Characterizing RNA secondary-structure features and their effects on splice-site prediction

Rezarta Islamaj Dogan
Computer Science Department
University of Maryland
College Park, MD 20742
rezarta@cs.umd.edu

Lise Getoor
Computer Science Department
University of Maryland
College Park, MD 20742
getoor@cs.umd.edu

W. John Wilbur
National Center for Biotechnology Information
NLM, NIH
Bethesda, MD 20894
wilbur@ncbi.nlm.nih.gov

Abstract— **RNA molecules are distinguished by their sequence composition and by their three-dimensional shape, called the secondary structure. The secondary structure of a pre-mRNA sequence may have a strong influence on gene splicing. In our previous work, we showed that a splice-site model employing sequence features built using our *feature generation algorithm* was very effective in predicting splice sites. The generated sequence features also contained biologically relevant features. In this paper, we extend the feature generation algorithm to construct secondary-structure features. These features capture the nucleotide pairing tendency in the splice-site neighborhood. We extend the splice-site model to include both pre-mRNA sequence and structure characteristics. The new model significantly outperforms the sequence-based features model. The identified secondary-structure features capture biologically relevant signals such as splicing silencers. We also found these signals to prefer specific regions around the splice-site neighborhood and we detail their preference.**

I. Introduction

The three-dimensional shape of proteins or nucleic acid sequences is called *secondary structure*. The secondary structure of RNA molecules is defined by the pairings of the nucleotides along the sequence. RNA secondary-structure characteristics are important in biology, because RNA sequences fold into structures that are critical to their biological functions. Moreover, RNA secondary-structure properties may help identify subsequences of nucleotides that interact with other molecules or complexes.

Human genes — and the genes of every eukaryotic organism — are composed of contiguous coding regions in the DNA sequence. The coding regions, *exons*, are separated by non-coding regions, *introns*. During the transcription process, the messenger RNA copies the portion of the DNA that contains a gene (pre-mRNA). After transcription, during the splicing process, the non-coding regions are excised from the pre-mRNA sequence. All the coding pieces, then, are ligated together into the final gene product (mRNA), ready to be translated into protein. The borders of the introns are called *splice sites*, the start of an intron is called the *donor splice site* and the end of an intron is called the *acceptor splice site*. Splicing takes place in several stages [1]. There are a number of proteins that recognize the splice-site locations and bind to the sequence facilitating the intron excision.

Splice-site prediction is the task of recognizing the actual boundaries of the protein-coding regions in the DNA sequence. Accurate splice-site prediction is a critical component of gene prediction. Gene prediction from DNA sequence data is an important goal in bioinformatics, not only to provide fast and reliable annotation of the large quantity of sequences data, but also to provide valuable biological insights. In our previous work, [2], [3], we developed a splice-site prediction model achieving significant accuracy improvements over existing methods. We also showed that the features generated using our algorithm correspond to biologically significant functional elements [4], [5].

In our splice-site prediction model, we have considered only sequence-based features. However, the splicing process is not a mere linear process. In fact, the correct identification of the splicing borders actually involves a large number of proteins. The affinity of sequence nucleotides to form pairing bonds may guide these proteins to their binding sites, thus having an important effect in the splicing process. To investigate this, we use an RNA secondary-structure prediction algorithm [6] to fold the training sequences into their secondary-structure form. Using the secondary-structure sequences, we extend our feature generation algorithm to generate structure-based features. These novel features capture the pairing tendency of the position-specific subsequences in splice-site neighborhood. The combined splice-site model of both sequence- and structure-based features improves splice-site prediction. The secondary-structure features also capture important biological properties.

The possibility of extracting useful information from RNA secondary structure for splice-site prediction was proposed by Patterson et al. in [7]. Their splice-site prediction model combined a sequence-based splice-site predictor score and a few structure-based metrics, such as the optimal folding energy score, the max-helix score, and a second-order Markov model to capture the pairing profile of a folded sequence. They suggested that there are structural cues that should be exploited by gene-finding algorithms. Our approach differs from [7] in that we searched the space of possible position-specific nucleotide pairings in order to find specific features that improved splice-site prediction. We also offer biological interpretation for the identified features. Our recent work demonstrated that our sequence-based splice-site predictor achieved much better results than the WAM model, which was used as the sequence-based predictor in their work.

