Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals

Gene Yeo [1 2] and Christopher B. Burge [1 3]

Abstract

We propose a framework for modeling sequence motifs based on the Maximum Entropy principle (MEP). We recommend approximating short sequence motif distributions with the Maximum Entropy Distribution (MED) consistent with low-order marginal constraints estimated from available data, which may include dependencies between non-adjacent as well as adjacent positions. Many Maximum Entropy models (MEMs) are specified by simply changing the set of constraints, and are utilized to discriminate between signals and decoys. Classification performance using different MEMs gives insight into the relative importance of dependencies between different positions. We apply our framework to large datasets of RNA splicing signals. Our best models outperform previous probabilistic models in the discrimination of human 5' (donor) and 3' (acceptor) splice sites from decoys. Finally, we discuss mechanistically-motivated ways of comparing models.

Key words: Maximum entropy, splice sites, non-neighboring dependencies, Markov models, maximal dependence decomposition, molecular sequence analysis, sequence motif

1 Introduction

Given a set of aligned sequences representing instances of a particular sequence motif, what model should be used to distinguish additional motif occurrences from similar sequences? This problem occurs commonly in computational biology with examples of DNA, RNA and protein sequence motifs. For example, it is important identify signal peptides in protein sequences and to to recognize true sites of RNA splicing from 'decoy' splice sites in primary transcript sequences. A number of statistical models have been developed to approximate distributions over sets of aligned sequences. For example, Markov Models (MMs) and Hidden Markov Models (HMMs) are commonly used in

[1]Department of Biology, Massachusetts Institute of Technology, 77 Massachusetts Avenue Building 68-223, Cambridge, MA, 02319

[2]Department of Brain and Cognitive Sciences, Center for Biological and Computational Learning, Massachusetts Institute of Technology, Cambridge, MA, 02319

[3]To whom correspondence should be addressed: cburge@mit.edu, fax: (617)452-2936, phone: (617)258-5997

bioinformatics[12], with applications in gene-finding and protein domain modeling [18].

We propose that the most unbiased approximation for modeling short sequence motifs is the Maximum Entropy Distribution (MED) consistent with a set of constraints estimated from available data. This approach has the attractive property that it assumes nothing more about the distribution than that it is consistent with features of the empirical distribution which can be reliably estimated from known signal sequences. In this paper we consider low-order marginal distributions as constraints, but other types of constraints can also be accomodated. Such models have been exploited in natural language processing [3], amino acid sequence analysis [5] and as a weighting scheme for database searches with profiles [19].

We introduce our approach in Section 2.1, define "constraints" in Section 2.2, and define Maximum Entropy models (MEM) in Section 2.3. In Section 2.4, we describe the use of Brown's iterative scaling [4] procedure of iterative scaling to obtain the MED consistent with a given set of constraints. In Section 2.5, we introduce a greedy-search information maximization strategy to rank constraints. This approach is applied to splice site recognition[7], an important problem in genetics and biochemistry, for which an abundance of high quality data are available. We focus on effectively modeling the 9 base sequence motif at the 5' splice site (5'ss), and the $\sim$23 base sequence motif at the 3' splice site (3'ss) of human introns, and not on the general problem of gene prediction. However, better modeling of the splice signals should lead to improved gene prediction and can be used to predict the splicing phenotypes of mutations that alter or create splice sites. The constraints for a MEM can also be ranked in importance. Finally, we propose a straightforward mechanistically-motivated way of comparing splice site models in terms of local optimality.

2 Methods

2.1 Maximum Entropy Method

Let X be a sequence of λ random variables $X = \{X_1, X_2, ..., X_\lambda\}$ which take values from the alphabet $\{A, C, G, T\}$. Let lower-case $x = \{x_1, x_2, ...x_\lambda\}$ represent a specific DNA sequence. Let $p(X)$ be the joint probability distribution $p(X_1 = x_1, X_2 = x_2, ..., X_\lambda = x_\lambda)$, and upper case $P(X = x)$ denote the probability of a state in this distribution (i.e. there are 4^λ possible states).

The principle of maximum entropy was first proposed by Jaynes [16] and states that of all the possible distributions in the hypothesis space that satisfy a set of constraints (component distributions, expected values or bounds on these values), the distribution that is the best approximation of the true distribution given what is known (and assuming nothing more) is the one with the largest Shannon entropy, H, given by the expression

$$H(\hat{p}) = -\sum \hat{p}(x) log_2(\hat{p}(x)) \tag{1}$$

where the sum is taken over all possible sequences, x. We will use logarithms to base 2, so that the entropy is measured in bits. Shannon entropy is a measure of the average uncertainty in the random variable X, i.e. the average number of bits needed to describe the outcome of the random variable. The set of constraints should therefore be chosen carefully and must represent statistics about the distribution that can be reliably estimated. It is possible to specify a set of constraints which are "inconsistent" in that they cannot be simultaneously satisfied (e.g. $\{P(A,A) = 3/4, P(T,T) = 1/2\}$). However, all constraint sets used here will be subsets of the marginal frequencies of the "empirical distribution" on sequences of length λ, and will therefore be consistent. The uniqueness of the MED for a consistent set of constraints was proved by Ireland and Kullback [15].

The principle of minimum cross-entropy or minimum relative entropy (MRE), first introduced by Kullback, is a generalization of the MEP that applies in cases when a background distribution q is known in addition to the set of constraints. Of the distributions that satisfy the constraints, the MRE distribution is the one with the lowest relative entropy (or KL-divergence), D, relative to this background distribution:

$$D_{KL}(\hat{p}) = \sum \hat{p}(x) log \frac{\hat{p}(x)}{q(x)} \tag{2}$$

Minimizing $D_{KL}(\hat{p})$ is equivalent to maximizing $H(\hat{p})$ when the prior q is a uniform distribution on the sequences of length λ. Shore and Johnson (1980) proved that maximizing any function but entropy will lead to inconsistencies unless that function and entropy have identical maxima [26]. This implies that if we believe that the constraints are correct and well estimated (and no other information is assumed), then the MED is the best approximation of the true distribution.

2.2 Marginal Constraints

For convenience, we consider two categories of constraints: "complete" constraints, which specify sets of position dependencies and "specific" constraints, which are constraints on (oligo-)nucleotide frequencies at a subset of positions.

