A Method for Identifying Splice Sites and Translational Start Sites in Eukaryotic mRNA

Steven L. Salzberg
Department of Computer Science
224 New Engineering Building
Johns Hopkins University
Baltimore, MD 21218 USA
salzberg@cs.jhu.edu

Abstract

This paper describes a new method for determining the consensus sequences that signal the start of translation and the boundaries between exons and introns (donor and acceptor sites) in eukaryotic mRNA. The method takes into account the dependencies between adjacent bases, in contrast to the usual technique of considering each position independently. When coupled with a dynamic program to compute the most likely sequence, new consensus sequences emerge. The consensus sequence information is summarized in conditional probability matrices which, when used to locate signals in uncharacterized genomic DNA, have greater sensitivity and specificity than conventional matrices. Species-specific versions of these matrices are especially effective at distinguishing true and false sites.

1 Introduction

As part of an automated system for finding coding regions in uncharacterized eukaryotic DNA, we have developed a new method for finding the signals that indicate the start of translation, the beginnings of introns (donor sites), and the ends of introns (acceptor sites). The basis of the method is the computation of conditional probabilities for each of the four bases that comprise DNA[1] in a fixed set of positions around each site. The standard method, by contrast, computes the probabilities of the bases in each position as if they were independent of adjacent bases. Instead, the new method is to compute, for each position, the probability of each base *given* the base in the previous position, where "previous" is defined as the adjacent base in the 5' direction. In the consensus pattern that emerges, the identity of each base is dependent on its neighbors.

The resulting conditional probability (CP) matrices indicate that for several positions in all three types of sites (start of translation, donor, and acceptor sites) the probability of a base occurring in a given position is sometimes strongly dependent on the previous base. This has a natural biological explanation, in that the mechanisms responsible for translation and splicing involve molecules that recognize and bind to sets of adjacent bases in the mRNA. Unlike matrices in which all probabilities are independent of adjacent positions, the consensus sequence cannot simply be "read off" by choosing the highest probability in each column. Instead, a dynamic program can be used to generate a consensus sequence from a CP matrix. The sequence that is produced by the program is the most likely sequence given the data in the matrix.

To generate the new consensus matrices and sequences, we have used a data set of 570 complete vertebrate coding sequences in conjunction with the CP matrices to generate new consensus sequences for the start site, donor site, and acceptor site in vertebrate DNA. These patterns are different from the patterns that would be produced by a standard consensus matrix and, mathematically speaking, are more likely than the patterns generated by a matrix of independent probabilities. For comparison, we have also generated traditional consensus matrices, which correspond closely to previously published matrices. Traditional consensus matrices tabulate the probability of each base b at position i, $P(b, i)$. The use of conditional probabilities requires the tabulation of $P(b_1, i|b_2, i-1)$, i.e., the probability of base b_1 in position i given base b_2 in the previous position. These matrices form the basis of an algorithm for signal detection that is equivalent to a first-order Markov chain method.

Many previous studies have attempted to characterize the sequences around the start, donor, and acceptor sites in eukaryotic organisms. Kozak [1, 2] has produced several comprehensive studies of the consensus sequence around the start of translation, and introduced the scanning model of ribosome progression that is now widely accepted. Senapathy et al. [3] and Mount et al. [4], among others, have characterized the sequence patterns around splice junctions, and more recent results have

[1] Although splicing and translation occur in mRNA, here we follow the convention of expressing sequences and sequence patterns in terms of ACGT rather than ACGU.

described the patterns and processes for non-consensus (AT-AC) splicing [5]. The consensus sequences uncovered in these previous studies have been most frequently described as matrices containing the probabilities of the four bases in the positions immediately surrounding the sites.

Specific computational systems for identifying splice junctions have been developed by many previous researchers. Brunak et al. [6] used a neural network that considered positional frequencies, binding energies, and other coding measures. Zhang and Marr [7] used a combination of features including dinucleotide frequencies to identify donor sites in *S. pombe*, and found evidence for pairwise correlations in the signals. Solovyev et al. [8] identify splice sites with a linear discriminant that combines triplet counts, octamer frequencies, G, GG, and GGG counts, and other measures. Because correct identification of sequence signals is a critical component of gene-finding systems, some of those systems have explicit models of start sites, donor sites, and acceptor sites enbedded in their algorithms. GeneID [9] uses a model of independent base probabilities around each site, as does GeneParser [10, 11].