We describe our data in Section II and the generation of structure-based features, using our feature generation algorithm, in Section III. Section III also summarizes the definitions of the sequence-based features used in the splice-site prediction model. We provide a detailed description of our experiments, using the novel features, in Section IV. Next, we discuss our findings and the possible biological relevance of the new features.

II. Data characteristics

The dataset used for feature generation was a collection of 162-nucleotide-long training sequences centered at the splice site. Both upstream and downstream regions were 80 nucleotides long and the sequence alphabet was {A,C,G,T}. The acceptor-site training data contained 20,996 positive instances and 200,000 negative instances, and the donor-site training data contained 20,761 true positive instances and 200,000 negative instances. We used these sequences to generate sequence-based features. For secondary structure characteristics, we need the three-dimensional shape. We used the RNA secondary-structure prediction algorithm, Afold [6], to fold all the training sequences into their three-dimensional form. Afold was modified so that given the training sequences as input, the result was a new set of sequences that, for each nucleotide, denoted whether it participated in paring bonds in the secondary form. Those constituted the secondary-structure sequences.

We wanted to understand if splicing was affected by the pairing tendency of the nucleotides in the close neighborhood of the splice site. To answer that question, we plotted the fraction of positive sequences having a paired k-nucleotide subsequence (k-mer) for each position of its length and compared it with that of the negative sequences. Those plots are shown in Figures 1 and 2. We were surprised to see that for acceptor splice-site sequences, the positive sequences showed a higher tendency to have paired k-mer sequences in the upstream region, with a clear peak of pairing tendency just before the actual splice-site position. The donor splice-site sequences, on the other hand, showed a tendency toward reduced k-mer pairings in the upstream region and a higher tendency for pairing in the downstream region.

These observations are of interest because they are consistent with the actual splicing scenario that takes place in living cells. These findings encouraged us to investigate the possible impact of secondary-structure features on splice-site prediction.

III. Feature Generation for Splice-Site Prediction

This section describes our feature generation algorithm (FGA) [2], [3]. FGA uses domain knowledge and data properties to construct and select useful features for the prediction task. Starting with an initial feature set, FGA iteratively calls a feature construction method to expand the current feature set, and a feature selection method to reduce the feature set size to manageable levels. After a specified number of iterations, the algorithm produces an output feature set. Those features are, in turn, used by a classification algorithm for the classification task. The classifier that consistently gave the best performance for our data was CMLS [8]. We used the 11-point average precision (11ptAvg Precision) to evaluate the performance of our algorithm. For any recall ratio, we calculated the precision at the threshold that achieved that recall ratio. The 11ptAvg is the average of precisions estimated at the recall values of 0%, 10%, 20%, …, 100%.

A. Feature Construction for Splice-site prediction

The first stage of the feature generation algorithm generates feature sets useful for splice-site prediction. Initially, we define the basic elements to construct features. In the case of pre-mRNA sequences, we use the nucleotide alphabet and sequence length to construct sequence-based features.

1) Feature Construction for Sequences: We considered several feature types that capture compositional and positional properties of sequences: *general k-mer, upstream/downstream k-mer, position-specific k-mer and conjunctive positional* features. We have described these features and their individual construction methods in [3]. Here we extend our algorithm to capture the secondary-structure characteristics of the splice-site sequence.

2) Feature Construction for Secondary-Structure Sequences: We define a novel feature type that captures the structure characteristics of the RNA sequences, the *position-specific paired k-mers*. A position-specific paired k-mer is a string of k nucleotides that, in the output sequence of the RNA secondary-structure algorithm, is predicted to form pairing bonds with other nucleotides in the sequence. To identify possible binding motifs for the proteins that affect splicing, we employ our feature generation algorithm to identify useful position-specific paired k-mer features.