2.2.1 "Complete" Constraints

Omitting the hats over the variables for convenience, let S_X be the set of all lower-order marginal distributions of the full distribution, $p(X = \{X_1, X_2, ..., X_\lambda\})$. A lower-order marginal distribution is a joint distribution over a proper subset of X. For example, for $\lambda = 3$,

$$S_X = \{p(X_1), p(X_2), p(X_3), p(X_1, X_2), p(X_2, X_3), p(X_1, X_3)\} \tag{3}$$

Define $S_s^m \subseteq S_X$, where superscript m refers to the *marginal-order* of the marginal distributions and the subscript s refers to the *skips* of the marginal distribution. In Equation 3, the first three elements are $1st$-order marginals (i.e. $m = 1$), and the last three elements are $2nd$-order marginals (i.e. $m = 2$): $p(X_1, X_2)$ and $p(X_2, X_3)$ are the $2nd$-order marginals with skip 0 ($s = 0$), and $p(X_1, X_3)$ is the 2nd-order marginal with skip 1 ($s = 1$). They are illustrated in our notation below:

$$S_0^1 = \{p(X_1), p(X_2), p(X_3)\}$$

$$S_0^2 = \{S_0^1, p(X_1, X_2), p(X_2, X_3)\}$$

$$S_1^2 = \{S_0^1, p(X_1, X_3)\}$$

$$S_X = S_{0,1}^2 = \{S_0^1, p(X_1, X_2)p(X_2, X_3), p(X_1, X_3)\}$$

For convenience, we include S_0^1 in S_s^m whenever the marginal order, $m > 1$. For an aligned set of sequences of length λ, the $1st$-order constraints (S_0^1) are the empirical frequencies of each nucleotide (A,C,G,T) at each position, and the Maximum Entropy Distribution consistent with these constraints is the Weight Matrix Model (WMM), i.e. all positions independent of each other [7]. On the other hand, if 2nd-order nearest-neighbor constraints (i.e. S_0^2) are used, the solution is a Inhomogeneous 1st-order Markov model (I1MM) (Appendix A). Consequently, different sets of constraints specify many different models. The performance of a model tells us about the importance of the set of constraints that was used.

2.2.2 "Specific" Constraints

"Specific" constraints are observed frequency values for a particular member of a set of "complete" constraints. Continuing with the example above, the list of 16 "specific" constraints for $p(X_1, X_3)$ are: $\{A \cdot A, A \cdot C, A \cdot G, A \cdot T, ..., T \cdot A, T \cdot C, T \cdot G, T \cdot T\}$, where $A \cdot A$ is the observed frequency of occurrence of the pattern ANA ($N = A, C, G$ or T).

2.3 Maximum Entropy Models

A Maximum Entropy Model (MEM) is specified with a set of complete constraints, and consists of two distributions, namely, the signal model ($p^+(X)$) and the decoy probability distribution ($p^-(X)$), both of which are the MEDs generated by iterative-scaling (Section 2.4) over constraints from a set of aligned signals and a set of aligned decoys of the same sequence length, λ, respectively. Given a new sequence, the MEM can be used to distinguish true signals from decoys based on the likelihood ratio, L,

$$L(X = x) = \frac{P^+(X = x)}{P^-(X = x)} \tag{4}$$

where $P^+(X = x)$ and $P^-(X = x)$ are the probability of occurrence of sequence x from the distributions of signals(+) and decoys(-), respectively. Following the Neyman-Pearson lemma, sequences for which $L(X = x) \geq C$, where C is a threshold that achieves the desired true-positive rate α, are predicted to be true signals.

2.4 Iterative Scaling to Calculate MED

In simple cases, the MED consistent with a set of constraints can be determined analytically using the method of Lagrange multipliers, but analytical solutions are not practical in most real world examples. Instead, the technique of iterative scaling is used. This technique was introduced by Lewis [21] and Brown [4], who showed that the procedure described below converges to the MED consistent with the given lower order marginal distributions. There is no limitation on the number or type of component distributions that can be employed [4]. Brown showed that at each step of the iteration, the approximation to the MED improves, using Equation(2) as a measure of closeness of the approximating distribution to the true distribution, but the proof of convergence is not rigorous (see [15] for a rigorous proof of convergence).

The iteration procedure begins with a uniform distribution with terms $P^0(X) = 4^{-\lambda}$, so all sequences of length λ are equally likely. Next, we specify a set of complete constraints and a corresponding list of specific constraints. Represent each member of the ordered list of specific constraints as Q_i, where i is the order in the list. The next step is to sequentially impose the specific constraints, Q_i, that the approximating distribution must satisfy. The terms relevant to the constraint at the j^{th} step of iteration have the form:

$$P^j = P^{j-1} \frac{Q_i}{\hat{Q}_i^{j-1}} \tag{5}$$

where P^{j-1} is a term at the $(j-1)$th step in the iteration while P^j is the corresponding term at the jth step, Q_i is the ith constraint in the list of "specific constraints" and $\hat{Q}_i^{j-1}$ is the value of the marginal corresponding to the ith constraint determined from the distribution p at the $j-1$th step. To illustrate, we return to our example in Equation 3 and apply constraint $Q_i = A \cdot A$ at the jth step:

$$P^j(X = ANA) = P^{j-1}(X = ANA) \frac{Q_i}{\hat{Q}_i^{j-1}} \tag{6}$$

where

$$\hat{Q}_i^{j-1} = \sum_{N \in \{A,C,G,T\}} P^{j-1}(X = ANA) \tag{7}$$

All terms not included in this sum (i.e. triplets not matching ANA) are iterated as follows:

$$P^j(X = VNW) = P^{j-1}(X = VNW) \frac{1 - Q_i}{1 - \hat{Q}_i^{j-1}} \tag{8}$$

for VNW such that $V \neq A$ or $W \neq A$, $N \in \{A, C, G, T\}$.

Note that enforcing satisfaction of a constraint at step j may cause a previous constraint to be unsatisfied until the previous constraint is applied again. This process is iterated until convergence or until a sufficiently accurate approximation is obtained.

2.5 Ranking Position Dependencies

As the iterations proceed, the entropy, H (Equation 1) of successive distributions $p(X)$ decreases from the maximum value $log_2(4^\lambda)$ to that of the MED. This makes intuitive sense- as more constraints are applied, the distribution contains more information, hence lower entropy. For our purposes, we say the entropy has converged when the difference in entropy between iterations becomes very small (e.g. $\mid \Delta H \mid \leq 10^{-7}$). A KL-divergence criterion gives similar results. We have found that convergence typically requires about 10-20 complete iterations of the constraints for a cutoff of $\mid \Delta H \mid \leq 10^{-7}$.

Applying different constraints reduces the entropy of the distribution by different amounts. Therefore, we can control the rate of convergence by changing the order in which the constraints are applied. We perform a greedy search to rank constraints by the amount that they reduce the entropy of the solution as described below.

Greedy-search Entropy-reduction Strategy

A first list ("bag of constraints") is initialized to contain all specific constraints. A second list, the "ranked list", is initially empty. At each iteration:

1. Initialize a uniform distribution.
2. Determine the MED consistent with all constraints from the "ranked list".
3. Apply the first constraint from the "bag of constraints". Determine the reduction in H relative to the distribution determined in step 2, ΔH_i. Repeat all the constraints separately, recording ΔH_i for each constraint.
4. Place the constraint with the largest ΔH_i in the "ranked list".
5. Repeat steps 1 to 4 until all constraints in the "bag of constraints" have been placed in the "ranked list".