Fickett [12] provides a recent review of the main work on identifying signals (both splice junctions and translation start sites), and points out that the best previous results used a combination of different types of evidence, both from the bases immediately surrounding the site and from sequences extending some distance away from the site. Fickett also describes a common variation on the consensus matrix known as the *position weight matrix*, which uses $P(b,i)/P(b)$ instead of just $P(b,i)$. This gives the relative frequency of each base in each position, and may offer advantages in regions of high or low GC content.

The main results of the current study, while consistent with Fickett's conclusion that the best methods combine many different coding measures, offers an alternative point of view. In the experiments below, we used only one coding measure, the conditional probability of the bases in the immediate vicinity of the splice and start sites, and obtained accuracies that are surprisingly good, given the limited information used. While it is true that higher accuracies may be obtained by using information from a larger window around the site [8], the bulk of the signal recognition ability comes from the local information. In addition, non-local coding measures are already used elsewhere in gene assembly programs [9, 11, 13], and the benefits of that information are reflected in better performance overall on the gene-finding problem. The results here show that, when it comes to identifying signals in DNA sequences, the majority of the required information is contained locally in the sequence pattern itself. This makes good sense biologically, since the translation and splicing machinery seems to operate primarily on mRNA that is near the sites.

Another conclusion of this study is that conditional probabilities are consistently better than independent positional probabilities, and sometimes strikingly so. This should prove very useful in gene-finding systems or other systems that use positional weight matrices. Our group's gene-finding systems VEIL [14] and MORGAN [15] already use conditional probabilities for site detection. Recently, the Genie system was modified to use dinucleotide probabilities instead of independent probabilities, and significant improvements in its overall gene-finding accuracy were reported based on this change alone [16]. Finally, based on the improvements shown below from species-specific matrices, one can recommend that future efforts to characterize splice junctions and start sites should emphasize the collection of large, high-quality datasets for each organism of interest.

2 Methods and Results

2.1 Sequence Data

The data for this study were originally collected by Burset and Guigo [17], who used it to study gene-finding systems. They collected a large set of genes, and carefully edited the set so as to remove sequences that were likely to contain erroneous annotation or to represent non-standard splicing mechanisms (e.g., alternative splicing or AT-AC splicing). Database entries were discarded if the protein coding region was ambiguous, if the sequences included pseudogenes, if alternatively spliced forms of the gene were listed, or if the gene contained no introns. This produced 1410 sequences, which were then further refined by discarding sequences whose protein coding region did not start with ATG, whose length was not a multiple of three, whose introns did not start with the dinucleotide GT and end with AG, or that had protein coding regions with an in-frame stop codon. (By using only "standard" GT-AG introns, the statistical methods for characterizing the splice sites should produce a clearer picture of the consensus pattern.) Sequences corresponding to immunoglobulins and histocompatibility antigens were discarded, and finally all sequences entered prior to January 1993 were discarded, leaving only relatively recent entries.

The resulting data set contains 570 complete protein coding sequences, each comprising one gene with at least one intron. The sequences contain a total of 2649 exons and 2079 introns. For the purpose of characterizing start sites, 562 patterns are available, because 8 of the sequences contain less than 4 bases prior to the start of translation. Because every intron has both a donor and acceptor site, there are 2079 subsequences available to compute the consensus for these sites.[2]

[2]The data set is available by ftp from ftp.cs.jhu.edu in the directory pub/salzberg/sitedata.

2.2 Conditional Probability Matrix Computation

To compute a matrix of conditional probabilities, the columns of the matrix are defined to be positions on either side of the site of interest. For example, the start site matrix in this study uses positions from -12 through +6 (12 positions upstream of the start codon through 4 positions downstream, where the start codon itself occupies positions 0–2). Each entry in the matrix contains the conditional probability of base x in position i given that base y is in position $i-1$ (the previous position), which is computed as:

$$P(x_i|y_{i-1}) = P(x_i \wedge y_{i-1})/P(y_{i-1})$$

A simple dinucleotide frequency count gives the probability of x and y occurring together in adjacent positions $i-1$ and i. For the first column only, the matrix contains the independent probabilities of the four bases, and all the remaining columns contain conditional probabilities. Note that the conditional probability matrix is *not* equivalent to a dinucleotide probability matrix; in a separate study (data not provided), dinucleotides were found to be inferior to conditional probabilities at identifying splice sites. (The dinucleotide matrix contained $P(x_i \wedge y_{i-1})$ for each position. These probabilities, rather than the conditionals, were multiplied together to produce a score, using the algorithm given in Section 2.4.)