Construction Method: This construction method starts with an initial set of position-specific paired k-mer features and expands them to a set of position-specific paired $(k+1)$-mers by appending the letters of the alphabet to each feature. As an example, assume $F_{initial}$ is $\{AACC_1\}$. This set contains one feature, the 4-mer "AACC" starting at the first sequence position. Each nucleotide of this feature was paired in the secondary structure. Now, we can extend it to the next level set of position-specific paired 5-mers, $F_{constructed} = \{AACCA_1, AACCC_1, AACCG_1, AACCT_1\}$. In that manner, we incrementally construct higher levels.

B. Feature Selection for Constructed Features

To reduce the size of our constructed feature sets, we considered different feature selection methods. *Information Gain* (IG) measures the number of bits of information obtained for category prediction by knowing the presence or absence of a feature. *Chi-Square* (CHI) statistic measures the lack of

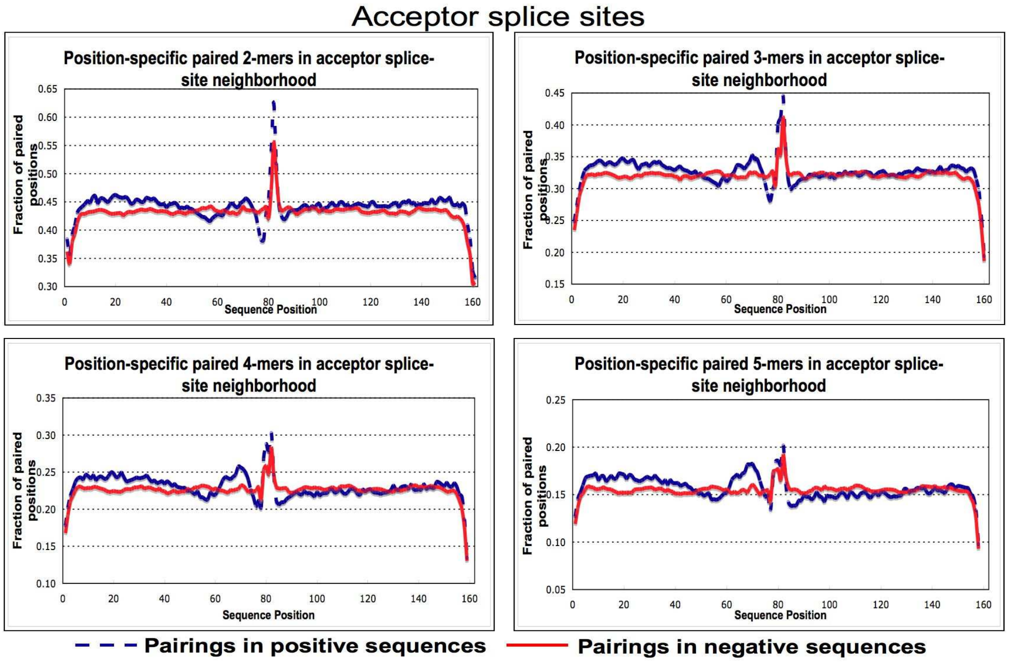

Fig. 1. Position-specific paired features found in true acceptor-site sequences (positive) vs. non-acceptor-site sequences (negative). The acceptor-site consensus “AG” is at positions [80,81] in the sequence. The upstream region, the sequence region to the left of the splice site, indicated pairing affinity in the true sequences.

