BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 4 2005, pages 471–482
doi:10.1093/bioinformatics/bti025

Prediction of splice sites with dependency graphs and their expanded bayesian networks

Te-Ming Chen[1], Chung-Chin Lu[1,*] and Wen-Hsiung Li[2]

[1]Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan and [2]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA

Received on October 27, 2003; revised on August 11, 2004; accepted on September 3, 2004
Advance Access publication September 16, 2004

ABSTRACT

Motivation: Owing to the complete sequencing of human and many other genomes, huge amounts of DNA sequence data have been accumulated. In bioinformatics, an important issue is how to predict the complete structure of genes from the genomic DNA sequence, especially the human genome. A crucial part in the gene structure prediction is to determine the precise exon–intron boundaries, i.e. the splice sites, in the coding region.

Results: We have developed a dependency graph model to fully capture the intrinsic interdependency between base positions in a splice site. The establishment of dependency between two position is based on a χ^2-test from known sample data. To facilitate statistical inference, we have expanded the dependency graph (which is usually a graph with cycles that make probabilistic reasoning very difficult, if not impossible) into a Bayesian network (which is a directed acyclic graph that facilitates statistical reasoning).

When compared with the existing models such as weight matrix model, weight array model, maximal dependence decomposition, Cai *et al.*'s tree model as well as the less-studied second-order and third-order Markov chain models, the expanded Bayesian networks from our dependency graph models perform the best in nearly all the cases studied.

Availability: Software (a program called DGSplicer) and datasets used are available at http://csrl.ee.nthu.edu.tw/bioinf/

Contact: cclu@ee.nthu.edu.tw

INTRODUCTION

Automated DNA sequencing has led to the rapid accumulation of huge amounts of DNA sequence data. This demands mathematical modeling, statistical methods and information technology to analyze the data.

Gene identification refers to the prediction of the complete gene structure, especially the precise exon–intron structure of a gene in an eukaryotic genomic DNA sequence. Genomic sequences with lengths in the order of many millions of base pairs are now being produced. Such a sequence consists of a collection of genes separated from each other by long stretches of intergenic regions. Currently, ∼30 000 genes have been estimated in the three billion base pairs of the human genome. That is, only 1.1% of the human genome seems to contain useful coding information (Lander *et al.*, 2001). However, there might be a fairly large number of human genes that remains to be identified. In response to this challenge, computational gene-finding prediction approaches have proliferated in recent years. However, their performance is still far from satisfactory (Mathe *et al.*, 2002; Zhang, 2002).

Gene identification can be regarded as an attempt to define precisely the sequential dependency on the basic biochemical processes of transcription, RNA processing and translation. The sequence properties of known genes may offer us clues about the intrinsic mechanisms of these processes (Burge and Karlin, 1997). How to model the biological signals, such as promoter elements, transcriptional and translational signals and splice sites is undoubtedly the key issue in the prediction of the complete gene structure. In this paper, we focus on the signals related to pre-mRNA splicing, i.e. the splice sites that include donor and acceptor sites are the most important elements for the prediction of precise exon–intron boundaries.

Splice signal detection

The cell recognizes a gene by utilizing different proteins to bind to different signals. Typically, there are several DNA segments required for a particular signal. We call these segments the members of the signal. However, not every member of a signal has a consensus sequence. We may and will assume that the differences between sequences for the members of a signal arose from a common ancestor via a stochastic process (Ewens and Grant, 2001), which suggests that the construction of statistical models for signals and genes is reasonable.

Several statistical models of donor and acceptor splice sites have been constructed in the past 20 years (Staden, 1984; Zhang and Marr, 1993; Burge and Karlin, 1997; Cai *et al.*, 2000; Arita *et al.*, 2002; Yeo and Burge, 2004). One of

*To whom correspondence should be addressed.

the earliest and most influential models is the weight matrix model (WMM) (Staden, 1984) that uses the position-specific compositional biases in splice sites. The WMM weights can be optimized using a neural network method (Brunak *et al*., 1991) developed for NetPlantGene (Hebsgaard *et al*., 1996) and NetGene2 (Tolstrup *et al*., 1997), and also adopted in NNSplice (Reese *et al*., 1997). Another method, called the weight array model (WAM) (Zhang and Marr, 1993), was developed to describe the dependencies between adjacent base positions by the inhomogeneous first-order Markov chain (1MC) model, and was later applied using the VEIL (Henderson *et al*., 1997) and MORGAN (Salzberg *et al*., 1998) software program.

Statistically significant dependencies between base positions in the donor and acceptor splice sites have been studied more recently (Burge and Karlin, 1997; Cai *et al*., 2000). Certain observed dependencies between splice site positions can be interpreted in terms of the spliceosome cycle between the structure of small nuclear RNPs (snRNPs) and the splice site region of the pre-mRNA (Mathews *et al*., 2000). Thus more complex splice signal models that are capable of capturing such dependencies, for instance, the maximal dependence decomposition (MDD) model in Genscan (Burge and Karlin, 1997) and Bayesian networks (Cai *et al*., 2000), have been developed. However, these more complex models do not achieve significant improvement in splice site discrimination over simpler models that assume only dependencies between adjacent positions. A possible reason is that these models do not fully capture the intrinsic interdependency between base positions either in the donor site or in the acceptor site. Significant improvement can possibly be achieved by combining one of the basic statistical models, such as WMM, WAM and MDD, of splice sites with other signal/content sensors and/or with rule-based filtering such as in GeneSplicer (Pertea *et al*., 2001), where the MDD model is combined with two second-order Markov chain (2MC) models to characterize coding/non-coding regions around splice sites in addition to a local maximal score filter. In this paper, we will develop a dependency graph model and its derivatives, and make an attempt to fully capture the intrinsic interdependency between base positions in a splice site.

METHODS

Test of dependency and the contingency tables

We used the χ^2-test as employed in MDD (Burge and Karlin, 1997) to establish the interdependency between positions in a splice signal.

To perform the null hypothesis test of independence on a pair of nucleotides at the i-th and the j-th positions of a splice site, we formed a 4×4 contingency table (Ewens and Grant, 2001), as shown in Table 1, by counting the observed number Y_{mn} of DNA sequences where the i-th nucleotide X_i was m and the j-th nucleotide X_j was n (for simplicity, we have encoded A, T, C, G as $1, 2, 3, 4$, respectively) from a sample of Y DNA sequences. The numbers Y_{mc} and Y_{rn} in Table 1 are row sums and column sums, respectively. It is clear that $\sum_{m=1}^{4} Y_{mc} = \sum_{n=1}^{4} Y_{rn} = Y$.

Table 1. A contingency table between two nucleotides in a splice site

$X_i \backslash X_j$	A	T	C	G	Total
A	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{1c}
T	Y_{21}	Y_{22}	Y_{23}	Y_{24}	Y_{2c}
C	Y_{31}	Y_{32}	Y_{33}	Y_{34}	Y_{3c}
G	Y_{41}	Y_{42}	Y_{43}	Y_{44}	Y_{4c}
Total	Y_{r1}	Y_{r2}	Y_{r3}	Y_{r4}	Y

The test statistic used is as follows:

$$\chi^2(X_i, X_j) = \sum_{m=1}^{4} \sum_{n=1}^{4} \frac{(Y_{mn} - E_{mn})^2}{E_{mn}}, \qquad (1)$$

where

$$E_{mn} = Y_{mc} Y_{rn} / Y$$

is the expected number of DNA sequences in which the i-th nucleotide X_i is m and the j-th nucleotide X_j is n from a sample of Y DNA sequences when the null hypothesis of independence was true. To determine the rejection region for the null hypothesis, we have specified a numerical value α for the Type I error of the test, according to a χ^2-distribution with degrees of freedom $(4-1)\cdot(4-1) = 9$, and then the critical point, K, was computed as follows:

$$P \text{ (null hypothesis is rejected when it is true)}$$
$$= P(\chi^2(X_i, X_j) \geq K \mid \text{null hypothesis}) = \alpha.$$