Tables 4,6, and 8 are the conditional probability matrices for the start sites, donor sites, and acceptor sites for the 570 vertebrate sequences. For comparison, Tables 5, 7, and 9 show the standard consensus matrices for the same set of sequence data. In these matrices, each entry represents the independent probability of a base occurring in that position. Note that the 570 sequences used to compile these tables contain a significant fraction of closely homologous sequences. No attempts were made to remove homologies, and a systematic analysis of how homology changes the statistics is beyond the scope of this paper. It is clear from even a cursory examination of the tables that the when the probability of a base is conditioned on the previous position, it often changes dramatically. For example, consider Tables 6 and 7 in position +3. In Table 7, adenine is observed to appear in that position in 71% of the donor sites. However, in Table 6 the picture is more complicated: if position +2 contains adenine, cytosine, or guanine, then adenine is still the most likely base at +3, but when position +2 contains thymine, then guanine follows in position +3 a full 63% of the time, while adenine's probability drops to 19%. Further examination of the CP matrices reveals numerous instances of this type of dependency between adjacent bases.

2.3 Consensus Sequences for Start, Acceptor, and Donor Sites

To find the consensus sequence from a conventional matrix, one can simply read it off by noting the highest probability that appears in each column. In a conventional matrix, each column contains just four probabilities, and the highest probability in the column is the base most likely to appear in that position. However, the consensus sequence cannot be read directly from a CP matrix. One can instead compute the most likely sequence, which is not the same as the consensus — the consensus can be thought of as the "typical" pattern for a site. For example, if a position contains thymine 40% of the time and cytosine 38% of the time, the most likely base is T, but the consensus is better represented as Y (pyrimidine).

To discover the most likely sequence from one of the CP matrices, one must use a dynamic program that finds the most probable path from left to right through the matrix P. The idea is that we must compute, for each position, the probability of the most likely sequence ending in A, C, G, and T respectively. This gives us a column of a new matrix M, where each of the n columns of M contains the probability of the most likely sequence ending in one of the four bases. To extend M by one more column, we use the CP matrix P to find the probability of the best path ending in A that had one of (A, C, G, T) in the previous location. More formally, M can be computed by:

$$M_{b,j} = \begin{cases} P(b = a, c, g, t) & \text{for } j = 1 \\ \max\limits_{b,x \in (a,c,g,t)} P(b_j|x_{j-1})(M_{x,j-1}) & \text{for } j > 1 \end{cases}$$

In this notation, $M_{b,j}$ refers to the row of M that corresponds to the base b, and j is the column corresponding to the j^{th} position of the pattern.[3]

The consensus sequences for the vertebrate data set, aligned with the most likely sequence from the conditional probability matrices shown in Tables 4–9, are:

```
Start site
CP matrix      C A A A C A G A C A C C ATG G T G
Indep. matrix  C * A C C * G C C A C C ATG G * G
Kozak, 87        (G C C)G C C R C C ATG G
Donor site
CP matrix        C A G GT G A G T G G G G G G
Indep matrix   A/C A G GT R A G T
Senapathy et al. A/C A G GT R A G T
Acceptor site
CP matrix      T T T T C T C T T T G C AG G
Indep. matrix  Y Y Y T Y Y Y Y Y Y * C AG G
Senapathy et al. T Y Y Y Y Y Y Y Y Y * C AG G
```

In the patterns, "Y" means either C or T (pyrimidine) and "R" means A or G (purine). Positions

[3] A C program to compute M and generate the most likely sequence from a CP matrix is available by ftp to ftp.cs.jhu.edu, in the file pub/salzberg/matrixdp.c.

where no base occurred at least 33% of the time are marked by *. The sequences shown for the CP matrix does not have *'s because these are the *most likely* sequences. The consensus patterns from the CP matrices are nearly identical to Kozak's consensus sequence, (GCC)GCC(A/G)CCATGG [1] and to Senapathy et al.'s [3] reported donor and acceptor sequences. The only difference is in the (GCC) at position -9 in the start sequence, where the new CP matrix indicates ACA. Here, despite the fact that cytosine has a higher independent probability than adenine at positions -9, -7, and -5, the most likely sequence has adenine in those positions when pairwise dependencies are taken into account.