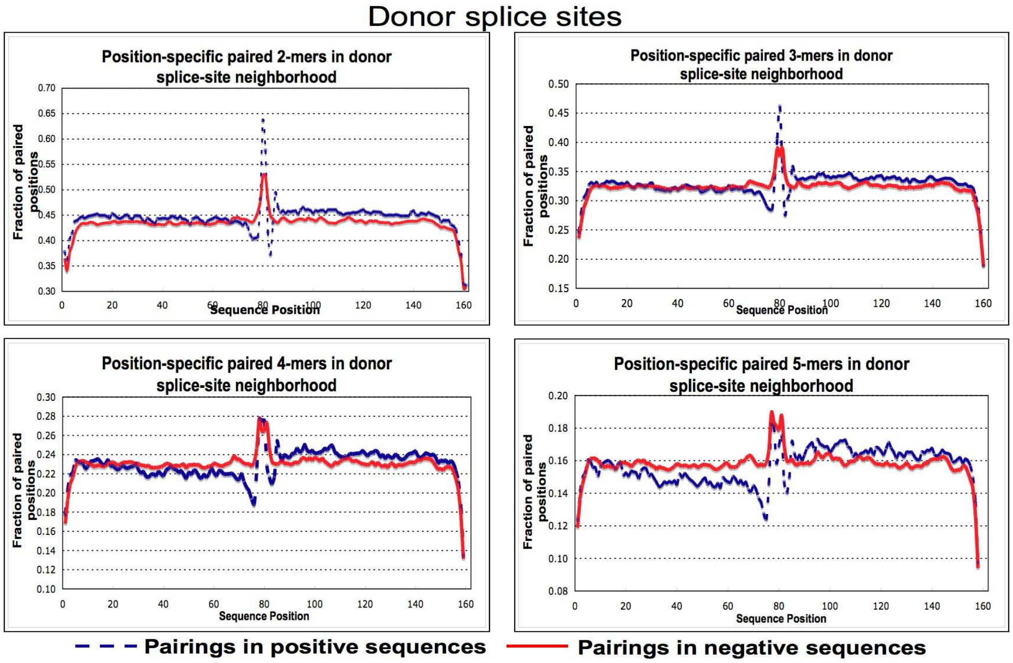

Fig. 2. Position-specific paired features found in true donor-site sequences vs. non-donor-site sequences. The donor-site consensus “GT” is at positions [80,81] in the sequence. The upstream region shows a lower pairing affinity compared to the downstream region, the sequence region to the right of the splice site. A smaller fraction of pairings was observed in true sequences, compared to negative sequences in the upstream region.

independence between feature f and the category c_i. *Mutual Information* (MI) is a criterion commonly used in statistical language modeling of word associations. The definitions of these values are the same as those presented by Yang and Pedersen in [9]. The other filtering method that we use, *KL-distance* (KL) criterion, measures the divergence between the distribution of features present in a training sequence and the categories that sequence may belong to. KL definition is given by Schneider in [10].

For each initial feature set, we iterate between the construction feature method to obtain more complex features, and a feature selection method to reduce the dimensionality of the constructed set. We perform this process for a predefined number of iterations. In this manner, we generate different feature sets, each useful for splice-site prediction.

Recursive Feature Elimination

After we generate the individual feature sets separately, we collect all features into a mixed set. Starting with the mixed set, we learn a prediction model using a classifier similar to linear support vector machines. The CMLS classifier [8] produces a decision boundary that discriminates between the two different categories. Each feature is assigned a weight during learning. These weights define the decision boundary and can be used for ranking. Features with zero weights, or weights very close to zero, are assumed to not contribute to the classification task [11], and are therefore eliminated. In this manner, we learn a new model and eliminate a fixed number of features after each iteration.

C. Splice-Site Prediction Model

Our generated features are of two major feature types: features capturing sequence properties and features capturing structure properties of the splice-site neighborhood. Using this natural separation, we use a classifier to learn sequence- and structure-features splice-site prediction models. Then, we define a new model for the splice-site prediction — a linear combination of the structure-features model and the sequence-features model:

$$Score_{seq} = c_0 + c_1 * Score_{structure} + c_2 * Score_{sequence}$$

The structure-model and sequence-model of splice-site prediction are used to score a held-out training-sequences set. Then, we use the classifier to learn the coefficients for the linear combination of the models. We give a detailed analysis of all the mentioned methods and their results for the problem of splice-site prediction, in the next section.

IV. Experiments and Discussion

To explore the splice-site prediction impact of the nucleotides showing high pairing potential, we conducted the sets of experiments described in this section. All the reported 11ptAvg precision values are the results of three-fold cross validations.

A. Position-specific paired k-mers

Similar to our position-specific sequence-based k-mer features [3], we constructed all the position-specific k-mers for k values ranging from 1 to 5. We scored the features, using the feature-selection methods and used the top $1,000$ features to predict both donor and acceptor splice sites. The results are shown in Table I.