Bayesian networks

Bayesian methods provide a formalism for reasoning about partial beliefs under conditions of uncertainty (Pearl, 1988). The basic expressions in Bayesian formalism are statements about conditional probabilities. We say two random variables X and Y are independent if $P(x \mid y) = P(x)$. The variables X and Y are conditionally independent given the random variable Z if $P(x \mid y, z) = P(x \mid z)$. From the Bayesian rule, the global joint distribution function $P(x_1, x_2, \ldots, x_n)$ of variables $X_1, X_2, \ldots, X_n$ can be represented as a product of local conditional distribution functions. That is,

$$P(x_1, x_2, \ldots, x_n) = P(x_1) P(x_2 \mid x_1) \cdots P(x_n \mid x_1, \ldots, x_{n-1}).$$

A Bayesian network for a collection $\{X_1, X_2, \ldots, X_n\}$ of random variables represents the joint probability distribution of these variables. The joint probability distribution, which is

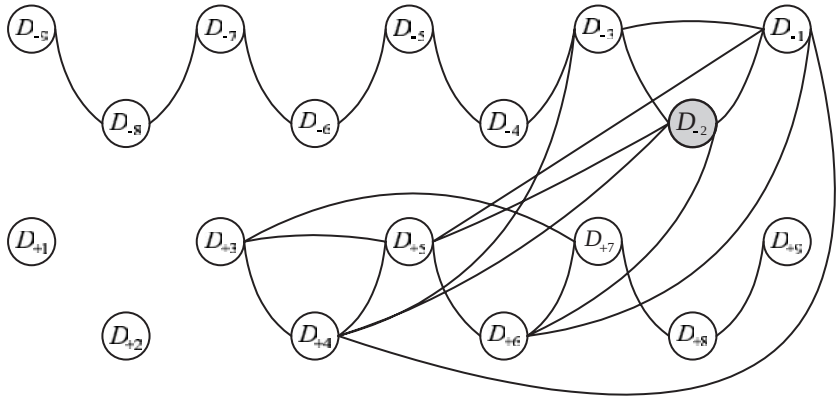

Fig. 1. The dependency graph for the donor site.

associated with a set of assertions of conditional independence among the variables, can be written as:

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid E_{x_i}),$$

where E_{x_i} is a subset of variables $x_1, \ldots, x_{i-1}$ on which x_i is dependent. Hence, a Bayesian network can be described as a directed acyclic graph consisting of a set of n nodes and a set of directed edges between nodes. Each node in the graph corresponds to a variable x_i and each directed edge is constructed from a variable in E_{x_i} to the variable x_i. If each variable has a finite set of values, to each variable x_i with parents in E_{x_i}, there is an attached table of conditional probabilities $P(x_i \mid E_{x_i})$.

MODEL ARCHITECTURE AND ALGORITHMS

Dependency graphs

The most outstanding observation made by P. Chambon is that almost all introns in pre-mRNA begin and end in the same way: the first two bases in an intron are GU and the last two are AG (Weaver, 1999). Although GU and AG are conserved at the donor site (the region surrounding the exon/intron boundary) and at the acceptor site (the region surrounding the intron/exon boundary), respectively, the bases at other positions are uncertain and require a statistical model.

We have chose a window of 18 base positions for the donor site, where nine consecutive bases are upstream from the exon/intron boundary and nine consecutive bases are downstream to the exon/intron boundary (see Results section). Further, a window of 36 base positions was chosen for the acceptor site, where 27 consecutive bases are upstream from the intron/exon boundary and nine consecutive bases are downstream to the intron/exon boundary. To be more precise, we denoted the two conserved bases of the donor site as D_{+1} and D_{+2}, i.e. $D_{+1} = \mathrm{G}$ and $D_{+2} = \mathrm{U}$. The bases in the downstream of the donor site were named as $D_{+3}, D_{+4}, \ldots, D_{+9}$ and to the upstream as $D_{-1}, D_{-2}, \ldots, D_{-9}$. On the other hand, we denoted the two conserved bases of the acceptor site as A_{-2} and A_{-1} i.e. $A_{-2} = \mathrm{A}$ and $A_{-1} = \mathrm{G}$. The bases in the downstream of the acceptor site were named as $A_{+1}, A_{+2}, \ldots, A_{+9}$ and to the upstream as $A_{-3}, A_{-4}, \ldots, A_{-27}$.

By constructing a contingency table from a sample of donor sites for each pair (D_i, D_j) of bases at distinct positions of the donor site, we tested the null hypothesis of independence for D_i and D_j with the χ^2-statistic $\chi^2(D_i, D_j)$ as in (1). We have set the Type I error of the test to $\alpha_D = 10^{-8}$ for the donor site (see Results section). According to the χ^2-distribution with nine degrees of freedom, the critical point for the rejection region of the test is $K_D = 55.4491$ for the donor site.

By rejecting the null hypothesis, we infer that the two base positions D_i and D_j are dependent if the χ^2-statistic $\chi^2(D_i, D_j) \geq K_D = 55.4491$. It is clear that each of the two conserved base positions D_{+1} and D_{+2} must be statistically independent of any other position.

The dependency graph of the donor site was constructed as follows. There were 18 nodes in the undirected graph, each corresponding to a base position of the donor site. An edge was established between two nodes in the graph if the two corresponding base positions of the donor sites were dependent, as inferred from the χ^2-test procedure. The dependency graph so obtained is shown in Figure 1, where D_{+1} and D_{+2} are isolated nodes. For future use, adjacent base positions D_j to each base position D_i in the dependency graph of the donor site were sorted from left to right according to the χ^2-values $\chi^2(D_i, D_j)$ varying from high to low (Table 2).

Using a similar procedure, we constructed the dependency graph of the acceptor site, as can be inferred from Table 3, where adjacent base positions A_j to each base position A_i in the dependency graph of the acceptor site were sorted from left to right in accordance with the χ^2-values $\chi^2(A_i, A_j)$ from high to low. The Type I error of the test for the acceptor site was chosen to be $\alpha_A = 10^{-3}$ with the critical point $K_A = 27.8772$ (see Results section).

Table 2. Adjacent base positions D_j to each base position D_i in the dependency graph of the donor site are sorted from left to right according to the χ^2-values of $\chi^2(D_i, D_j)$ from high to low

D_{-9}	D_{-8}
D_{-8}	D_{-9}, D_{-7}
D_{-7}	D_{-6}, D_{-8}
D_{-6}	D_{-5}, D_{-7}
D_{-5}	D_{-6}, D_{-4}
D_{-4}	D_{-5}, D_{-3}
D_{-3}	D_{-2}, D_{-4}, D_{-1}, D_{+4}
D_{-2}	D_{-3}, D_{-1}, D_{+4}, D_{+5}, D_{+6}
D_{-1}	D_{+6}, D_{-2}, D_{+4}, D_{+5}, D_{-3}
D_{+3}	D_{+5}, D_{+4}, D_{+7}
D_{+4}	D_{+5}, D_{+3}, D_{-2}, D_{-1}, D_{-3}
D_{+5}	D_{+4}, D_{+3}, D_{+6}, D_{-2}, D_{-1}
D_{+6}	D_{-1}, D_{+5}, D_{+7}, D_{-2}
D_{+7}	D_{+8}, D_{+6}, D_{+3}
D_{+8}	D_{+9}, D_{+7}
D_{+9}	D_{+8}

Expanded Bayesian networks

Although the dependency graph can fully capture the intrinsic interdependency between base positions in the donor site or in the acceptor site, it is difficult, if not impossible, to perform statistical inference based on the dependency graph. This is because there are cycles in the dependency graph of the donor site as shown in Figure 1 and also in the dependency graph of the acceptor site, which can be inferred from Table 3.

In contrast, as a directed acyclic graph, a Bayesian network is suitable for statistical reasoning as performed in the Cai *et al*. (2000) tree model. Whereas the Bayesian networks constructed by Cai *et al*. (2000) cannot capture the cyclic dependency among base positions as described in the dependency graph.