It is important to emphasize that the ranking of a constraint depends on the constraints ranked before, so that this algorithm is not guaranteed to determine the optimal subset of k constraints for $2 \leq k \leq N-1$, where N is the total number of constraints. Another possible criterion for ranking

(instead of ΔH_i) is ΔKL_i defined as the reduction in relative entropy (Equation 2). Constraints can also be ranked in larger groups, instead of one at a time, thus speeding up the process.

3 Splice Site Recognition

The success of gene finding algorithms such as Genscan [8], HMMgene [17] and Genie [20] is critically dependent on finding the signals that mark exon-intron boundaries, which are recognized in cells by the nuclear pre-mRNA splicing machinery. The two strongest contributing signals are the donor or 5' splice site (5'ss) and the acceptor or 3' splice site (3'ss), which demarcate the beginning and end of each intron, respectively.

In [28], a number of algorithms that predict human splice sites were compared, indicating, as might be expected, that algorithms which use global and/or local coding information and splice signals (HMMgene and NetGene2) perform better than algorithms that only use the splice signals themselves (NNSPLICE, SpliceView and GeneID-3). Here, we focus on modeling the discrete splicing signals of specific length, with the understanding that once these have been optimally modeled, they could be incorporated into more complex exon or gene models if desired.

A number of models have been developed that can be estimated from reasonably sized sets of sequences[7]. Weight Matrix models (WMMs) assume independence between positions. Although this assumption is frequently violated in molecular sequence motifs [6], WMMs are widely used because of their simplicity and the small number of sequences required for parameter estimation (SpliceView and GeneID-3 score splice sites based on Weight Matrix models [25]). Inhomogeneous first-order Markov models (I1MMs) account for nearest-neighbor dependencies which are often present in sequences and usually can discriminate sites more accurately than WMMs. However, I1MMs ignore dependencies between non-adjacent positions, which may also be present. Higher-order Markov models account for more distant neighboring dependencies, but the number of parameters that have to be estimated and hence the required number of training samples increases exponentially with Markov order.

Decision tree approaches, such as the Maximal Dependence Decomposition (MDD) [7] used in Genscan and GeneSplicer [24] reduce the parameter estimation problem by partitioning the space of signals such that each leaf of the tree contains a sufficiently-sized subset of the sites and the strongest dependencies between positions are modeled at the earliest branching points when the most data are available. Cai and colleagues applied probabilistic tree networks and found that simple first-order Markov models are surprisingly effective for modeling splice sites[10]. Arita and colleagues utilize the Bahadur expansion to approximate training of Boltzmann machines to model all pair-wise correlations in splice sites and found no improvement compared to first-order Markov models for 5'ss, but better performance for the 3'ss [1]. Our work is related to the latter two approaches in that

we introduce a general family of models in which Markov models appear as natural members. It is worth noting that in addition to (non-)adjacent pairwise dependencies, MEMs can accommodate third-order or higher-order dependencies.

3.1 Construction of Transcript Data

To avoid using computationally predicted genes, available human cDNAs were systematically aligned to their respective genomic loci by using a gene annotation script called GENOA (L.P. Lim and C.B.B. unpublished). To simplify the analysis, genes identified by this script as alternatively spliced were excluded. We used a total of 1821 non-redundant transcripts that could be unambiguously aligned across the entire coding region, spanning a total of 12,715 introns (hence 12,715 5'ss and 12,715 3'ss). Our training and test data sets comprise disjoint subsets of these data. We use sequences at positions {-3 to +6} of the 5'ss (i.e. last 3 bases of the exon and the first 6 bases of succeeding intron), which have the GT consensus at positions {+1,+2}, and the sequences at positions {-20 to +3} of the 3'ss with the AG consensus at positions {-2,-1} (see Table 1). These splice sites are recognized by the major class or U2-type spliceosome that is universal in eukaryotes. We excluded 5'ss that have the GC consensus and 5'ss or 3'ss that matched the consensus patterns for splicing by the minor class or U12-type spliceosome. Decoy splice sites are sequences in the exons and introns of these genes that match a minimal consensus but are not true splice sites e.g. decoy 5' splices sites are non-splice sites matching the pattern N_3GTN_4 and decoy 3'ss are non-splice sites that match the pattern $N_{18}AGN_3$ [9]. Table 1 here

4 Results and Discussion

4.1 Models of the 5' splice site

The various models tested are listed in Table 2. The text abbreviations are in the first column, where "me" stands for maximum entropy, "s" stands for skip and "x" stands for the maximum skip; the first number is the marginal order and the second is the skip number or maximum skip number. Figure 1 and Table 2 together illustrate the improvement in performance resulting from use of more complex constraints. From the ROC analysis (Figure 1 and Appendix C), it is clear that me2s0 (equivalent to a I1MM), does much better than the me1s0 (equivalent to a WMM), as has been observed previously [7], indicating that nearest-neighbor contributions are important in human 5'ss. Our best model according to ROC analysis and maximum correlation coefficient analysis (Appendix B) for the 5' site is the me2x5 model, which takes into account all pair-wise dependencies. The MDD model used in Genscan [8] performs slightly better than the me2s0/I1MM model. Analysis using maximum 'approximate correlation' (see Appendix B) rather than maximum

correlation coefficient gave similar results.

We observe that the me2x5 model shows significant improvement over the me1s0/WMM model: the false positive rate at 90% sensitivity was reduced by approximately a factor of 2. The correlation coefficients are not large, which likely reflects properties of the human pre-mRNA splicing mechanism, in which 5'ss recognition relies heavily on other signals, such as enhancers and silencers, distinct from the splice signal itself [14] [2]. Figure 1 here Table 2 here

4.1.1 Ranked Constraints

The top 20 2nd-order constraints determined for models me2s0 and me2x5 using the greedy-search algorithm are listed in Tables 3 and 4. Figure 2A illustrates the faster increase in information content of the model when the constraints were applied in ranked order (Table 3), versus a random ordering of constraints. Furthermore, higher performance is achieved with ranked constraints versus a similar number of randomly ordered constraints (Figure 2B). Of course, when all the constraints are used, there is no difference in performance. Clearly, certain pairs of positions contain more information useful for discrimination. Also, the information content of the distribution is related to the performance of the model, i.e. the performance increases with increasing information content of the model. It is useful that the rankings of the dependencies are not just on the level of positions, but also at the level of (oligo)nucleotide sequence, a feature not seen in [10]. Some of these effects could reflect preferences of trans-acting factors which may bind cooperatively to different 5'ss positions. Table 3 here Table 4 here Figure 2 here

