy with the matched model was 92.90%.

c.v.	*mean accuracy (%)*		
run	*mixed*	Δ	p
1	91.61	$+1.29 \pm 1.18$	0.006
2	92.15	$+0.76 \pm 0.76$	0.014
3	92.27	$+0.63 \pm 0.62$	0.012
4	92.25	$+0.66 \pm 0.71$	0.014
5	91.84	$+1.06 \pm 0.99$	0.010
6	92.12	$+0.78 \pm 0.73$	0.009
7	91.99	$+0.91 \pm 0.74$	0.009
8	92.22	$+0.68 \pm 0.71$	0.028
9	92.07	$+0.84 \pm 0.47$	0.006
10	92.53	$+0.38 \pm 0.63$	0.072
mean	92.10	+0.80	-

much it differs from the mean accuracy with the matched models. Properly matching WAM score and structure metrics improved accuracy in all ten of the cross-validation runs, with $p < 0.05$ in nine of the ten runs and with improvements exceeding 1% with $p < 0.01$ in two of them. On average, the improvement was 0.8%, approximately a 10% reduction in misclassification rate.

Figure 1 presents three views of the structure metrics we developed. In each graph, the solid and dotted lines are the mean and standard deviation, respectively, of the metric, calculated across all 10 cross-validation groups. The top graph shows the $\log_{10}$ likelihood ratio of a base pairing (either stacked or unstacked) with any other base within the folding window. From this graph it can be observed that there is an approximately 25% smaller chance of a pair forming at position -5 in an acceptor splice site than in a non-site. This effect is reversed in the region from -2 to +1, where acceptors demonstrate a 25% greater chance of pairing.

The middle graph of Figure 1 further investigates this trend by showing the probability of a stacked pair, S, at a given position, conditioned on the previous two positions being OP, i.e., an unpaired base followed by a paired base. This can be interpreted as the probability that a helix will continue given that it has recently started. From positions -30 to -7 there is a roughly equal chance that such a continuation will occur regardless of whether the sequence is an acceptor or not. Then at positions -6 to -4 the likelihood of a helix continuing drops by approximately 20% in acceptors