To resolve the dilemma, we expanded the dependency graph to form a Bayesian network by allowing a base position in the dependency graph to appear more than once in the Bayesian network as nominally distinct nodes. The basic procedure to build such a Bayesian network for the donor site is as follows:

(1) Calculate the sum $S_i = \sum_{j \in N(i)} \chi^2(D_i, D_j)$ of χ^2-statistics for each base position D_i in the donor site, where $N(i)$ is the set of indices of all base positions adjacent to D_i in the dependency graph of the donor site (here and after, two base positions of a splice site are called adjacent if there is an edge connecting them in the dependency graph of the splice site).

(2) Assign a base position D_i with the largest sum $S_i = \max_j S_j$ to be the root of the Bayesian network.

(3) Expand the dependency graph from the rooted base position to the adjacent base positions as the first layer of the Bayesian network. The root itself forms the zeroth layer of the Bayesian network.

(4) Further expand the dependency graph from each base position in the first layer of the Bayesian network to their adjacent base positions to form the second layer of the Bayesian network.

(5) Repeatedly expand the dependency graph as in Steps 3 and 4 until all base positions and the two directions in any edges in the dependency graph have been reached at least once.

As can be seen from Figure 1 (Table 3), a node in the expanded Bayesian network may have at most five (22) parent nodes. Since each node represents a variable of four possible bases, there will be up to $4^6 = 4096$ ($4^{23} \simeq 7.0 \times 10^{13}$) parameters to be estimated in establishing the conditional probability table for such a node in learning the expanded Bayesian network donor site (acceptor site) model for inference.

Considering the size of the training datasets used and to prevent overfitting the parameters of the statistical inference models, we have modified Step 4 of the basic procedure to limit each node in the expanded Bayesian network to have a maximum number p of parent nodes (we will consider $p = 1, 2, 3$) so that there are at most $4^{p+1} = 16, 64, 256$ for $p = 1, 2, 3$, respectively, instead of 4^6 or 4^{23}, parameters needed to be estimated for a conditional probability table.

To introduce the modification, we tag an adjacent index j in the adjacent index set $N(i)$ of each base position D_i in the dependency graph with an ordered pair $[n_j, \chi^2(D_i, D_j)]$, where n_j is the number of times dynamically recorded that D_j has been used as a parent node of D_i during the expansion process and $\chi^2(D_i, D_j)$ is the χ^2-statistic between D_i and D_j. We gave a total order to the tags as follows: $[n_j, \chi^2(D_i, D_j)] > [n_k, \chi^2(D_i, D_k)]$ if either $n_j < n_k$ or $n_j = n_k$ and $\chi^2(D_i, D_j) > \chi^2(D_i, D_k)$. The n_js were set to zeroes as initial values. Now, we state the modification of Steps 3 and 4 given in the above method as follows:

(3′) As in Step 3 given above. For each node D_i in the first layer, increase the first entry of the tag to the rooted base position [which must be in the list $N(i)$] by one since the rooted position has been used as a parent node of D_i.

(4′) As in Step 4 given above. If there are more than p parent nodes for a node in the second layer, keep p of the links to the parent nodes with the largest p tags and delete the rest. For each node D_i in the second layer, update the tag to each parent base position D_j in $N(i)$ by incrementing n_j by one.

The construct of the ordered tags was to ensure that the potential parent base positions for a base position in

Table 3. Adjacent base positions A_j to each base position A_i in the dependency graph of the acceptor site are sorted from left to right according to the χ^2-values $\chi^2(A_i, A_j)$ varying from high to low

A_{-27}	A_{-26}, A_{-25}, A_{-21}, A_{-24}, A_{-23}, A_{-15}, A_{-17}, A_{-3}, A_{-10}, A_{-4}
A_{-26}	A_{-27}, A_{-25}, A_{-24}, A_{-6}, A_{-20}, A_{-5}, A_{-16}, A_{-10}
A_{-25}	A_{-24}, A_{-23}, A_{-26}, A_{-27}, A_{-13}, A_{-6}, A_{-19}, A_{-5}, A_{-8}, A_{-21}, A_{-18}, A_{-14}, A_{-4}
A_{-24}	A_{-25}, A_{-23}, A_{-22}, A_{-26}, A_{-6}, A_{-15}, A_{-18}, A_{-14}, A_{-27}, A_{-16}, A_{-3}, A_{-5}, A_{-19}, A_{-8}, A_{-20}
A_{-23}	A_{-22}, A_{-25}, A_{-24}, A_{-21}, A_{-17}, A_{-19}, A_{-12}, A_{-16}, A_{-6}, A_{-5}, A_{-10}, A_{-20}, A_{-27}, A_{-14}, A_{-11}
A_{-22}	A_{-23}, A_{-21}, A_{-24}, A_{-20}, A_{-10}, A_{-16}, A_{-6}, A_{-14}, A_{-3}
A_{-21}	A_{-20}, A_{-22}, A_{-23}, A_{-19}, A_{-14}, A_{-15}, A_{-11}, A_{-27}, A_{-4}, A_{-16}, A_{-25}, A_{-17}
A_{-20}	A_{-21}, A_{-19}, A_{-14}, A_{-6}, A_{-22}, A_{-18}, A_{-3}, A_{-26}, A_{-10}, A_{-23}, A_{-5}, A_{-4}, A_{-8}, A_{-16}, A_{-17}, A_{-24}
A_{-19}	A_{-20}, A_{-18}, A_{-17}, A_{-23}, A_{-10}, A_{-21}, A_{-6}, A_{-7}, A_{-14}, A_{-15}, A_{-25}, A_{-24}
A_{-18}	A_{-17}, A_{-19}, A_{-16}, A_{-14}, A_{-20}, A_{-12}, A_{-24}, A_{-6}, A_{-8}, A_{-25}, A_{-13}
A_{-17}	A_{-16}, A_{-18}, A_{-15}, A_{-19}, A_{-23}, A_{-4}, A_{-7}, A_{-27}, A_{-12}, A_{-21}, A_{-13}, A_{-9}, A_{-20}
A_{-16}	A_{-17}, A_{-15}, A_{-14}, A_{-3}, A_{-18}, A_{-6}, A_{-8}, A_{-12}, A_{-23}, A_{-10}, A_{-5}, A_{-26}, A_{-22}, A_{-7}, A_{-11}, A_{-24}, A_{-21}, A_{-20}, A_{-9}
A_{-15}	A_{-14}, A_{-16}, A_{-17}, A_{-11}, A_{-24}, A_{-21}, A_{-13}, A_{-27}, A_{-19}, A_{-3}
A_{-14}	A_{-15}, A_{-13}, A_{-16}, A_{-3}, A_{-20}, A_{-12}, A_{-10}, A_{-5}, A_{-6}, A_{-18}, A_{-21}, A_{-19}, A_{-8}, A_{-24}, A_{-23}, A_{-22}, A_{-7}, A_{-25}
A_{-13}	A_{-12}, A_{-14}, A_{-6}, A_{-15}, A_{-25}, A_{-7}, A_{-4}, A_{-8}, A_{+4}, A_{-5}, A_{-3}, A_{-17}, A_{-18}
A_{-12}	A_{-13}, A_{-11}, A_{-14}, A_{-8}, A_{-10}, A_{-18}, A_{-3}, A_{-16}, A_{-23}, A_{-6}, A_{-5}, A_{-17}
A_{-11}	A_{-12}, A_{-10}, A_{-6}, A_{-9}, A_{-15}, A_{-7}, A_{-21}, A_{-4}, A_{-16}, A_{-23}
A_{-10}	A_{-9}, A_{-11}, A_{-8}, A_{-6}, A_{-14}, A_{-3}, A_{-12}, A_{-19}, A_{-16}, A_{-20}, A_{-5}, A_{-22}, A_{-23}, A_{-27}, A_{-26}
A_{-9}	A_{-10}, A_{-8}, A_{-11}, A_{-4}, A_{-17}, A_{-16}
A_{-8}	A_{-9}, A_{-7}, A_{-10}, A_{-3}, A_{-5}, A_{-6}, A_{-16}, A_{-12}, A_{-14}, A_{-13}, A_{-18}, A_{-25}, A_{-4}, A_{-20}, A_{-24}
A_{-7}	A_{-8}, A_{-6}, A_{-11}, A_{-19}, A_{-13}, A_{-17}, A_{-16}, A_{-14}
A_{-6}	A_{-5}, A_{-7}, A_{-10}, A_{-4}, A_{-3}, A_{-20}, A_{-11}, A_{-16}, A_{-14}, A_{-8}, A_{-13}, A_{-24}, A_{-26}, A_{-19}, A_{-23}, A_{+5}, A_{-12}, A_{-18}, A_{-25}, A_{-22}, A_{+1}, A_{+7}
A_{-5}	A_{-6}, A_{-4}, A_{-14}, A_{-8}, A_{-23}, A_{-3}, A_{-26}, A_{-10}, A_{-12}, A_{-16}, A_{-13}, A_{-20}, A_{-25}, A_{-24}
A_{-4}	A_{-5}, A_{-6}, A_{-3}, A_{+1}, A_{-17}, A_{-21}, A_{-13}, A_{-11}, A_{-20}, A_{-9}, A_{-8}, A_{-25}, A_{-27}
A_{-3}	A_{-8}, A_{-14}, A_{-16}, A_{-6}, A_{-4}, A_{-10}, A_{-12}, A_{-5}, A_{-20}, A_{-24}, A_{-13}, A_{-15}, A_{-27}, A_{-22}
A_{+1}	A_{+2}, A_{-4}, A_{+3}, A_{-6}
A_{+2}	A_{+3}, A_{+1}, A_{+8}
A_{+3}	A_{+2}, A_{+4}, A_{+9}, A_{+6}, A_{+1}
A_{+4}	A_{+5}, A_{+3}, A_{-13}
A_{+5}	A_{+6}, A_{+4}, A_{-6}
A_{+6}	A_{+5}, A_{+7}, A_{+3}
A_{+7}	A_{+6}, A_{+8}, A_{-6}
A_{+8}	A_{+9}, A_{+7}, A_{+2}
A_{+9}	A_{+8}, A_{+3}