The similarity between the consensus patterns produced by CP and independent probability matrices confirms that the CP matrices, despite their dramatic differences in the details of their entries, capture similar summary information about the patterns used to create them. However, as we discuss next, they can produce substantial improvements when used in signal detection methods.

2.4 Detecting Signals with CP Matrices

Consensus matrices can be used for signal detection in the following manner. For any pattern of anonymous DNA, one must compute a score based on its probability of being a true instance of a start, donor, or acceptor site. This score can be compared to the scores of known true sites to determine if the anonymous pattern is also a true site. Independent probability matrices have been used in this manner in a number of well-known gene finding systems, including GeneParser [11] and GeneID [9], and in the newer system MORGAN [15]. Very recently, Reese et al. [16] changed the splice site recognition function in the Genie gene-finding system from independent probability matrices to dinucleotide probabilities, and they report a significant increase in overall accuracy from this change alone.

The scoring function estimates the probability that a new sequence is a true site, which we can write as $P(T|S)$; i.e., the probability of a true site T given a sequence $S = (s_1, s_2, \ldots, s_n)$. A consensus matrix contains the probability of a sequence given that it is a true site, or $P(S|T)$ (this follows from the fact that only true sites are used to create these matrices). Thus to compute $P(T|S)$, we use Bayes' Law:

$$P(T|S) = P(S|T)P(T)/P(S)$$

When comparing a set of patterns to detect true sites, we can treat the underlying prior $P(T)$ as a constant. $P(S)$ is normally estimated by multiplying the individual base probabilities for $s1, s2, \ldots, s_n$, and $P(S|T)$ is the product of the entries in the matrix. Note that this approach, because it multiplies the individual probabilities of the sequence of bases, implicitly assumes that these probabilities are independent.

For conditional probability matrices, the scoring function is similar, with the difference being that the score $P(S|T)$ in the CP matrix is really a 1-state Markov chain model, computed by multiplying the conditional probabilities of each successive base, given the previous base in the sequence. Thus the CP matrix takes into account the dependencies between adjacent bases in the sequence. When estimating $P(S)$, we use the 16 prior conditional probabilities for each base given the four possible bases in the previous position. We then compute $P(S)$ as

$$P(s_1)\prod_{i=2}^{n} P(s_i|s_{i-1})$$

The 16 priors can be computed based on the entire data set, or on each coding sequence separately. Experiments using both methods (data not shown) revealed that using the entire data set to compute the priors was superior, so this method was used for the experiments below.

We compared the two scoring methods, conventional matrices and CP matrices, as follows. First, all true sites from the data set were scored using both methods. Then these scores were sorted to determine a detection threshold.[4] For example, if the lowest-scoring true site is used to set the threshold, then no true sites will be missed, giving a sensitivity of 100% (equivalent to a false negative rate of 0%). This threshold is then used for every other subsequence in the data set, which contains 2.88 million bp, to determine how many false sites will score above the threshold (the false positive rate).

Table 1 shows the signal detection rates for start sites, donor sites, and acceptor sites on the complete set of vertebrate sequences. The left side of the table contains sensitivity and false positive rates for conditional probability matrices and the right side shows the same values for conventional matrices. Sensitivity is defined as the probability of correctly identifying a true site, and the false positive rate (FP) is 1 minus the probability of correctly rejecting a site as false. The table reports the number of true sites missed (1−Sensitivity) and the number of false sites that passed the threshold (false positives). To provide a further comparison, the table also gives the correlation coefficient (CC) for each threshold, computed as

$$\frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FN) * (TN + FP) * (TP + FP) * (TN + FN)}}$$

Note that maximizing the CC is not the right way to set thresholds for problems such as this, where there are far more negative examples than positive ones. Thus a CC of 0.51 can be obtained using conditional probabilities if one is willing to miss 30% of true donor sites, but if this matrix is used within a gene finding system, the threshold should probably be set to miss as few true sites as possible.

[4] The sorted scores and thresholds are available by ftp in the same directory as the complete data set, at ftp.cs.jhu.edu in the directory pub/salzberg/sitedata.