4.2 Models of the 3' splice site

The 3'ss sequence motifs is much longer than the 5'ss, ∼23 bases. For notational simplicity, we define the index of each position in the sequence starting from 1 to 21, excluding the invariant AG dinucleotide. To avoid the impractical task of storing and iterating over $4^{21} \approx 4 \times 10^{12}$ possible sequences, we may first break up the sequences into 3 consecutive non-overlapping fragments of length 7 each (fragments 1 to 3: positions 1 to 7, 8 to 14 and 15 to 21 respectively), build individual MEDs for the 3 fragment subsets (see Equation 9), and score new sequences by a product of their likelihood ratios (Equation 4).

$$P'(X) = P(X_1, ..., X_7)P(X_8, ..., X_{14})P(X_{15}, ..., X_{21}) \tag{9}$$

However, using Equation 9 ignores dependencies between segments. The resulting loss in performance is illustrated in Figure 3 (compare me2s0 and mm1 curves). Again, the me1s0 is equivalent to a WMM. To retain the dependencies of the nucleotides between the segments while avoiding computer memory issues, we propose the following approach. Six other fragments are modeled (fragments 4 and 5: positions 5 to 11 and 12 to 18 respectively; fragments 6 to 9: positions 5 to 7,

8 to 11, 12 to 14, 15 to 18 respectively). We then multiply the likelihood ratios for fragments 1 to 5 and divided by the likelihood ratios of fragments 6 to 9. For dependencies within 7 bases, this approach "covers" all the positions.

$$P_{overlap}(X) = \frac{P'(X)P''(X)}{P(X_5, ..., X_{11})P(X_{12}, ..., X_{18})} \tag{10}$$

where

$$P''(X) = P(X_5, X_6, X_7)P(X_8, X_9, X_{10}, X_{11})P(X_{12}, X_{13}, X_{14})P(X_{15}, X_{16}, X_{17}, X_{18})$$

The performance of this "overlapping" Maximum Entropy model is illustrated in Figure 3 (labeled modified me2s0), and performs similarly to the corresponding Markov model. Models me3s0 and me4s0 were modified analogously. Previous researchers have found that nearest-neighbor dependencies were sufficient to specify good models for 3'ss sites ([7],[10]). In fact, we found that a 2nd-order Markov model of the 3'ss site performs better than a 1st-order Markov model, but that a 3rd-order model performs worse than a 1st-order Markov model, presumably because of parameter estimation and/or sample size issues for 3rd-order transition probabilities of the form $P(L|IKJ)$ where I, J and/or K are purines (low frequency in most 3'ss positions). This observation motivates our procedure for segmenting the signal into 9 fragments, which use only 2nd-order constraints and neglects some long-range dependencies (such as between positions 1 and 21). It is possible to segment the signal in a way that captures such long-range dependencies (not shown). However we found that adding dependencies beyond 2-nucleotide separations does not significantly change the performance (Table 5 and Figure 4).

Figure 3 here
Figure 4 here
Table 5 here

4.3 Clustering Splice Site Sequences

The MDD model [7] [8] demonstrated that appropriate subdivision of the data can lead to improved discrimination. Here, we ask whether MEMs can be improved by first clustering the data into subsets. First, we generated a symmetric dissimilarity matrix D, where d_{ij} is the number of mismatches between splice site sequences i and j in the list of training set sequences. Next, we implemented hierarchical clustering on D using Ward's method. Results for our set of 5'ss are shown in Figure 5 and Figure 6.

Interestingly, we observe that the highest contributors to the information content (excluding the GT consensus) in cluster 1 come from the 3rd, 4th, 5th and 6th bases in the intron, whereas the last two bases in the exon contribute the most in cluster 2, indicating that clusters 1 and 2 represent "right-handed" and "left-handed" versions of the 5'ss motif respectively. These two classes of 5'ss might be recognized by different sets of trans-factors e.g. U6 snRNP would generally interact more strongly with "right-handed" 5'ss, while U5 snRNP should interact preferentially with "left-handed"

5'ss [9]. We can combine separately trained models in the following manner:

$$P_{combined}(X) = P(X|M_1)P(M_1) + P(X|M_2)P(M_2) \tag{11}$$

where $P_{combined}(X)$ is the probability of generating sequence X under the combined model, $P(X|M_1)$ and $P(X|M_2)$ are the conditional probabilities of generating X given the model constructed using cluster 1 and cluster 2 sequences, respectively, and $P(M_1)$ and $P(M_2)$ are the prior probabilities of cluster 1 and 2, respectively. The performance of combined 5'ss models are illustrated in Figure 7. Separating the sequences into the 2 clusters and modeling them separately with WMMs and then combining the models performs significantly better than using a WMM derived from all the sequences. However, modeling the separate clusters with me2x5 and I1MM models does not show significant improvements compared to modeling the entire cluster. Apparently, the more complicated models are able to capture cluster-specific information using the entire set of sequences. Figure 8 shows the motifs for 3'ss clusters which appear to separate into T-rich versus C/T-rich pyrimidine tracts. Combined 3'ss models showed a similar effect as with the 5'ss models (data not shown).

Figure 5 here
Figure 6 here
Figure 7 here
Figure 8 here

5 Applications of Splice Site Models

The specificity of pre-mRNA splicing hinges on highly conserved base pairing between the 5' splice site (5'ss) and U1/U6 small nuclear RNAs as well as interactions with U1C protein [11] and U5 snRNA [23]. It is unclear whether decoy splice sites are recognized by the splicing machinery. A study showing that intronic 5' decoy sites are activated when cells are heat shocked demonstrates that intronic decoys may be functional under special conditions [22]. Therefore, decoys could potentially be real splice sites, but may be blocked by the presence of RNA secondary structures [29], or have suboptimal location relative to splicing enhancers and repressors [14] [13]. Nevertheless, a good computational model should generally assign higher scores (i.e. log-likelihood ratios) to real 5'ss and lower scores to decoys, when all other factors are equal.

5.1 Proximal 5'ss decoys in introns

We have used several measures to compare the performance of different models, all of which involve comparing the sensitivity of the models for a given false positive rate (Appendices B and C). This essentially sets a global threshold, C (see Section 2.3) in deciding whether a sequence is or is not a true splice site. However, the splice site recognition machinery does not appear to use a global setting- in some cases weak splice sites are used when positioned in close proximity to splicing enhancers. This suggests a local decision rule for splice site detection, i.e. the most important factor may be whether the true splice site has higher score than decoys in its proximity.