the dependency graph would be utilized uniformly in the expansion process with a little emphasis on those with high interdependency. Table 2 has been used to facilitate the dynamic ordering of tags used in Steps (3′) and (4′). The expanded Bayesian network for the donor site as a result of the modified procedure with $p = 2$ is shown in Figure 2. Let random variable $X_i^{(l)}$ be associated with the base position D_i in the l-th layer of the expanded Bayesian network of the donor site. Let $E_{X_i^{(l)}}$ be the set of parent random variables of the variable $X_i^{(l)}$ as shown in Figure 2. For a specific DNA sequence $S = (d_{-9}, \ldots, d_{-1}, d_1, \ldots, d_9)$ of a tested potential donor site, we let $x_i^{(l)} = d_i$ for all l and for all i. Then the probability $P(S|M)$ of having S based on the expanded Bayesian network model for a donor site is defined as:

$$P(S|M) \triangleq \frac{\prod_{l,i} P(x_i^{(l)}|E_{x_i^{(l)}})}{\sum_{\{y_i^{(l)}\}} \prod_{l,i} P(y_i^{(l)}|E_{y_i^{(l)}})},$$

where the denominator is the sum of all possible base configurations $\{y_i^{(l)}\}$ of the random variables $X_i^{(l)}$ induced from all possible donor site DNA sequences and used as a normalization factor.

A similar procedure can be used to build an expanded Bayesian network for the acceptor site from the dependency

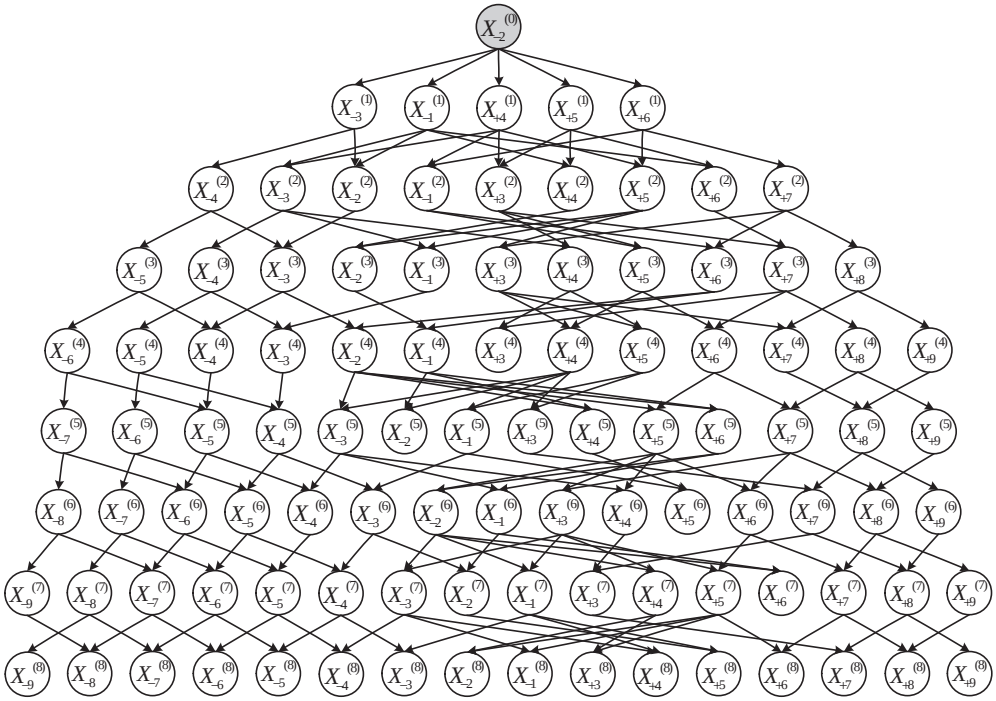

Fig. 2. The Bayesian network expanded from the dependency graph of the donor site with maximum number of parent nodes $p = 2$.

graph, but not shown here because of the high complexity. This procedure can be accomplished with the sorted lists of adjacent base positions to each base position in the dependency graph of the acceptor site as given in Table 3.

RESULTS

Splice site datasets

To build reliable expanded Bayesian networks for the detection of human splice sites, high-quality datasets must be used. We extracted a collection of 2381 real donor sites and 2381 real acceptor sites from a set of 462 annotated multiple-exon human genes at http://www.fruitfly.org/sequence/human-datasets.html. We excluded the splice sites that contained base positions not labeled with A, T, C, G but with other symbols. Finally, there were 2379 real donor sites and 2380 real acceptor sites which were used as the true dataset. We also extracted a large collection of 283 062 pseudo donor sites and 400 314 pseudo acceptor sites from the 462 annotated genes and used it as the false dataset. Each of these pseudo donor/acceptor sites has $D_{+1} = \mathrm{G}, D_{+2} = \mathrm{U}/A_{-2} = \mathrm{A}, A_{-1} = \mathrm{G}$ but is not a real donor/acceptor site according to the annotation.

Model learning

To prepare a machinery to determine whether a tested splice site is real or pseudo, we used the true training data to train a true expanded Bayesian network splice site model M_{T} and the pseudo training data to train a false expanded Bayesian network model M_{F}. Each node in an expanded Bayesian network is associated with a conditional probability table of at most 4^{p+1} parameter entries to be estimated. The maximum-likelihood estimation procedure is used, which amounts to calculate the relative frequency.

Test score

The score, $\mathrm{Score}_M(S)$, of a tested potential splice site S under the two contrast models M_{T} and M_{F} is the log-odds ratio defined as follows:

$$\mathrm{Score}_M(S) = \log\left[\frac{P(S \mid M_{\mathrm{T}})}{P(S \mid M_{\mathrm{F}})}\right],$$

where $P(S \mid M_{\mathrm{T}})$ and $P(S \mid M_{\mathrm{F}})$ are the probability of having the tested potential splice site S based on the true splice site model M_{T} and the probability based on the false splice site model M_{F}, respectively. With an empirically determined threshold score T, the tested potential splice site S will be claimed real if the log-odds score is no less than T; otherwise, it will be claimed pseudo.