For each line in the table, thresholds were set so that increasing numbers of true start sites would be missed, and these same thresholds were then used against the complete database. For example, in the second line of Table 1, a threshold was set using the CP matrix that would correctly identify 545/562 true start sites, missing only 17. This same threshold would lead to 15,307 other sites being labeled as true sites, which is 0.53% of the total. Another threshold was set using the conventional matrix so that it too would identify 545 true start sites. This threshold would then lead to 20,346, or 0.71% of other sites being incorrectly labeled. The data from Table 1 is shown graphically in Figures 1–2, which illustrate that the CP matrix method consistently beats the conventional matrices for any level of sensitivity.

The table and figures show that CP matrices give a consistent advantage over conventional matrices. For start sites, CP matrices give a 25-35% reduction in false positives for a given sensitivity level. For donor sites, the differences range from 30% to 50%, while for acceptor sites, CP matrices provide a benefit ranging from 15% to over 30%.

The results above use a single data set, which gives an optimistically biased estimate of the benefit of CP matrices. As a more stringent test of these results, the data were divided into separate training and test sets. 456 sequences (80%) were randomly selected for training, and the remaining 114 (20%) were used for testing. All the conditional probability matrices were re-constructed using only the training data, and thresholds were set using the training data. The same thresholds were then used for the test data to measure the false negative and false positive rates. The results are given in Table 2, and shown graphically in Figures 3–5. Not surprisingly, the threshold setting for false negatives (true sites missed) on the training data was sometimes accurate, but sometimes inaccurate at estimating the false negative rate on the test data, for both types of matrices. However, the main purpose of this additional experiment is to see if the difference between CP and conventional matrices holds when the matrices are used on a separate test set. As the table shows, the false positive rate is 20-40% lower for most threshold settings, and the CC is always lower as well. Although the differences are not as great as in 1, the CP matrices still show a consistent improvement over independent probability matrices.

Finally, we investigated how these differences change when one uses a database from a single species. It turns out that the CP matrices provide an even greater advantage if the sequences come only from human, rather than from a wide range of vertebrates. We extracted a subset of 93 human sequences from the original 570 sequences and repeated the experiments above. Experimental results for this subset, which contains 606,097bp, 439 exons, and 346 introns, are shown in Table 3. There are two improvements noticeable in the human-specific table. First, the false positive rate (FP) for all three types of sites improves substantially when compared to conventional matrices matrices. The improvement is a factor of four for 100% sensitivity for start sites, where the number of false positives fell from 8499 to 2256. At less sensitive levels, this difference is even greater in some cases. For donor sites the differences are not as large, but at some sensitivity levels the number of false positives is almost cut in half.

Second, a more striking difference can be observed by comparing the numbers for CP matrices only between Table 3 and Table 1. This comparison shows that the false positive rate of CP matrices for human-only data is much lower than for vertebrate data. For example, consider start site recognition at the 100% sensitivity level. Here there were 1% false positives in the vertebrate data, versus 0.37% for the human data. At 90% sensitivity, the false positive rate fell to 0.26% for vertebrates, while it fell even further, to 0.08%, for human sequences. Thus the species-specific improvement in the false positive rate seems to be around a factor of four. Note that Table 3, like Table 1, shows differences between the methods when using a single data set. Although the data set is too small to experiment with a separate test set here, these differences are likely to decrease on separate test data, as they did in Table 2.

The most likely sequence patterns from the human-only CP matrices can be computed using dynamic programming, just as in Section 2.3. These sequences, compared to those from the vertebrate data, are:

```
Start site
Human C A A A C A G A C A C C ATG G T G C
Vert.   C A A A C A G A C A C C ATG G T G C
Donor site
Human C A G GT G A G T G G C A A G G G G
Vert.   C A G GT G A G T G G G G G G G G G
Acceptor site
Human T T T T C C C C C C A C AG G
Vert.   T T T T C T C T T T G C AG G
```

Note that although the start sequence consensus is identical, the donor site has three differences and the acceptor site has five.

2.5 Comparisons

Although comparisons are difficult to make without using identical data sets, a very rough comparison might be informative. The splice site detection method of Solovyev et al. is reported to be the best known method, and their tests also used human-only data. Because their method was only used for donor and acceptor sites, and not for start sites, we will only compare those numbers here. A description of their algorithm is beyond the scope of this discussion, but in brief it is a straightforward linear discriminant function based on a set of complex features. The feature include: triplet composition in an 80-base window around donor sites, a 10-base consensus matrix, the number of G bases, GG pairs, and

GGG triplets in a 50-base region of the intron, and octanucleotide frequency measures for a 114-base window around the site. They use a similar set of features for acceptor sites. They report an overall accuracy for donor site prediction of 97%, with $CC = 0.63$. However, they do not give a breakdown into false positives and false negatives, and because the number of pseudo-sites is vastly greater than the number of true sites (97.8% of their test data was pseudo-sites), it is hard to compare their numbers to those reported here. One of their figures indicates that they obtained a 3% false positive rate for 96% sensitivity, which for their data would indicate approximately 900 false positives. For acceptor sites, they report a 4% false positive rate at 96% sensitivity, which would yield approximately 3600 false positives (they had a substantially more pseudo-acceptor sites).