We collected 4000 features from position-specific paired k-mer sets for k from 2 to 5. To this set, we added position-specific paired 1-mer features (648 for a 162 nucleotide long sequence). We applied recursive feature elimination on those sets of features, as shown in Tables II(a) and II(b). Compared with individual results of our sequence-based features, the 11ptAvg precision performance of the position-specific paired k-mers was very promising. It clearly showed that such a feature carried an important amount of information, which could possibly help to further understand the splicing mechanism. Also, when combined with a previously identified sequence-based features model, it might provide a model that could substantially increase our ability to predict splice sites from stretches of un-annotated RNAs.

B. Splice-site prediction with sequence and structure-based features

We selected a set of features from the position-specific paired k-mers to combine with our previously identified acceptor and donor sequence-features sets [3]. The mixed model for acceptor-site prediction contained a collection of 3000 sequence features 3148 structure features. The mixed model for donor-site prediction contained a collection of 1675 sequence features and 3148 structure features. These models produced the following 11ptAvg precision results: 89.74% for acceptor splice sites and 89.46% for donor splice sites. Although producing a low rate of false positives and ranking well, those results have not produced better predictions, compared with our sequence-based feature model (see our previously published work for a comprehensive description of those results). In fact, at this point, our sequence-based feature models have produced higher splice-site prediction 11ptAvg precisions: 90.35% for acceptor splice-sites and 90.61% for donor splice sites.

To understand the importance of the secondary-structure features for splice-site prediction, we conducted the following experiments. Starting with the whole set of sequence and structure features, we applied recursive feature elimination, eliminating 200 features for each iteration. Tables III(a) and III(b) show the splice-site prediction results for both acceptor and donor datasets in our experiments. In those tables we also list the number of features that describe sequence composition or structure characteristics for each mixed feature set. We also trained the classifier and built prediction models for each separate sequence- and structure-feature set and reported the individual 11ptAvg precisions.

From the results shown in Table III we made several observations. First, the sequence composition was of primary importance in defining a splice site. The 11ptAvg results of

TABLE I

Feature-generation comparison for position-specific paired k-mer features for k from 1 to 5 for acceptor and donor splice-site prediction. We give the 11ptAvg precision for each set when all the features are used and when top-1000 features are used for different selection methods.

Acceptor-Site Models

K-mer	All	IG	KL	MI	CHI
1	61.79	-	-	-	-
2	64.46	62.11	61.84	46.62	62.13
3	59.82	55.05	-	43.46	54.96
4	51.04	42.93	36.98	40.17	43.02
5	44.13	38.72	27.17	37.20	-

(a)

Donor-Site Models

K-mer	All	IG	KL	MI	CHI
1	61.07	-	-	-	-
2	66.08	61.88	61.78	44.29	61.92
3	-	54.73	53.09	47.91	54.61
4	51.21	44.06	41.30	39.42	43.40
5	45.29	43.14	35.12	41.37	43.70

(b)

TABLE II

Splice-site prediction results for position-specific paired k-mer features for different stages of recursive feature elimination using CMLS. We start with 4648 features for (a) acceptor and (b) donor, where 648 is the number of position-specific paired nucleotides and 4000 are the CHI-selected features for k values from 2 to 5. For each iteration we reduce the number of features by 500 and report the 11ptAvg for splice-site prediction.

Nr of Features	11ptAvg (Acceptor)
4648	66.81
4148	66.84
3648	66.91
3148	66.74
2648	66.33
2148	65.24
1648	64.39
1148	61.80
648	58.47

(a)

Nr of Features	11ptAvg (Donor)
4648	69.77
4148	69.82
3648	69.17
3148	69.03
2648	68.55
2148	67.68
1648	65.81
1148	65.28
648	63.10

(b)

models built only on sequence features consistently showed high values. Second, specific nucleotide pairings of particular locations could be the key to the discovery of important binding sites. The 11ptAvg results of models built only on structure features were several orders of magnitude higher than random (10%). And third, the secondary-structure information improves splice-site prediction, in addition to the sequence-based features. For example, as shown in Table III(a), when the number of features was reduced to 3048, the addition of paired position-specific features increased the 11ptAvg from 89.69%, which was the result of sequence-based features, to 90.36%. This result was statistically significant with alpha 0.005.