Measures of predictive accuracy

A tested potential splice site was called a true positive (TP) if it was predicted true and actually true; a false negative (FN) if predicted pseudo but actually true; a true negative (TN) if predicted pseudo and actually pseudo; and a false positive (FP) if predicted true but actually pseudo.

Table 4. The consensus region of the donor site where each base position has 1 or 2 nt (in bold) with total compositional percentage not <60%

Position	D_{-3}	D_{-2}	D_{-1}	D_{+1}	D_{+2}	D_{+3}	D_{+4}	D_{+5}	D_{+6}	D_{+7}
A%	**33**	**59**	9	0	0	**49**	**71**	7	15	**25**
G%	19	13	**79**	**100**	0	**46**	12	**84**	**22**	**36**
C%	**36**	14	3	0	0	3	8	5	16	22
T%	12	14	8	0	**100**	2	9	5	**47**	17
Consensus	A/C	A	G	G	T	A/G	A	G	G/T	A/G

It is common to use the two measures of false negative (FN) rate and false positive (FP) rate defined as

$$\text{FN rate} = \frac{\#\text{FN}}{\#\text{TP} + \#\text{FN}},$$
$$\text{FP rate} = \frac{\#\text{FP}}{\#\text{TN} + \#\text{FP}},$$

to report the predictive accuracy of a splice site inference model (Cai *et al*., 2000; Pertea *et al*., 2001). Note that the sensitivity and the specificity of the inference model are equal to one minus the FN rate and one minus the FP rate, respectively (Khodarev *et al*., 2003).

Cross-validation

We used a 5-fold cross-validation in our dataset to estimate the splice site detection accuracy of all the models studied (Pertea *et al*., 2001). Each model was cross-validated by randomly partitioning the data into five subsets. Then we tested each subset (called the testing data) with the parameters trained by the other four subsets (called the training data) under the splice site model, and took the average of the five predictive accuracy measures corresponding to the five testing/training data pair. We also verified the training data with the model trained by themselves in the same manner.

Type I error selection

We considered five different values of the Type I error α, $10^{-8}, 10^{-6}, 10^{-3}, 10^{-2}$ and 10^{-1}, in the χ^2-test for the construction of the dependency graphs for the donor site and the acceptor site. According to the χ^2-distribution with nine degrees of freedom, the critical point K for the rejection of the null hypothesis is 55.4491, 44.8109, 27.8772, 21.6660 and 14.6837, respectively.

We compared the predictive accuracy of the corresponding five expanded Bayesian network predictive models with at most p parents for each window as will be described in the next section and for each $p = 1, 2, 3$. Then, we chose a Type I error for each window and for each p with the best performance for the donor site and for the acceptor site, respectively.

Table 5. The consensus region of the acceptor site where each base position has 1 or 2 nt (in bold) with total compositional percentage not <60%

Position	A%	G%	C%	T%	Consensus
A_{-27}	22	18	**31**	**30**	C/T
A_{-26}	22	19	31	28	
A_{-25}	22	16	**30**	**32**	C/T
A_{-24}	20	17	**32**	**32**	C/T
A_{-23}	22	17	**30**	**31**	C/T
A_{-22}	21	17	**32**	**31**	C/T
A_{-21}	19	16	**33**	**32**	C/T
A_{-20}	18	16	**32**	**34**	C/T
A_{-19}	16	16	**33**	**35**	C/T
A_{-18}	14	15	**34**	**36**	C/T
A_{-17}	13	17	**33**	**38**	C/T
A_{-16}	13	14	**35**	**38**	C/T
A_{-15}	11	12	**35**	**42**	C/T
A_{-14}	9	13	**37**	**41**	C/T
A_{-13}	9	12	**35**	**45**	C/T
A_{-12}	8	11	**36**	**45**	C/T
A_{-11}	8	11	**33**	**48**	C/T
A_{-10}	7	11	**37**	**46**	C/T
A_{-9}	7	12	**39**	**42**	C/T
A_{-8}	9	12	**41**	**38**	C/T
A_{-7}	8	9	**42**	**41**	C/T
A_{-6}	7	7	**45**	**41**	C/T
A_{-5}	7	6	**39**	**48**	C/T
A_{-4}	22	22	34	21	
A_{-3}	4	0	**74**	22	C/T
A_{-2}	**100**	0	0	0	A
A_{-1}	0	**100**	0	0	G
A_{+1}	23	**53**	14	10	G

Window selection

To determine a proper window for the donor site and a proper window for the acceptor site for the purpose of computational prediction, we gathered statistics of 50 bases upstream of the exon/intron boundary and 50 bases downstream of the intron/exon boundary, respectively. We found that there was a consensus region between 3 bases upstream and 7 bases downstream of the exon/intron boundary and another between 27 bases upstream and 1 base downstream of the intron/exon boundary, respectively, as shown in Tables 4 and 5 where each of the base positions had 1 or 2 nt with the total compositional percentage not <60%.

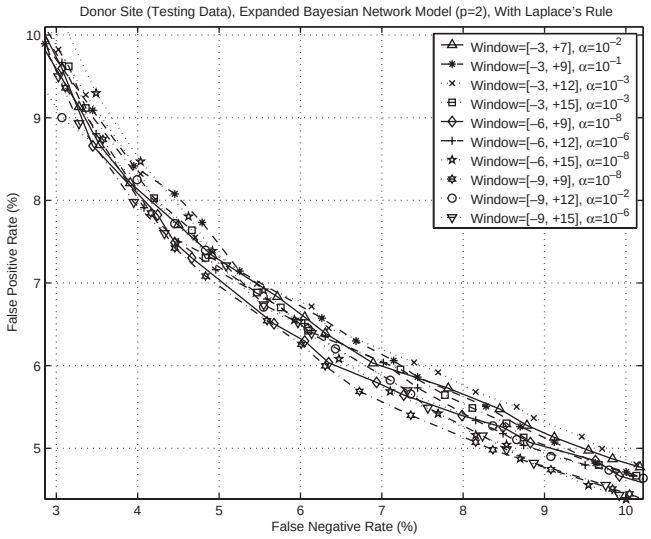

Fig. 3. Comparison of predictive accuracy of the 10 expanded Bayesian network models for the testing data of the donor site corresponding to the 10 different windows.

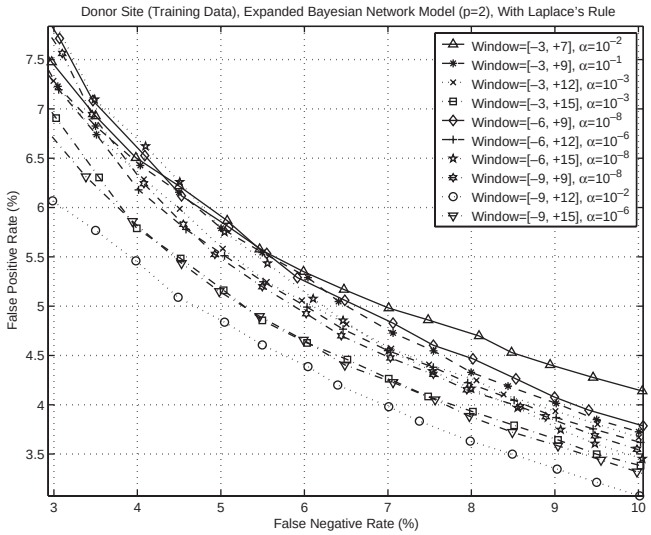

Fig. 4. Comparison of predictive accuracy of the 10 expanded Bayesian network models for the training data of the donor site corresponding to the 10 different windows.