The false positive rates in the Solovyev study only counted sites already containing a GT or AT as potential donor or acceptor sites. Table 3 counts all sites when computing false positive rates, so to make a rough comparison, false positive rates in Table 3 should be multiplied by 16. Thus at a sensitivity of 95%, the table shows that the CP matrices have a 5.6% false positive rate for pseudo-donor sites, and 9.9% for pseudo-acceptor sites. This is not quite as good as the linear discriminant method, but it is surprisingly close given how much less information is used. Clearly, though, some non-local information can be useful: for example, the branch site occurs some distance upstream of the 3' acceptor, and the local matrices do not capture this site. In addition, the coding region side of any site cannot contain in-frame stop codons, so the presence of stop codons can be used to rule out many false positives. Thus if the only goal is to identify splice sites, a method based on both local and non-local information should be used, but if a position weight matrix is being used in the context of a larger system, then the data presented here suggest that the matrix should be replaced with a CP matrix.

Although CC values are often used in comparisons, they are not the best standard to use here. As shown in the tables above, the highest CC values are obtained for relatively high false negative rates, because of the skewed composition of the data. For the same reason, overall accuracy is not a good indicator of performance either. For example, in the human sequences, at threshold level that missed 40% of the donor sites, the CC is 0.54 and the overall accuracy is 99.9%. But if the matrix is being used as part of a gene-finding effort, it might be too conservative to set the threshold so high. By combining a more generous threshold with other constraints, such as internal codon or hexamer frequencies and open reading frame requirements, one should be able to use the CP matrices to achieve better exon recognition than is currently being obtained by gene-finding methods that use independent position weight matrices. (For example, the MORGAN system [15] uses these CP matrices with a threshold that misses less than 1% of true sites, and uses other coding measures to distinguish between true exons and pseudo-exons with high accuracy.)

3 Discussion

The identification of sequence patterns is essential to understanding the machinery behind translation and splicing of mRNA. Identification of the most likely base at each position around a splice site is the first step in characterizing these patterns. The current study uses the growing amount of sequence data to go one step further towards characterizing splice sites. The conditional probability matrices computed in this study show numerous important dependencies between adjacent bases around start sites, donor sites, and acceptor sites. Although the overall consensus pattern changes only slightly with the use of these new matrices, the ability to detect true sites accurately improves substantially. As more data accumulates, it should be possible to refine these matrices further and develop even better methods for site recognition.

The results above indicate that the further improvements in splice site recognition can be had by construction of a species-specific conditional probability matrix. If there is not enough data available for a species, then a CP matrix encompassing a larger family of organisms is still preferable to a matrix of independent probabilities. As the amount of DNA sequence for all organisms grows, it should become possible to develop accurate matrices tailored to many individual species. Besides providing better characterizations of the sites, these matrices should also help to improve the performance of gene finding systems.

There are at least two possible explanations for the different performance of the human-only sequence patterns and the more general vertebrate sequence patterns. One is that the patterns are different simply because the human sequences are closer evolutionarily, and therefore have not diverged as much as the patterns across the complete data set. A second, more interesting explanation is that the mechanisms of splicing themselves may be slightly different in humans; i.e., there may be some specialized aspects of translational initiation and RNA splicing that are made evident in the sequences that appear in the genome. This latter question is an important issue for further investigation.

Acknowledgements

Thanks to Simon Kasif, Art Delcher, and the anonymous reviewers for many helpful suggestions. This material is based upon work supported by the National Science foundation under Grant No. IRI-9223591, and by the National Center for Human Genome Research at the National Institutes of Health under Grant No. K01-HG00022-1.

References

[1] M. Kozak. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Research*, 15(20):8125–8148, 1987.

[2] M. Kozak. A consideration of alternative models for the initiation of translation in eukaryotes. *Critical Reviews in Biochemistry and Molecular Biology*, 27:385–402, 1992.