C. New prediction model with sequence- and structure-based information

The results in Tables III(a) and III(b) suggested that adding the structure-based features in the large mix of features does not produce a visible difference in the splice-site prediction results. Instead, in order to profit from the information encoded in the newly generated features, we used the combined model. The combined model initially learns two different splice-site models; one based on the structure features and one based on the sequence ones. To illustrate this, we selected the feature set of size 3000 in Table III. This set contained 1679 position-specific paired k-mers (structure features) and 1321 general, upstream, downstream and position-specific k-mers and conjunctive positional features (sequence features). The 11ptAvg result for splice-site prediction of the structure-based features model was 60.42% and the 11ptAvg of the sequence-based features model was 90.19%. We learned the new splice-site prediction model as a linear combination of the structure-features model and the sequence-features model:

$$Score_{seq} = w_0 + w_1 * Score_{structure} + w_2 * Score_{sequence}$$

We trained the classifier and learned the weights that defined the linear combination model. The linear combination model produced an 11ptAvg precision of 91.46% for donor splice-site prediction. This result was an improvement over the 90.36% obtained when using the whole set of 3000 donor features (mixed), and over the 90.19% obtained when using only the sequence features, as shown in Table III. This improvement is statistically significant for alpha 0.005.

V. Biological significance

Figures 1 and 2 showed that the nucleotide pairing tendency in the positive sequences supported the actual splicing scenario. In our splice-site prediction experiments, we generated features that captured the pairing tendency of nucleotides in specific positions in the sequence. In this section, we focus on the generated secondary-structure features and search for known splicing regulator signals.

The biological signals that are present in the splice-site neighborhood fall into these categories. Exonic splicing enhancers (ESE) are signals that activate the nearby splicing sites. Exonic splicing silencers (ESS) act as suppressors to the splicing activity. Both enhancing and silencing effects are accomplished via the different types of proteins that bind to the ESE and ESS signals. Fairbrother et al. [12] identified 238 candidate ESE 6-mers, the RescueESE set. Goren et al. [13] identified a set of 285 candidate splicing regulator 6-mers, the ESR set. And Wang et al. [14] derived a set of 176 candidate ESS 6-mers, the FasESS set.

TABLE III

ACCEPTOR(A) AND DONOR (B)SPLICE-SITE PREDICTION 11PTAVG RESULTS. RECURSIVE FEATURE ELIMINATION IS PERFORMED FOR MIXED FEATURES MODELS OF ACCEPTOR AND DONOR SITES. FOR EACH ITERATION WE REDUCED THE NUMBER OF FEATURES BY 200. AFTER EACH ITERATION, WE COUNTED THE NUMBER OF STRUCTURE- AND SEQUENCE-BASED FEATURES THAT WERE SELECTED AND BUILT SEPARATE PREDICTION MODELS FOR EACH. THESE RESULTS ARE ALSO LISTED.

Acceptor Models (No.Features and 11ptAvg)

Mix Model		Structure		Sequence	
5848	89.74	2941	66.55	2907	90.35
5048	90.05	2400	64.23	2648	90.02
4448	90.76	1981	62.83	2467	90.27
4048	90.55	1668	60.26	2380	90.26
3448	90.37	1227	58.52	2221	90.09
3048	90.36	957	55.41	2091	89.69
2448	90.25	583	45.84	1865	89.68
2048	89.51	376	37.60	1672	89.30
1448	89.12	153	32.04	1295	88.51
1048	88.42	57	24.00	991	87.79

(a)

Donor Models (No.Features and 11ptAvg)

Mix Model		Structure		Sequence	
4823	89.46	3148		1675	90.61
4000	89.83	2482	64.68	1518	90.22
3400	90.13	2009	62.11	1391	90.26
3000	90.36	1679	60.42	1321	90.19
2400	90.76	1206	57.00	1194	90.20
2000	90.75	933	50.58	1067	90.23
1600	90.57	677	44.25	923	90.13
1000	90.15	335	34.08	665	89.82
600	89.46	183	25.64	417	89.20

(b)