Then keeping the reasonable complexity in mind, we examined 10 extensions of the consensus region of the donor site to select a proper window for the donor site. We compared the predictive accuracy of the corresponding 10 expanded Bayesian network predictive models with $p = 2$ for the training data and the testing data of the donor site, as shown in Figures 3 and 4, respectively. In this comparison, we used the best choice of Type I error for each window in the construction of the dependency graph for the donor site. Although the window 9 bases upstream to 12 bases downstream of the exon/intron boundary had the best predictive performance for the training data of the donor site, the window 9 bases upstream to 9 bases downstream of the exon/intron boundary had the best predictive performance for the testing data of the donor site. Considering the computational complexity, we selected the window nine bases upstream to nine bases downstream of the exon/intron boundary as the right choice for the donor site (in this case, the best Type I error α is 10^{-8}). We also examined the same 10 candidates of window under WMM, WAM, MDD, Cai *et al.*'s tree model, the 2MC and third-order Markov chain (3MC) models, and the expanded Bayesian network models with $p = 1$ and $p = 3$ and the best window for each model was determined.

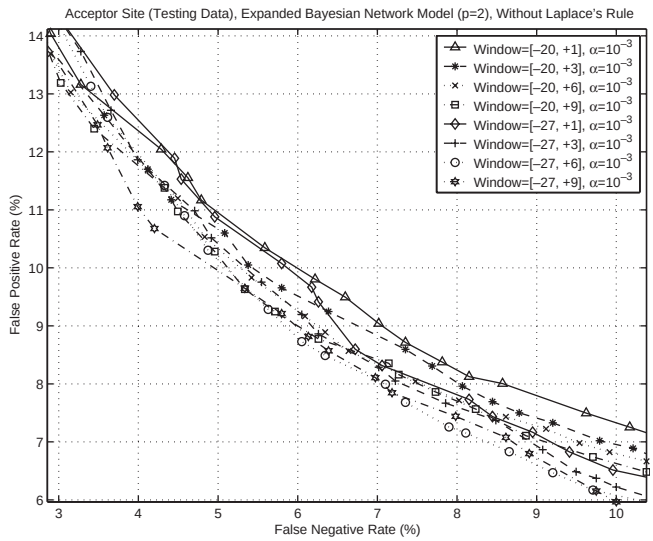

Fig. 5. Comparison of predictive accuracy of the eight expanded Bayesian network models for the testing data of the acceptor site corresponding to the eight different windows.

We also examined eight extensions of the consensus region of the acceptor site and compared the predictive accuracy of the corresponding eight expanded Bayesian network predictive models with $p = 2$ for the training data and the testing data of the acceptor site, as shown in Figures 5 and 6, respectively. In this comparison, we also used the best choice of Type I error for each window in the construction of the dependency graph for the acceptor site. It is apparent that the window 27 bases upstream to 9 bases downstream of the intron/exon boundary is the most suitable one for the acceptor site (in this case, the best Type I error α is 10^{-3}). Similarly, we examined the same eight candidates of window with WMM, WAM, MDD, Cai *et al.*'s tree model, the 2MC and 3MC models, and the expanded Bayesian network models with $p = 1$ and $p = 3$ and selected the best window for each model.

Laplace's rule

When the training dataset is not large enough, some probability parameters in a probabilistic predictive model will be

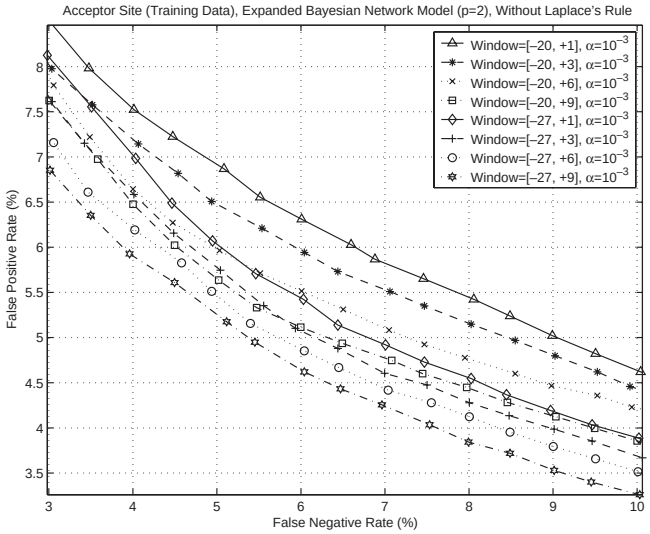

Fig. 6. Comparison of predictive accuracy of the eight expanded Bayesian network models for the training data of the acceptor site corresponding to the eight different windows.

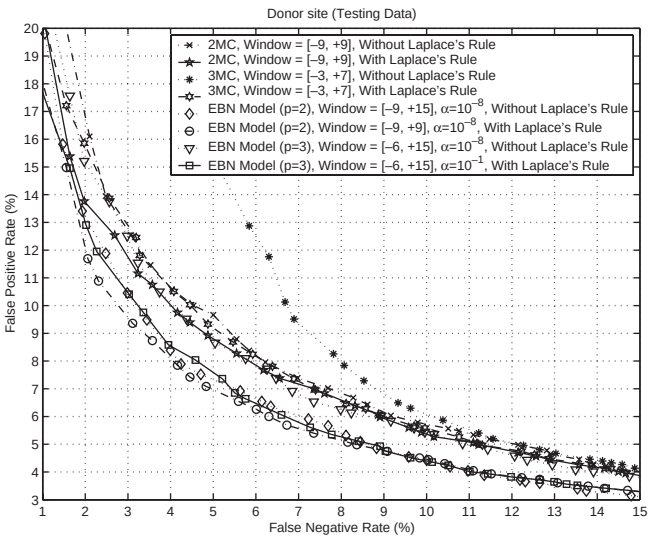

Fig. 7. Comparison of predictive accuracy for the testing data of the donor site using the 2MC and 3MC models, and the expanded Bayesian network models with $p = 2, 3$ with/without the Laplace's rule.

estimated as zeroes due to the non-existence of the corresponding base configurations in the training dataset and the predictive accuracy of the probabilistic model may be diminished. This often occurs when a higher-order Markov chain model or an expanded Bayesian network with higher p is built. One well-known approach to cope with this problem is to derive the relative frequencies by adding some fake extra counts to the true counts observed for each base configuration (Durbin *et al*., 1998). An extra count for each base configuration is called a pseudocount. The simplest pseudocount

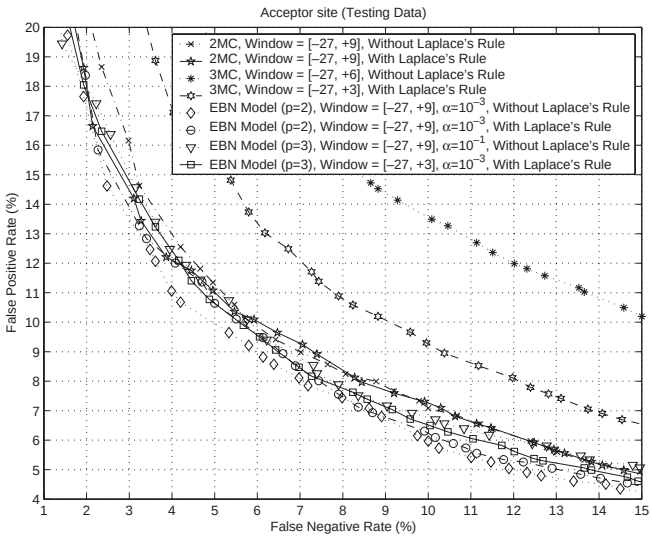

Fig. 8. Comparison of predictive accuracy for the testing data of the acceptor site using the 2MC and 3MC models, and the expanded Bayesian network models with $p = 2, 3$ with/without the Laplace's rule.

method is Laplace's rule: to add one pseudocount for each base configuration. In Figures 7 and 8, it is observed that the predictive accuracy of the 3MC model for splice sites is not acceptable without using the Laplace's rule and improves markedly with the use of Laplace's rule. The predictive accuracy of the expanded Bayesian network model with $p = 3$ is much better with the Laplace's rule than without it for the donor site, but only slightly better for the acceptor site. The predictive accuracy of the 2MC model, and the expanded Bayesian network model with $p = 2$ remains almost the same with or without the Laplace's rule while both models performed slightly better with the Laplace's rule, except that the expanded Bayesian network model with $p = 2$ performs better without the Laplace's rule for the acceptor site. For determining the best Type I error and/or the best window for each model in the previous subsections, we have compared the predictive accuracy with or without the Laplace's rule.