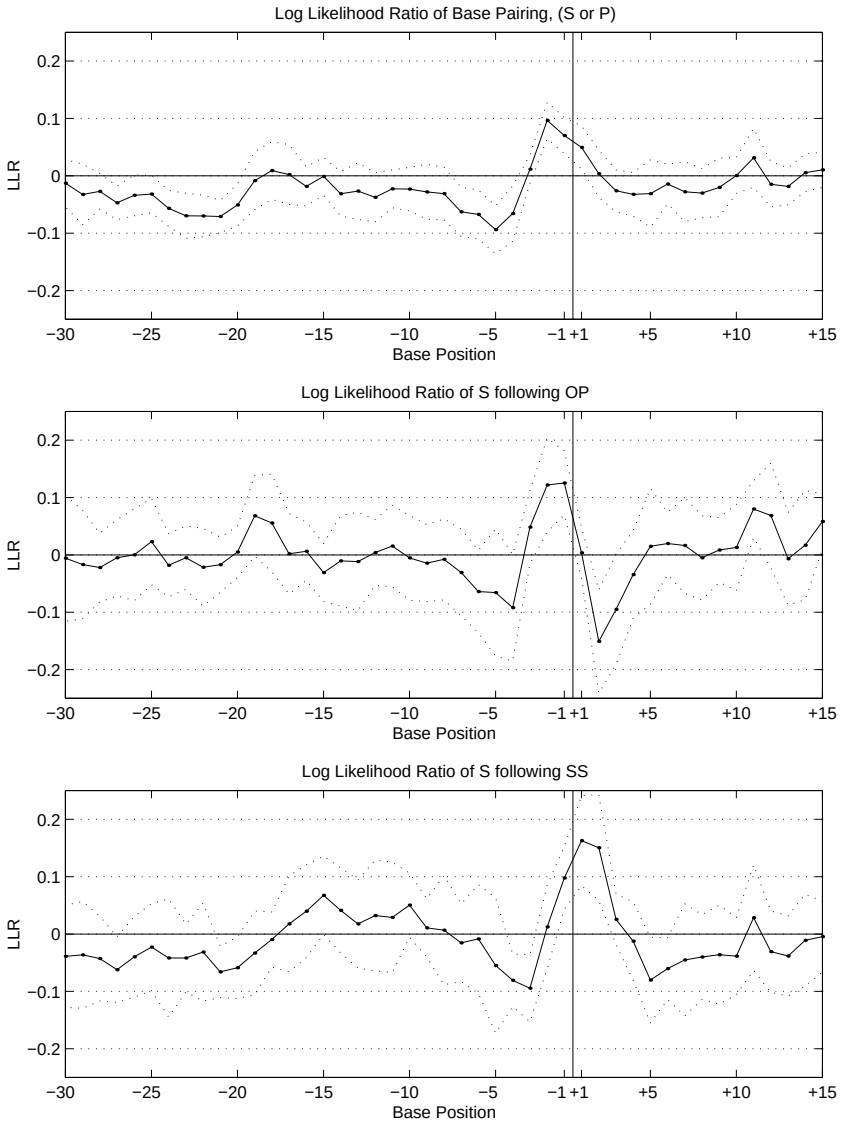

Figure 1: Log likelihood ratios (LLRs) of three structural patterns occurring at different positions relative to the acceptor splice site (vertical line). Top: LLR that a base pairs (either stacked or unstacked) with any other in the folding window. Middle: LLR that a base forms a stacked pair, conditioned on a helix start one position upstream. Bottom: LLR that a base continues a helix that begins three or more positions upstream. Solid and dotted lines show mean $\pm$ one standard deviation across the cross-validation groups.

relative to non-acceptors. Just before the splice site, the trend reverses for the acceptors, suggesting a bias toward helix formation at position -3 and -2. Shortly after the splice site there is a bias away from helix formation.

Finally, the bottom graph shows the conditional probability of a stacked pair, S, directly following two other stacked pairs, SS. This can be interpreted as the probability that a helix of length 3 or greater will continue. There is a bias at positions -5 to -3 prior to the splice site for termination of helices, but once a helix extends into the splice site region, there is a strong bias toward continuation. After the splice site, the bias is reversed briefly.

Collectively, these three graphs suggest that acceptor sequences are more likely than our non-acceptor sequences to form a short helix at the splice site.

4 Conclusion

We have presented evidence that valuable information can be extracted from predictions of pre-mRNA structure that aid in the location of acceptor splice sites. Multiple machine learners were able to utilize this information to produce statistically significant improvements in accuracy. While specific structural signatures were not detected, general trends toward helix formation in the region of the splice site suggest the possibility of greater exploitation of structural cues by gene finding algorithms.

Similar structural biases were observed at the donor splice site in the same data set, but not with sufficient strength that statistical significance could be ascribed to them. Future research is warranted toward the development of models that capture structural features at the donor splice site, as well as improving upon the acceptor site models we have presented and their biological interpretation.

Acknowledgments

This research was partially supported by grant NSF-DBI 9974498. DJP was supported by a National Defense Science and Engineering Graduate Fellowship, USA. Thanks to Benno Schwikowski and the anonymous referees for helpful comments.

References

1. M. Burset and R. Guigó. Evaluation of gene structure prediction programs. *Genomics*, 34(3):353–367, 1996.
2. R. Guigó, P. Agarwal, J.F. Abril, M. Burset, and J.W. Fickett. An assessment of gene prediction accuracy in large DNA sequences. *Genome Research*, 10(10):1631–1642, 2000.