[3] P. Senapathy, M.B. Shapiro, and N.L. Harris. Splice junctions, branch points, and exons: sequence statistics, identification, and applications to genome project. *Methods in Enzymology*, 183:252–278, 1990.

[4] S. Mount, C. Burks, G. Hertz, G. Stormo, O. White, and C. Fields. Splicing signals in *drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Research*, 20:4255–4262, 1992.

[5] S. Mount. AT-AC introns: An ATtACk on dogma. *Science*, 271(5256):1690–1692, 22 March 1996.

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

[7] M.Q. Zhang and T.G. Marr. A weight array method for splicing signal analysis. *Computer Applications in the Biosciences (CABIOS)*, 9(5):499–509, 1993.

[8] V.V. Solovyev, A.A. Salamov, and C.B. Lawrence. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.*, 22:5156–5163, 1994.

[9] R. Guigo, S. Knudsen, N. Drake, and T. Smith. Prediction of gene structure. *J. Mol. Biol.*, 226:141–157, 1992.

[10] E. E. Snyder and G. D. Stormo. Identification of coding regions in genomic DNA sequences: An application of dynamic programming and neural networks. *Nucleic Acids Research*, 21(3):607–613, 1993.

[11] E. E. Snyder and G. D. Stormo. Identification of coding regions in genomic DNA. *Journal of Molecular Biology*, 248:1–18, 1995.

[12] J. Fickett. The gene identification problem: an overview for developers. *Computers and Chemistry*, 20:103–118, 1996.

[13] Y. Xu, R. Mural, J.R. Einstein, M. Shah, and E. Uberbacher. Grail: A multi-agent neural network system for gene identification. *Proc. of the IEEE*, 84(10):1544–1552, 1996.

[14] J. Henderson, S. Salzberg, and K. Fasman. Finding genes in human DNA with a hidden Markov model. *Journal of Computational Biology*, 1997. to appear.

[15] S. Salzberg, X. Chen, J. Henderson, and K. Fasman. Finding genes in DNA using decision trees and dynamic programming. In *ISMB-96: Proc. Fourth Internatl. Conf. Intelligent Systems for Molec. Bio.*, pages 201–210, Menlo Park, CA, 1996. AAAI Press.

[16] M. Reese, F. Eeckman, D. Kulp, and D. Haussler. Improved splice site detection in genie. In *RECOMB 97*, pages 232–240. ACM Press, 1997.

[17] M. Burset and R. Guigo. Evaluation of gene structure prediction programs. *Genomics*, 34(3):353–367, 1996.

A Consensus Matrices

This appendix contains both conditional probability (CP) matrices and independent probability matrices for start sites, donor sites, and acceptor sites. All the probabilities are based on the same database of 570 vertebrate sequences. These matrices are all available electronically by ftp at the site ftp.cs.jhu.edu. Connect to the directory pub/salzberg and get the file sitematrices.h. The C code to compute the most likely sequence from one of these matrices is available at the same site, in the files cpmatrix.h and matrixdp.c.

Signal Detection (All Vertebrate Sequences)								
	True Sites Missed		CP Matrix			Conventional Matrix		
			False Sites Labeled True			False Sites Labeled True		
	Num	%	Num	%	CC	Num	%	CC
Start	0	0	28,021	0.97	0.14	35,791	1.24	0.12
Site	17	2.8	15,307	0.53	0.18	20,346	0.71	0.16
Detection	31	5.4	11,599	0.40	0.20	18,010	0.62	0.16
(562 true	56	10	7,438	0.26	0.24	11,282	0.39	0.20
sites)	169	30	2,442	0.08	0.31	3,666	0.13	0.26
	281	50	716	0.02	0.38	1,346	0.05	0.29
Donor	0	0	133,816	4.6	0.12	143,480	5.0	0.12
Site	3	0.1	46,051	1.6	0.21	81,536	2.8	0.16
Detection	29	1.4	22,437	0.78	0.29	34,146	1.2	0.23
(2079 true	104	5.0	11,772	0.41	0.37	18,471	0.64	0.30
sites)	208	10	7,753	0.27	0.42	12,668	0.44	0.34
	624	30	2,412	0.08	0.51	3,702	0.13	0.44
Acceptor	0	0	165,963	5.7	0.11	213,312	7.4	0.09
Site	3	0.1	104,695	3.6	0.14	163,228	5.6	0.11
Detection	30	1.4	36,436	1.3	0.23	52,447	1.8	0.19
(2079 true	103	5.0	21,410	0.74	0.28	29,496	1.0	0.24
sites)	208	10	14,796	0.51	0.32	17,623	0.61	0.29
	623	30	5,393	0.19	0.39	6,517	0.23	0.36

Table 1: Sensitivity and false positive rates of start, donor, and acceptor site detection for a range of different threshold values, using conditional probability (CP) matrices and conventional independent probability matrices.