We compared models by scoring possible 5'ss in a dataset of $\sim 12,600$ human introns. Better models

should result in a larger number of introns with no higher-scoring decoys downstream of the real 5'ss. Figure 9 shows that our best 5'ss model, me2x5, results in the greatest number of introns which have no higher-scoring decoys downstream of the real 5'ss, i.e. 69 introns more than the MDD model, and 639 more than the WMM. Moreover, the me2x5 model gives the lowest number of introns that have a first higher-scoring decoy (fhsd) in the intron within 250 bases from the upstream real 5'ss - me2x5 predicted 75 fewer such introns than MDD, and 686 fewer introns than the WMM. The three models result in approximately the same number of introns where the fhsd occurred further than 250 bases from the real 5'ss. On inspection of the length distribution of these introns, we observed that the median length for these introns were $\sim 2,770 - 2,900$ bases, whereas the rest of the introns had a median length of $\sim 650 - 750$ bases, suggesting that global optimality of splice site motifs is less important in long introns. Figure 9 here

5.2 Ranking and Competing 5'ss

The top 20 highest-scoring 5'ss sequences ranked by the me2x5 model are listed in Table 6, with their corresponding ranks by the MDD, 1IMM and WMM models and, in the last column, the "odds ratio" defined as the frequency of occurrence of the sequence as a splice site divided by its occurrence as a decoy. Different models result in significatly different rankings of the signals. Figure 10 shows that the top scoring sequences are well correlated between models, but the lower scoring sequences vary much more. Table 6 here
Figure 10 here

5.3 Predicing Splicing Mutations in the ATM gene

Ataxia-telangiectasia (A-T) is an autosomal recessive neurological disorder caused by mutations in the *ATM* gene. Recently, our Maximum Entropy 5'ss and 3'ss models have been utilized to predict the consequences of genomic mutations in the *ATM* gene that perturb splicing with promising results [30].

6 Conclusions

We recommend using the Maximum Entropy Distribution as the least biased approximation for the distribution of short sequence motifs consistent with reliably estimated constraints. We show that this approach grants us the flexibility of generating many different models simply by utilizing different sets of constraints. Our greedy-search strategy ranks constraints at the resolution of paired nucleotides at specific positions. This can be useful for determining correlations with binding factors. We demonstrate on a simple example that using the constraints in order of their ranking increases the rate of convergence to the MED, increases the information content of the distribution and

improves performance much faster than using randomly ordered constraints. The ranking of these constraints may reflect biological dependencies between nucleotides at different positions in the motif. Our best models using simply dinucleotide marginal distributions outperform previous models, e.g. WMMs and IMMs. These models themselves are MEDs when position-specific frequencies or nearest-neighbor dinucleotide frequencies are used as constraints. MEMs are relatively easy to use, e.g. the 5'ss model is stored as a 16,384-long vector in lexicographic order. We have developed a 3'ss "overlapping" Maximum Entropy model using an approach which combines multiple sub-models that performs better than models utilizing only nearest-neighbor dependencies. We show that the MED takes into account possible sub-clustering information in the data. We use a straightforward biologically-motivated way to compare models in terms of local optimality. Importantly, the MED framework described can be applied to other problems in molecular biology where large datasets are available, including classification and prediction of DNA, RNA and protein sequence motifs.

7 Acknowledgements

We thank Philip Sharp and Tomaso Poggio for helpful discussions, and Uwe Ohler and the anonymous reviewers for comments on the manuscript. This work was supported by grants from the NIH and NSF (C. B. B.) and the Lee Kuan Yew Scholarship from the government of Singapore (G. Y.).

8 Appendix

A Inhomogeneous Markov Models

A kth-order Inhomogeneous Markov Model can be generated as follows:

$$p_{kMM}(X) = p(X_1, ..., X_k) \prod_{i=k+1}^{\lambda} p(X_i | X_{i-1}, ..., X_{i-k}). \tag{12}$$

where $X = \{X_1, X_2, ...X_\lambda\}$, k is the order and $p(X_i|X_{i-1}, ..., X_{i-k})$ is the conditional probability of generating a nucleotide at position i given the previous k nucleotides. As before, conditional probabilities and marginals are estimated from the corresponding frequencies of occurrences of nucleotide combinations at the specified positions.

It is important to note that the maximum entropy distributions using nearest-neighbor constraints of marginal-order $(k+1)$ are equivalent to kth-order Markov Models. In every case, the performance of the MED for constraints S_0^k was equivalent to that of a $(k-1)th$ order Markov model. Thus the class of Markov models is a subset of the class of solutions specified by MEM.

B Performance Measures

Table 7 here

Table 7 illustrates a confusion matrix, which contains information about how well a model performs given an independent test set with real splice sites (positives) and decoys (negatives). N is the total number of test sequences, i.e. $N = TP + FP + FN + TN$. Standard Measures of accuracy such as Correlation Coefficient (CC), Approximate Correlation (AC), Sensitivity (Sn) and False Positive Rate (FPR) are defined below:

$$CC = \frac{(TP \times TN) - (FN \times FP)}{(TP + FN)(TN + FP)(TP + FP)(TN + FN))^{\frac{1}{2}}}$$

$$AC = \frac{1}{2}(\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN}) - 1$$

$$Sn = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{TN + FP}$$

C ROC analysis

Receiver Operating Curve (ROC) Analysis [27] is an effective way of assessing the performance of models when used in a binary hypothesis test. In our case, a sequence x is predicted as a splice site if the likelihood ratio, L, is greater than a threshold, C (Equation 4). A ROC curve is a graphical representation of Sn (on the y-axis) versus false positive rate (FPR) (on the x-axis) as a function of C, and has the following useful properties:

1. It shows the tradeoff between sensitivity and false positive rate (increases in sensitivity are generally accompanied by an increase in false positives).
2. The closer the curve follows the left-hand border and then the top border of the ROC plot, the more accurate the model. The area under the curve is a measure of test accuracy.
3. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the model.

Analogous to the ROC analysis, we can plot the other standard measures as described above against changing values of the threshold, C. The maximum point on the curves will indicate the best setting for C and gives a performance measure which can be used to compare models. Hence we can define CC_{max} to be the maximum correlation coefficient i.e. the highest point on the curve, and $C_{CC_{max}}$ is the threshold where CC_{max} is obtained. AC_{max} and $C_{AC_{max}}$ can be defined similarly.

References

[1] M. Arita, K. Tsuda, and K. Asai. Modeling splicing sites with pairwise correlations. *Bioinformatics*, 18(2):S27–S34, 2002.

[2] B.J. Lam, K.J. Hertel. A general role for splicing enhancers in exon definition. *RNA*, 10:1233–41, 2002.

[3] A. Berger, S. Pietra, and V. Pietra. A maximum entropy approach to natural language processing . *Computational Linguistics*, 22(1):39–71, 1996.

[4] D. Brown. A Note on approximations to discrete probability distributions. *Information and Control*, 2:386–392, 1959.

[5] E. Buehler and L. Ungar. Maximum entropy methods for biological sequence modeling . *Workshop on Data Mining in Bioinformatics, BIOKDD*, 2001.

[6] M. Bulyk, P. Johnson, and G. Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res*, 30(5):1255–61, 2002.