Because the secondary-structure features generated by the FGA algorithm captured the pairing information of different nucletides and their preferred location, we hypothesize that, these specific paired features may have discovered ESE and ESS sites in the splice-site neighborhood. To test that, we compared them with the published ESE and ESS sets [12]–[14]. Our generated features contained, at most, position-specific paired 5-mers. Therefore, to compare with the exonic splicing regulator sets, we derived all the 5-mers contained in the 6-mer sets. The RescueESE set contained 208, the ESR set contained 297 and the FasESS set contained 142 unique 5-mers. We computed the overlap between the FGA-generated 5-mer sets and the 5-mers in the regulator sets. For each overlap, we computed the p-value, based on the hypergeometric distribution. The results are shown in Table IV.

The set of FGA-generated 5-mers of the downstream donor region produced a significant overlap with the FasESS set of splicing silencer signals. The splicing silencer signals are more subtle signals and therefore more difficult to discover. The upstream donor region 5-mers produced a significant overlap with the ESR set of splicing regulator signals. To investigate these signals further, we selected the 5-mer features, that produced the overlap, and searched their exact positions in the splice-site neighborhood. We divided the neighborhood into six regions: the *far, near* and *close* regions upstream or downstream from the annotated splice-site position. The far region upstream or downstream denoted the interval $50 - 80$ nucleotides away from the splice site. The near region denoted the interval from 20 to 50 nucleotides and the close region denoted the 20 nucleotides upstream or downstream the splice site. We grouped the overlapped 5-mer features into these six regions and we listed them in Table V. This detailed description has not been done before and we hope it will be of value to biologists. Although some of the signals appear in more than one region, it is interesting to note that, the weight of the features also changed with position, sometimes even switching sign.

VI. CONCLUSIONS

In this paper we presented an extension to our feature generation algorithm to construct features that capture the three-dimensional characteristics of genomic sequences. This algorithm was applied to the problem of splice-site prediction, and a new splice-site predictor model was proposed. The new model employed features that captured both sequence composition and structural shape characteristics of splice-site sequences. The linear combination of structure-features model with sequence-features model improved the splice-site prediction accuracy significantly. Moreover, the features employed by the structure-based model were found to overlap significantly with splicing regulator motifs. We divided the 160-nucleotide splice-site neighborhood into six regions, and mapped the position preference of the identified biologically relevant signals. This detailed description is likely to be valuable to biologists. In our future work, we plan to investigate other biologically relevant information, such as, the identification of features that capture the tendency not to create a pairing bond and their particular position.

ACKNOWLEDGMENTS

This research was supported in part by an appointment to the National Center for Biotechnology Information Scientific Visitors Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education.

REFERENCES

[1] S. M. Mount, "Messenger rna splicing signals," *Encyclopedia of Life Sciences*, 2001.

[2] R. Islamaj, L. Getoor, and W. J. Wilbur, "Feature generation algorithm: Application to splice-site prediction," in *International Workshop on Feature Selection for Data Mining: Interfacing Machine Learning and Statistics*, 2006.

[3] ——, "A feature generation algorithm for sequences with application to splice-site prediction," in *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2006.

[4] R. I. Dogan, L. Getoor, W. J. Wilbur, and S. M. Mount, "Spliceport - an interactive splice site analysis tool," *Nucleic Acids Research*, 2007.

[5] ——, "Features generated for splice-site prediction correspond to functional elements," submitted.

TABLE IV

THE SET OF POSITION-SPECIFIC PAIRED 5-MERS IN THE FINAL SPLICE-SITE MODEL IS DIVIDED INTO UPSTREAM AND DOWNSTREAM FEATURES SETS. THESE SETS WERE COMPARED WITH THE SETS OF 5-MERS FOUND IN RESCUE-ESE, FAS-ESS AND ESR SETS OF EXONIC SPLICING REGULATORS. FOR EACH COMPARISON WE FOUND THE OVERLAP AND CALCULATED THE P-VALUE. THE NUMBERS IN BOLD SHOW SIGNIFICANT RELATIONSHIPS.