Predictive accuracy comparison

The two predictive accuracy measures, FN rate and FP rate, are reported for WMM, WAM, MDD, Cai *et al*.'s tree model, the 2MC and 3MC models, and the expanded Bayesian network models with $p = 1, 2, 3$ for the donor site and the acceptor site are shown in Figures 9 and 10 and Figures 11 and 12, respectively.

For the testing splice site data shown in Figures 9 and 11, the predictive accuracy of the expanded Bayesian network model with $p = 2$ (EBN2) was superior to that of all the other predictive models in all the cases examined, except for false positive rates $\geq$12% for the donor site and $\geq$17% for the

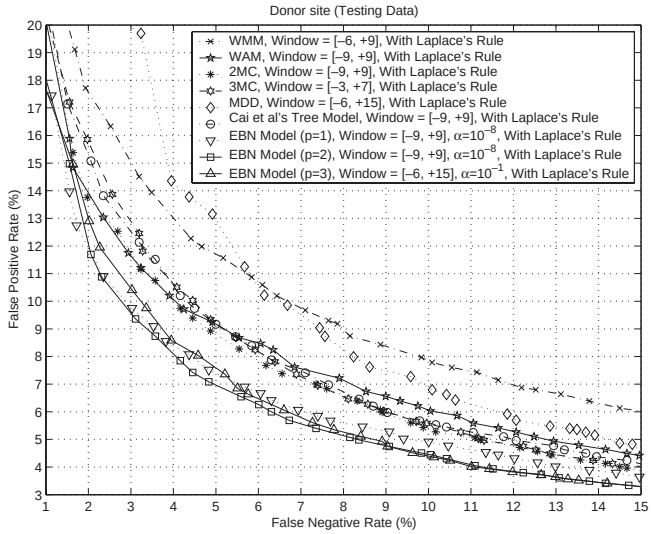

Fig. 9. Comparison of predictive accuracy for the testing data of the donor site using the WMM, WAM, MDD, Cai *et al*.'s tree model, the 2MC and 3MC models, and the expanded Bayesian network models with $p = 1, 2, 3$.

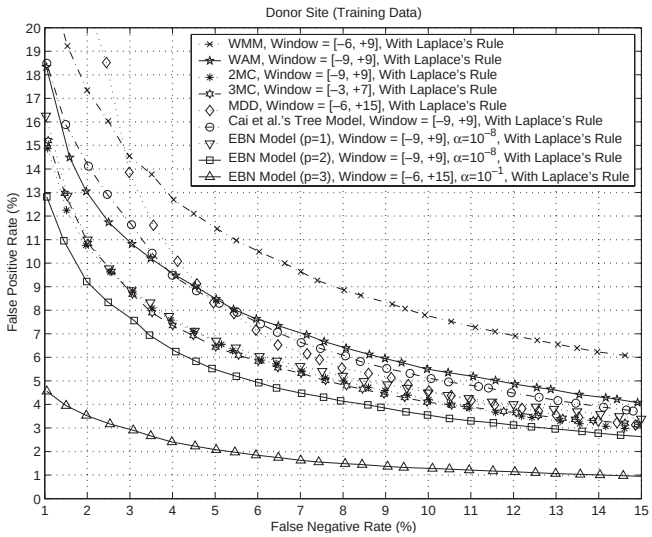

Fig. 10. Comparison of predictive accuracy for the training data of the donor site using the WMM, WAM, MDD, Cai *et al*.'s tree model, the 2MC and 3MC models, and the expanded Bayesian network models with $p = 1, 2, 3$.

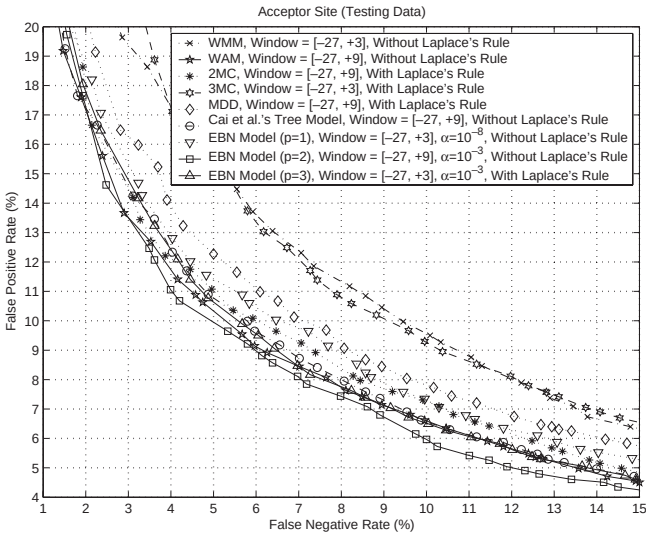

Fig. 11. Comparison of predictive accuracy for the testing data of the acceptor site using the WMM, WAM, MDD, Cai *et al*.'s tree model, the 2MC and 3MC models, and the expanded Bayesian network models with $p = 1, 2, 3$.

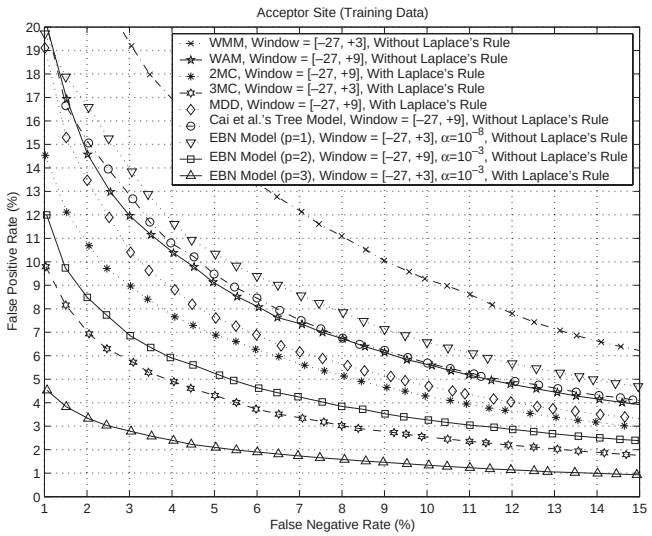

Fig. 12. Comparison of predictive accuracy for the training data of the acceptor site using the WMM, WAM, MDD, Cai *et al*.'s tree model, the 2MC and 3MC chain models, and the expanded Bayesian network models with $p = 1, 2, 3$.

acceptor site, respectively, where EBN2 was just among the best ones. For the training splice site data shown in Figures 10 and 12, the predictive accuracy of the expanded Bayesian network model with $p = 3$ (EBN3) is superior to that of all the other predictive models in all the cases examined. Note that while the predictive accuracy of EBN3 was superior to that of EBN2 for the training data, it was inferior for the testing data, which shows that EBN3 may overfit the splice site datasets. Also note that EBN2 outperforms the expanded Bayesian network model with $p = 1$ (EBN1) for almost all the cases examined. In particular, the sensitivity/specificity of EBN2 can reach up to 94%/94% for the testing data and 96%/94% for the training data of the donor site, and 93%/92% for the testing data and 95%/95% for the training data of the acceptor site.

We next compared the predictive accuracy of the Markov chain models considered. First, we observed their predictive accuracy for the donor site. For the testing data, the 2MC model has the best predictive accuracy among all the Markov chain models while WMM (which can be regarded as the 0MC model) has the worst. WAM (which is the 1MC model) and 3MC model are slightly inferior to 2MC with WAM approaching 2MC when specificity decreases and 3MC approaching 2MC when specificity increases. For the training data, 2MC and 3MC have almost the same predictive accuracy and are superior to WAM, which is in turn superior to WMM. Apparently, 3MC overfits the donor site dataset, and 2MC had the best predictive accuracy for the donor site among all the Markov chain models.