[7] C. Burge. Chapter 8. Modeling dependencies in pre-mRNA splicing signals. *S.L. Salzberg, D.B. Searls, S. Kasif (Eds.), Computational Methods in Molecular Biology, Elsevier Science*, pages 129–164, 1998.

[8] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268:78–94, 1997.

[9] C. Burge, T. Tuschl, and P. Sharp. Chapter 20. Splicing of precursors to mRNAs by the spliceosomes. *R. Gesteland and T. Cech and J. Atkins (Eds.), The RNA World , Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York*, pages 525–560, 1999.

[10] D. Cai, A. Delcher, B. Kao, and S. Kasif. Modeling splice sites with bayes networks . *Bioinformatics*, 16(2):152–158, 2000.

[11] H. Du and M. Rosbash. The U1 snRNP protein U1C recognizes the 5' splice site in the absence of base pairing. *Nature*, 419(6902):86–90, 2002.

[12] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Biological sequence analysis: probabilistic models of proteins and nucleic acids . *Cambridge University Press*, 1998.

[13] W. Fairbrother and L. Chasin. Human genomic sequences that inhibit splicing. *Molecular and Cellular Biology*, 20(18):6816–6825, 2000.

[14] W. Fairbrother, R. Yeh, P. Sharp, and C. Burge. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297(5583):1007–13, 2002.

[15] C. Ireland and S. Kullback. Contingency tables with given marginals . *Biometrika*, 55(1):179–188, 1968.

[16] E. Jaynes. Information theory and statistical mechanics I . *Physics Review*, 106:620–630, 1957.

[17] A. Krogh. Two methods for improving performance of an HMM and their application for gene finding. *T. Gaasterland et al(Eds.), Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology (ISMB). AAAI Press, Cambridge, UK*, pages 179–186, 1997.

[18] A. Krogh, M. Brown, I. Mian, K. Sjolander, and D. Haussler. Hidden markov models in computational biology. Applications to protein modeling. *Journal of Molecular Biology*, 235(5):1501–31, 1994.

[19] A. Krogh and G. Mitchison. Maximum entropy weighting of aligned sequences of proteins or DNA. *In Proc. Third Int. Conf. Intelligent Systems for Molecular Biology (ISMB), Eds. C. Rawlings et al. AAAI Press*, pages 215–221, 1995.

[20] D. Kulp, D. Haussler, M. Reese, and F. Eeckman. A generalized hidden markov model for the recognition of human genes in DNA. *D.J. States et al(Eds.), Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Cambridge, UK*, pages 134–142, 1996.

[21] P. Lewis. Approximating probability distributions to reduce storage requirements. *Information and Control*, 2:214–225, 1959.

[22] E. Miriami, J. Sperling, and R. Sperling. Heat shock affects 5' splice site selection, cleavage and ligation of CAD pre-mRNA in hamster cells, but not its packaging in hnRNP particles. *Nucleic Acids Research*, 22:3084–3091, 1994.

[23] A. Newman. The role of U5 snRNP in pre-mRNA splicing. *EMBO*, 16(19):5797–800, 1997.

[24] M. Pertea, X. Lin, and S. Salzberg. GeneSplicer: a new computational method for splice site prediction . *Nucleic Acids Research*, 29(5):1185–1190, 2001.

[25] M. Shapiro and P. Senapathy. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implicatinos in gene expression. *Nucleic Acids Research*, 15(17):7155–7174, 1987.

[26] J. Shore and R. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-Entropy . *IEEE Transactions on Information Theory*, 26(1):26–37, 1980.

[27] J. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240(4857):1285–1293, 1988.

[28] T. Thanaraj. Positional characterisation of false positives from computational prediction of human splice sites. *Nucleic Acids Research*, 28(3):744–754, 2000.

[29] L. Varani, M. Hasegawa, M. Spillantini, M. Smith, J. Murrell, B. Ghetti, A. Klug, M. Goedert, and G. Varani. Structure of tau exon 10 splicing regulatory element RNA and destabilization by mutations of frontotemporal dementia and parkinsonism linked to chromosome 17. *Proc Natl Acad Sci USA*, 96(14):8229–34, 1999.

[30] L. Eng, G. Coutinho, S. Nahas, G. Yeo, R. Tanouye, T. Dork, C.B. Burge, and R.A. Gatti Non-classical splicing mutations in the coding and non-coding regions of the ATM gene: maximum entropy estimates of splice junction strengths. submitted. 2003.

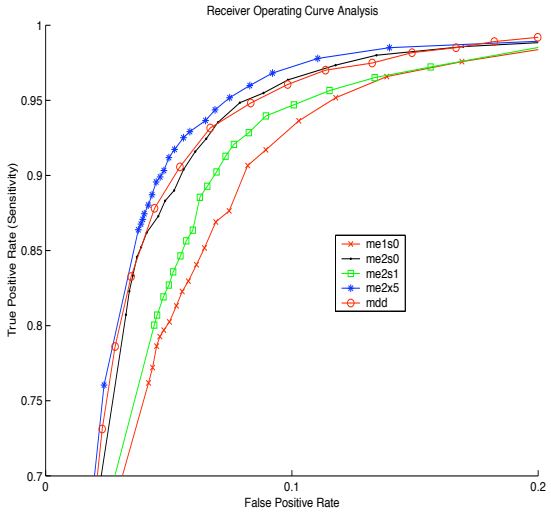

Figure 1: 5'ss: ROC curves for me2x5, me2s0, me2s1, me1s0 and MDD. The curves for the other models are not plotted, but can be inferred from Table 2.

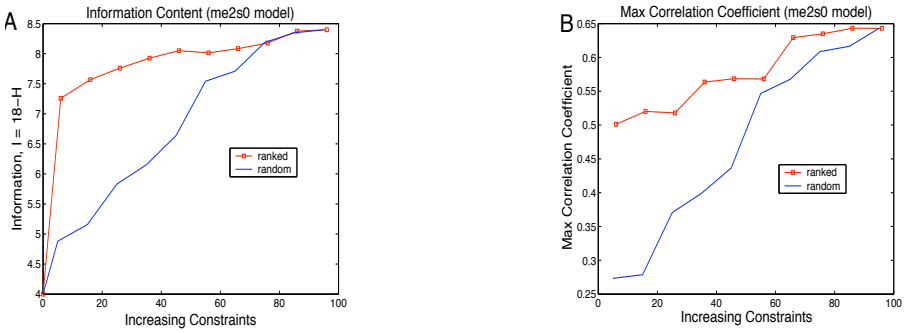

Figure 2: (A): Information content (I=18-H) of me2s0 model as constraints are added. If the constraints are ranked, the information content increases at a higher rate than if randomly ranked constraints are used. The x-axis corresponds to the model using the top N constraints. (B): Maximum Correlation Coefficient as a function of constraints. Ranked constraints added sequentially led to better performance with fewer constraints, compared to a random ordering of constraints. The model is me2s0 (excluding 1st order marginals).