		Rescue-ESE (208)		FAS-ESS (142)		ESR (297)	
FGA set	size	Overlap	P-value	Overlap	P-value	Overlap	P-value
Don-5mer-downstream	160	26	0.93553	46	**4.291e-08**	50	0.27672
Don-5mer-upstream	128	32	0.09996	16	0.725485	62	**5.043e-07**

TABLE V

THE FGA-GENERATED POSITION-SPECIFIC *paired* 5-MER FEATURES THAT OVERLAPPED WITH FASESS AND ESR SETS. THE FEATURES ARE GROUPED INTO SIX REGIONS: *far*, *near* AND *close* UPSTREAM OR DOWNSTREAM THE SPLICE-SITE LOCATION. THE FAR REGION COVERS FEATURES THAT APPEAR IN POSITIONS $50 - 80$ NUCLEOTIDES, THE NEAR REGION $20 - 50$ AND THE CLOSE REGION COVERS FEATURES IN POSITIONS $0 - 20$ NUCLEOTIDES AWAY FROM SPLICE SITE.

Region	5-mer Features overlapping with Fas-ESS signals
Far - upstream	CCTGG, GCTGC, TGCTG, TTGTG
Near - upstream	CCCTG, CCTGC, CCTGG, CCTTC, CGAGG, CGTGG, GCCAT, GCGGC, TGGAG
Close - upstream	CCAGG, CCAGT, CCATC, CCTGG, CTGCA, CTTCC, GGCAA
Close - downstream	AAGTT, AGATG, AGATT, AGGTG, AGGTG, AGTAT, AGTGA, AGTTG, AGTTT GGTAG, GGTGT, GTATA, GTTCA, GTTGT, GTTTG, GTTTT, AAGGG, AAGTG AGGGT, AGGTA, AGTAG, AGTCC, AGTGG, AGTTA, GATTA, GTAGG, GTGGC GTTCT, GTTCT, GTTGG, TGGGA, TGGGG, TTTCT, AGGGG, CTGGG, GGGGG
Near - downstream	GAGGG, GGGAG, GGGGA, GGGTG, GTGGG, CGGGG, GGAGG GGGGT, GGTGG, GGGGG, CTGGG, AGGGG

[6] A. Y. Ogurtsov, S. A. Shabalina, A. S. Kondrashov, and M. A. Roytberg, "Analysis of internal loops within the rna secondary structure in almost quadratic time," *Bioinformatics*, vol. 22, no. 11, 2006.

[7] D. J. Patterson, K. Yasuhara, and W. L. Ruzzo, "Pre-mrna secondary structure prediction aids splice site prediction," in *Pacific Symposium on Biocomputing*, 2002.

[8] T. Zhang and F. J. Oles, "Text categorization based on regularized linear classification methods," *Information Retrieval*, vol. 4, no. 1, pp. 5–31, 2001.

[9] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Machine Learning, Proceedings of 14th International Conference (ICML 1997)*, 1997, pp. 412–420.

[10] K.-M. Schneider, "A new feature selection score for multinomial naive bayes text classification based on kl-divergence," in *Meeting of the Association of Computational Linguistics (ACL)*, 2004, pp. 186–189.

[11] X. H.-F. Zhang, K. A. Heller, I. Hefter, C. S. Leslie, , and L. A. Chasin, "Sequence information for the splicing of human pre-mrna identified by support vector machine classification," *Genome Research*, vol. 13, no. 12, pp. 2637–2650, 2003.

[12] W. Fairbrother, G. Yeo, R. Yeh, P. Goldstein, M. Mawson, P. Sharp, and C. Burge, "Rescue-ese identifies candidate exonic splicing enhancers in vertebrate exons," *Nucleic Acids Research*, vol. 1, no. 1;32, pp. W187–90, 2004.

[13] A. Goren, O. Ram, M. Amit, H. Keren, G. Lev-Maor, I. Vig, T. Pupko, and G. Ast, "Comparative analysis identifies exonic splicing regulatory sequences-the complex definition of enhancers and silencers," *Molecular Cell*, vol. 23;22, pp. 769–81, 2006.

[14] Z. Wang, M. Rolish, G. Yeo, Tung, M. Mawson, and C. Burge, "Systematic identification and analysis of exonic splicing silencers," *Cell*, vol. 119, pp. 831–845, 2004.