3. M. Pertea, X. Lin, and S.L. Salzberg. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Research*, 29(5):1185–1190, 2001.
4. S. M. Mount. A catalogue of splice junction sequences. *Nucleic Acids Research*, 10(2):459–472, 22 January 1982.
5. R. Staden. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Research*, 12:505–519, 1984.
6. M. Q. Zhang and T. G. Marr. A weight array method for splicing signal analysis. *Computer Applications in the Biosciences*, 9(5):499–509, 1993.
7. C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268:78–94, 1997.
8. Melissa J. Moore, Charles C. Query, and Phillip A. Sharp. Splicing of precursors to mRNA by the spliceosome. In Raymond F. Gesteland and John F. Atkins, editors, *The RNA World: the nature of modern RNA suggests a prebiotic RNA world*, number 24 in Cold Spring Harbor Monograph Series, pages 303–357. Cold Spring Harbor Laboratory Press, 1993.
9. Jonathan P. Staley and Christine Guthrie. Mechanical devices of the spliceosome: Motors, clocks, springs, and things. *Cell*, 92:315–326, 1998.
10. David Solnick. Alternative splicing caused by RNA secondary structure. *Cell*, 43:667–676, December 1985.
11. Domenico Libri, Anna Piseri, and Marc Y. Fiszman. Tissue-specific splicing in vivo of the β-tropomyosin gene: Dependence on an RNA secondary structure. *Science*, 252:1842–1845, June 1991.
12. Béatrice Clouet d'Orval, Yves d'Aubenton Carafa, Joëlle Marie, and Edward Brody. Determination of an RNA structure involved in splicing inhibition of a muscle-specific exon. *Journal of Molecular Biology*, 221:837–856, 1991.
13. Béatrice Clouet d'Orval, Yves d'Aubenton Carafa, Pascal Sirand-Pugnet, Maria Gallego, Edward Brody, and Joelle Marie. RNA secondary structure repression of a muscle-specific exon in HeLa cell nuclear extracts. *Science*, 252:1823–1828, June 1991.
14. James O. Deshler and John J. Rossi. Unexpected point mutations activate cryptic 3' splice sites by perturbing a natural secondary structure within a yeast intron. *Genes & Development*, 5:1252–1263, 1991.
15. Andrés F. Muro, Massimo Caputi, Rajalakshmi Pariyarath, Franco Pagani, Emanuelle Buratti, and Francisco E. Baralle. Regulation of fibronectin EDA exon alternative splicing: Possible role of RNA secondary structure for enhancer display. *Molecular and Cellular Biology*, 19(4):2657–2671, April 1999.
16. Luca Varani, Masato Hasegawa, Maria Grazia Spillantini, Michael J. Smith, Jill R. Murrell, Bernardino Ghetti, Aaron Klug, Michel Goedert, and Gabriele Varani. Structure of tau exon 10 splicing regulatory element RNA and desta-

bilization by mutations of frontotemporal dementia and parkinsonism linked to chromosome 17. *Proceedings of the National Academy of Science USA*, 96:8229–8234, July 1999.

17. K.U. Mir and E.M. Southern. Determining the influence of structure on hybridization using oligonucleotide arrays. *Nature Biotechnology*, 17:788–792, 1999.
18. M. Amarzguioui, G. Brede, E. Babaie, M. Grøtli, B. Sproat, and H. Prydz. Secondary structure prediction and *in vitro* accessibility of mrna as tools in the selection of target sites for ribozymes. *Nucleic Acids Research*, 28(21):4113–4124, 2000.
19. M. Zuker, D.H. Mathews, and D.H. Turner. Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. In J. Barciszewski and B.F.C. Clark, editors, *RNA Biochemistry and Biotechnology*, NATO ASI Series, pages 11–43. Kluwer Academic Publishers, 1999.
20. D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288:911–940, 1999.
21. J. R. Quinlan. *C4.5: Programs for Empirical Learning*. Morgan Kaufmann, 1993.
22. J. R. Quinlan. Bagging, boosting, and C4.5. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 725–730, Cambridge, MA, 1996.
23. V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.
24. V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
25. C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–67, 1998.
26. M. Hearst (Ed.), S.T. Dumais, E. Osuna, J. Platt, and B. Schölkopf. Trends & controversies: Support vector machines. *IEEE Intelligent Systems*, 13(4):18–28, 1998.
27. Martin Reese, David Kulp, Andrew Gentles, and Uwe Ohler. GENIE gene finding data set. http://www.fruitfly.org/sequence/human-datasets.html.
28. C.B. Burge. Modeling dependencies in pre-mRNA splicing signals. In *Computational Methods in Molecular Biology*, pages 129–64. Elsevier Science, 1998.
29. J. McCaskill. The equilibrium partition function and base pair bindings probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.
30. Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic models of proteins and nucliec acids*. Cambridge, 1998.
31. W.S. Noble (formerly W.N. Grundy). svm 1.1. http://www.cs.columbia.edu/~noble/svm/doc/.