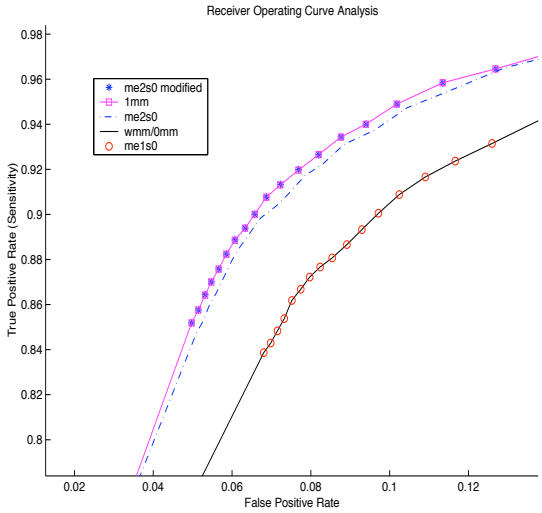

Figure 3: 3'ss: ROC curves for me2s0,me2s0 modified, me1s0, 0mm (0IMM) and 1mm (1IMM) models of the 3'ss. The curve labeled me2s0 was constructed by segmenting the 21 base long sequence set into 3 consecutive non-overlapping fragments of length 7 each. The curve labeled me2s0 (modified) was constructed as described in the text, and is equivalent to the 1st-order Markov model (I1MM).

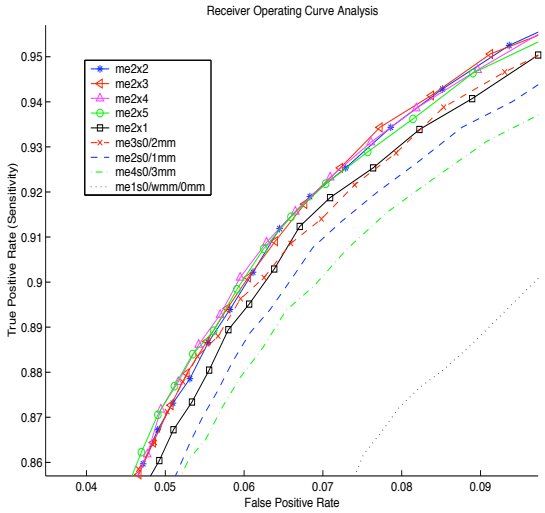

Figure 4: 3'ss: ROC curves for modified me2x5, me2x4, me2x3, me2x2, me2x1, me4s0, me3s0, me2s0 and me1s0 models of the 3'ss.

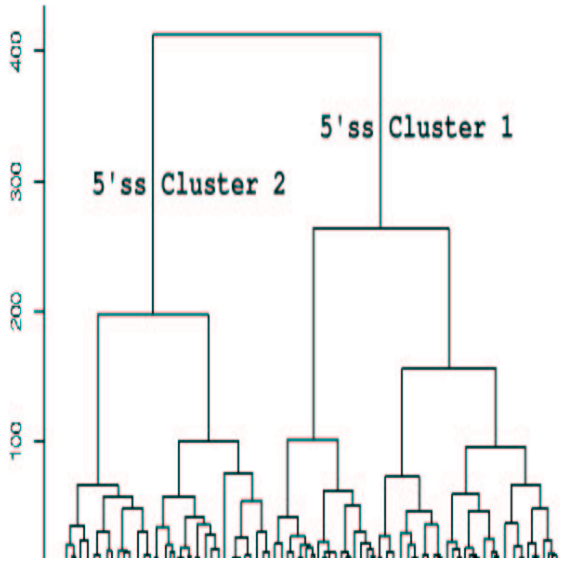

Figure 5: Truncated Dendrogram for 5'ss sequences (hierarchical clustering using ward's method). The two major clusters contain 7,260 and 5,367 sequences, respectively.

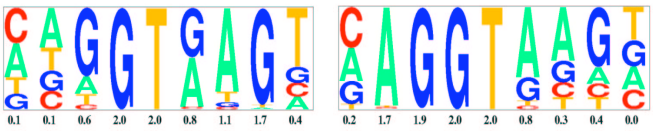

Figure 6: Sequence motifs for 5'ss cluster 1 (left) and 2 (right) created with the Pictogram program: http://genes.mit.edu/pictogram.html. The height of each letter is proportional to the frequency of the corresponding base at the given position, and bases are listed in descending order of frequency from top to bottom. The information content (relative entropy) for each position relative to a uniform background distribution is also shown.

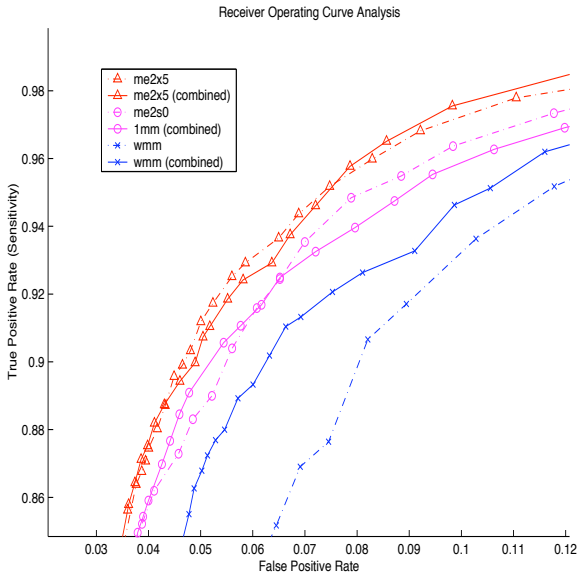

Figure 7: 5'ss: ROC curves for me2x5, 1IMM, WMM and me2x5 (combined), 1IMM (combined) and WMM (combined).

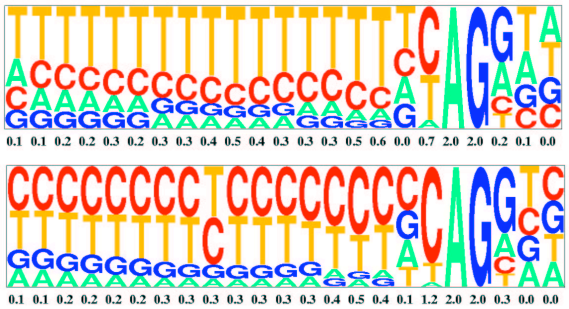

Figure 8: Sequence motifs for 3'ss cluster 1 (top) and 2 (bottom)

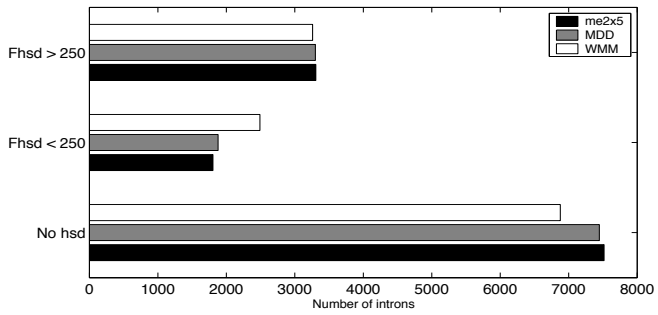

Figure 9: Bar Chart showing the number of introns that have no higher scoring decoy (hsd) than the real upstream 5'ss, and the number of introns that have a first higher scoring decoy (Fhsd) within 250 bases from the real 5'ss or greater than 250 bases from the real 5'ss.