Second, we observed the predictive accuracy of Markov chain models for the acceptor site. For the testing data, WAM has the best predictive accuracy among all the Markov chain models. WMM and 3MC have comparable predictive accuracy and were much inferior to WAM, while 2MC was slightly inferior to WAM. For the training data, the order of predictive accuracy of the Markov chain models was apparent with 3MC being the best, followed by 2MC and WAM, and WMM being the worst. Apparently, both 2MC and 3MC overfit the acceptor site dataset and WAM had the best predictive accuracy for the acceptor site among all the Markov chain models.

We compared the predictive performance between the Markov chain models and the expanded Bayesian network models. For the testing data of the donor site, the predictive accuracy of all the Markov chain models was inferior to that of all the expanded Bayesian network models with $p = 1, 2, 3$. For the testing data of the acceptor site, the predictive accuracy of WAM was second only to that of EBN2 (and better than all the other models, including MDD and Cai *et al*.'s tree model). The EBN3 has slightly inferior predictive accuracy compared to WAM, only when the FN rate $\leq$7%. The 2MC was slightly inferior to EBN3 but slightly superior to EBN1. The WMM and 3MC have much inferior predictive accuracy. Now for the training data of the donor site, the order of predictive accuracy is EBN3 $>$ EBN2 $>$ 2MC $\approx$ 3MC $\approx$ EBN1 $>$ WAM $>$ WMM from the best to the worst. And for the training data of the acceptor site, the order of predictive accuracy is EBN3 $>$ 3MC $>$ EBN2 $>$ 2MC $>$ WAM $>$ EBN1 $>$ WMM.

For the prediction of the testing donor site data, MDD and WMM are the worst two models, whereas MDD is better than WMM when specificity is high and worse when specificity is low. For the prediction of the testing acceptor site data, MDD is superior to the worst two models WMM and 3MC and inferior to all other models. For the prediction of the training donor site data, MDD approaches 2MC as specificity increases but becomes the worst as specificity decreases. For the prediction of the training acceptor site data, MDD is superior to WAM but inferior to 2MC.

For the testing data of the donor site, the predictive accuracy of Cai *et al*.'s tree model is slightly inferior to that of 2MC, almost the same as that of 3MC, and slightly better than that of WAM when the false positive rates are $<$9% and slightly worse when the false positive rates are $>$9%. For the testing data of the acceptor site, the predictive accuracy of Cai *et al*.'s tree model is slightly inferior to that of WAM but becomes almost the same when the false positive rates are $<$7%. For the training data of the donor site and the acceptor site, Cai *et al*.'s tree model has about the same predictive accuracy with WAM. These results match and validate the observations made by Cai *et al*. (2000) between WAM and Cai *et al*.'s tree model. However, we observed that the splice site predictive accuracy of Cai *et al*.'s tree model is inferior to EBN2.

DISCUSSION

In this study, we developed a dependency graph model to fully capture the intrinsic cyclic interdependency between base positions in a splice site. Each dependency graph is expanded into a Bayesian network to facilitate the learning of a machinery for determining whether a tested potential splice site is real or pseudo. Compared with the previously published splice site models, such as WMM, WAM, MDD, Cai *et al*.'s tree model and the less-studied 2MC and 3MC models, this approach for the modeling of splice sites achieves the best results for all interesting cases, under the two predictive accuracy measures of FN rate and FP rate as shown in Figures 9–12.

The representation of the donor (acceptor) site by a window around the exon/intron (intron/exon) boundary has been studied extensively. We found that the window from 9 bases upstream to 9 bases downstream of the exon/intron boundary best represents the donor site and the window from 27 bases upstream to 9 bases downstream of the intron/exon boundary best represents the acceptor site.

The interdependency between base positions in the representative window of a splice site can be seen from the dependency graphs of the donor and the acceptor splice sites. As shown in Figure 1, strong interdependency among bases D_{-3}, D_{-2}, D_{-1}, D_{+3}, D_{+4}, D_{+5}, D_{+6}, D_{+7} of the donor site was observed and conforms to the consensus region of the donor site as indicated in Table 4. This implies that the spliceosome binds the donor site mainly on the bases downstream of the exon/intron boundary which conforms to our biological knowledge derived from experiments. Similarly, as inferred from Table 3, strong interdependency among bases from A_{-27} to A_{-3} is observed and conforms to the consensus region of the acceptor site as indicated in Table 5. This implies that the spliceosome binds the acceptor site mainly on the bases upstream of the intron/exon boundary, which too conforms to our biological knowledge derived from experiments.

ACKNOWLEDGEMENTS

We thank Chao-Chung Chang and Chen-Wei Hsu for their computer programming skills in the automatic construction of dependency graphs and their expanded Bayesian networks and in the automatic execution of the predictive models studied. We thank the editor and the reviewers for their valuable suggestions to improve the presentation of this paper. This work was supported by a grant (NSC91-3112-B-007-003) from the National Research Program of Genomic Medicine, National Science Council, Taiwan.

REFERENCES

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

Brunak,S., Engelbrecht,J. and Knudsen,S. (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.*, **220**, 49–65.

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

Cai,D., Delcher,A., Kao,B. and Kasif,S. (2000) Modeling splice sites with Bayes networks. *Bioinformatics*, **16**, 152–158.

Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Protein and Nucleic Acids*. Cambridge University Press, Cambridge, MA.

Ewens,W.J. and Grant,G.R. (2001) *Statistical Methods in Bioinformatics: An Introduction*. Springer-Verlag, NY.

Hebsgaard,S.M., Korning,P.G., Tolstrup,N., Engelbrecht,J., Rouzé,P. and Brunak,S. (1996) Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.*, **24**, 3439–3452.

Henderson,J., Salzberg,S. and Fasman,K. (1997) Finding genes in human DNA with a hidden Markov model. *J. Comput. Biol.*, **4**, 127–141.

Khodarev,N.N., Park,J., Kataoka,Y., Nodzenski,E., Khorasani,L., Hellman,S., Roizman,B., Weichselbaum,R.R. and Pelizzari,C.A. (2003) Receiver operating characteristic analysis: a general tool for DNA array data filtration and performance estimation. *Genomics*, **81**, 202–209.

Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C. and Baldwin,J. Devon, K., Dewar, K., Doyle, M., and FitzHugh, W. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Mathe,C., Sagot,M., Schiex,T. and Rouzé,P. (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.*, **30**, 4103–4117.

Mathews,C.K., van Holde,K.E. and Ahern,K.G. (2000) *Biochemistry*, 3rd edn. Addison Wesley Longman, San Francisco, CA.

Pearl,J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.

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

Reese,M.G., Eeckman,F.H., Kulp,D. and Haussler,D. (1997) Improved splice site recognition in Genie. *J. Comput. Biol.*, **4**, 311–324.

Salzberg,S., Delcher,A., Fasman,K. and Henderson,J. (1998) A decision tree system for finding genes in DNA. *J. Comput. Biol.*, **5**, 667–680.

Staden,R. (1984) Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.*, **12**, 505–519.

Tolstrup,N., Rouzé,P. and Brunak,S. (1997) A branch point consensus from *Arabidopsis* found by non-circular analysis allows for better prediction of acceptor sites. *Nucleic Acids Res.*, **25**, 3159–3163.

Weaver,R.F. (1999) *Molecular Biology*. WCB McGraw-Hill, NY.

Yeo,G. and Burge,C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, **11**, 377–394.

Zhang,M.Q. (2002) Computational prediction of eukaryotic protein-coding genes. *Nat. Rev. Genet.*, **3**, 698–709.

Zhang,M.Q. and Marr,T.G. (1993) A weight array method for splicing signal analysis. *Comput. Appl. Biosci.*, **9**, 499–509.