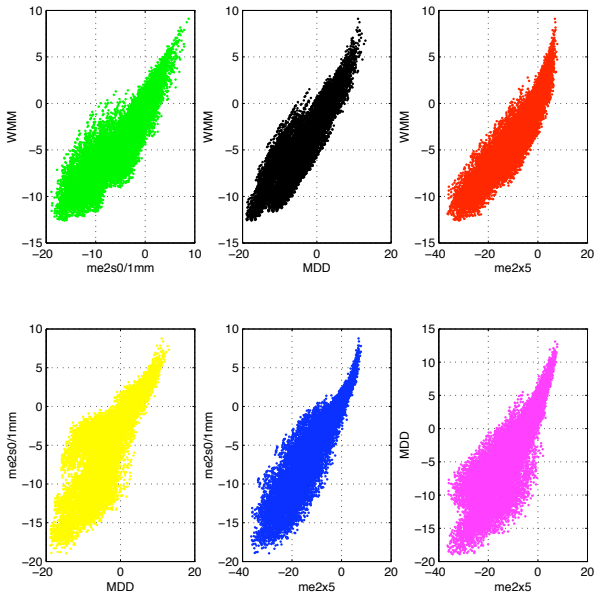

Figure 10: Scores of 5'ss sequences by different models plotted against each other. Model names are labeled in the x and y-axes.

	Real 5'ss	Decoy 5'ss	Real 3'ss	Decoy 3'ss
Train	8,415	179,438	8,465	180,957
Test	4,208	89,717	4,233	90,494
Total	12,623	269,155	12,698	271,451

Table 1: Number of sequences in 5'ss and 3'ss training and test sets.

Models	Constraints	CC
me2x5	$S^2_{1,2,3,4,5}$	0.6589
me2x4	$S^2_{1,2,3,4}$	0.6552
me2x3	$S^2_{1,2,3}$	0.6533
me5s0	S^5_0	0.6527
me2x2	$S^2_{1,2}$	0.6399
me4s0	S^4_0	0.6390
mdd	-	0.6493
me2s0	S^2_0	0.6425
me3s0	S^3_0	0.6422
me2s1	S^2_1	0.5971
me2s2	S^2_2	0.6010
me2s4	S^2_4	0.5861
me2s3	S^2_3	0.6031
me2s5	S^2_5	0.5924
me1s0	S^1_0	0.5911

Table 2: 5'ss Models ranked by ROC analysis(top to bottom), and the corresponding maximum Correlation Coefficients (CC).

Rank	ΔH_i	ΔKL_i
1	.AGgt....	.AGgt....
2	...gt.AG.	...gt.AG.
3	TA.gt....	TA.gt....
4	...gtTA..	...gtTA..
5	...gt..GT	...gt..GT
6	...gtGT..	...gtGT..
7	...gt.CG.	...gt.CG.
8	..TgtT...	..GgtG...
9	...gt.GG.	..AgtC...
10	...gtGG..	...gtCG..
11	...gtCG..	...gt..AC
12	.TAgt....	...gtCC..
13	..AgtC...	..CgtT...
14	...gt.AT.	CT.gt....
15	...gt.GT.	AG.gt....
16	...gt..CG	.TGgt....
17	...gt..AT	...gt.GT.
18	...gt..CT	.TAgt....
19	..GgtG...	..TgtC...
20	..CgtA...	...gtGG..

Table 3: Top 20 ranked constraints for me2s0. Lower letters refer to donor consensus positions. Capitalized letters are positional dependencies. All first order constraints were imposed as default.

Rank	ΔH_i	Sign
1	..Ggt..G.	-
2	...gt.AG.	+
3	.AGgt....	+
4	C..gt...C	+
5	...gtAA..	-
6	..GgtT...	+
7	..GgtC...	+
8	..GgtA...	-
9	...gtTA..	-
10	..Tgt..T.	-
11	..Tgt..A.	-
12	.G.gt..C.	-
13	...gtC.G.	+
14	.C.gt..C.	-
15	.T.gt..C.	-
16	..Cgt..A.	-
17	..Cgt..T.	-
18	..Agt..T.	-
19	..Agt..A.	-
20	..Cgt..G.	+

Table 4: Top 20 ranked constraints for me2x5 for 5'ss. + and - indicate whether the dinucleotide is more or less frequent than expected under independence assumption, respectively. All first order constraints were imposed by default.

Models	Constraints	CC
me2x2	$S^2_{1,2}$	0.6291
me2x3	$S^2_{1,2,3}$	0.6290
me2x4	$S^2_{1,2,3,4}$	0.6252
me2x5	$S^2_{1,2,3,4,5}$	0.6229
me2x1	S^2_1	0.6259
me3s0	S^3_0	0.6300
me2s0	S^2_0	0.6172
me4s0	S^4_0	0.6075
me1s0	S^1_0	0.5568

Table 5: 3'ss Models ranked by ROC analysis(top to bottom), and the corresponding maximum Correlation Coefficients (CC).

Sequence	me2x5	MDD	me2s0	WMM	odds ratio
ACGGTAAGT	1	2	5	26	184
TCGGTAAGT	2	3	12	114	77
ACGGTGAGT	3	17	18	90	11
GCGGTAAGT	4	14	10	56	3
ACGGTACGT	5	67	14	292	331
TCGGTGAGT	6	28	34	304	9
CAGGTAAGG	7	26	15	3	13
GAGGTAAGT	8	34	6	4	38
ATGGTAAGT	9	12	46	19	95
AAGGTAAGT	10	10	2	2	12
GACGTAAGT	11	41	136	86	10
CCGGTAAGT	12	1	7	17	22
CCGGTGAGT	13	7	22	68	18
CAGGTACGG	14	99	32	79	68
CAGGTAAGT	15	20	1	1	8
CAGGTAAGA	16	25	13	7	14
CGGGTAAGT	17	15	17	15	2
AAGGTACGT	18	54	8	46	233
AACGTAAGT	19	19	101	38	96
CAGGTGAGT	20	27	4	8	21

Table 6: Ranks of 5'ss sequences by different models.

	predicted positive	predicted negative
real positive	true positives, TP	false negatives, FN
real negative	false positives, FP	true negatives, TN

Table 7: Confusion Matrix