	Training	Test Data					
		CP Matrix			Conventional Matrix		
	FN (%)	FN (%)	FP (%)	CC	FN (%)	FP (%)	CC
Start	0	1	1.05	0.13	0	1.30	0.12
Sites	1	1	0.86	0.14	0	1.21	0.12
(114 test	5	6	0.42	0.19	6	0.64	0.16
sites)	10	14	0.27	0.22	12	0.38	0.19
	20	27	0.14	0.25	20	0.22	0.23
Donor	0	0	4.74	0.11	0	5.10	0.11
Sites	1	0	0.84	0.26	2	1.28	0.21
(385 test	5	10	0.39	0.34	7	0.63	0.28
sites)	10	16	0.25	0.39	13	0.43	0.31
	20	27	0.14	0.43	27	0.22	0.36
Acceptor	0	0	5.59	0.10	0	7.47	0.09
Sites	1	1	1.45	0.20	0	2.15	0.17
(385 test	5	4	0.77	0.27	4	1.06	0.23
sites)	10	9	0.53	0.30	9	0.64	0.28
	20	23	0.28	0.34	21	0.35	0.31

Table 2: False negative (FN) rates, false positive (FP) rates, and the correlation coefficient (CC) for site detection on a separate test set of 114 sequences.

Signal Detection (93 Human Sequences)								
	True Sites Missed		CP Matrix			Conventional Matrix		
			False Sites Labeled True			False Sites Labeled True		
	Num	%	Num	%	CC	Num	%	CC
Start	0	0	2256	0.37	0.20	8499	1.41	0.10
Site	2	2.1	1518	0.25	0.23	5714	0.94	0.12
Detection	5	5.4	843	0.14	0.30	3922	0.65	0.14
(93 true	10	11	468	0.08	0.36	2221	0.37	0.18
human	20	22	293	0.05	0.39	1256	0.21	0.21
sites)	46	50	37	0.01	0.53	183	0.03	0.31
Donor	0	0	6252	1.03	0.23	10621	1.76	0.18
Site	8	2.3	3440	0.57	0.29	4814	0.80	0.25
Detection	17	4.9	2143	0.35	0.36	3656	0.60	0.28
(346 true	35	10	1300	0.22	0.42	2319	0.38	0.33
sites)	69	20	644	0.11	0.49	1361	0.23	0.37
	138	40	220	0.04	0.54	372	0.06	0.46
Acceptor	0	0	11057	1.83	0.17	17162	2.84	0.14
Site	6	1.7	6519	1.08	0.22	9936	1.64	0.18
Detection	17	4.9	3738	0.62	0.28	6112	1.01	0.22
(346 true	35	10	2626	0.43	0.31	3512	0.58	0.27
sites)	69	20	1824	0.30	0.32	2104	0.35	0.31
	137	40	763	0.13	0.36	1006	0.17	0.32

Table 3: Sensitivity and false positive rates of start, donor, and acceptor site detection for a set of 93 human DNA sequences using CP and conventional matrices computed from those sequences.

-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	
.23	.24	.42	.27	.16	.30	.16	.20	.16	.44	.28	.29	1.0	0.0	0.0	0.0	.38	.11	.37	$P(a_i\|a_{i-1})$
.23	.27	.24	.32	.57	.29	.08	.22	.67	.06	.45	.17	0.0	0.0	0.0	0.0	.14	.19	.27	$P(c_i\|a_{i-1})$
.23	.45	.24	.28	.23	.30	.68	.45	.14	.47	.15	.50	0.0	0.0	0.0	0.0	.40	.59	.27	$P(g_i\|a_{i-1})$
.23	.05	.10	.14	.04	.11	.08	.13	.03	.03	.13	.05	0.0	1.0	0.0	0.0	.08	.11	.08	$P(t_i\|a_{i-1})$
.40	.35	.30	.25	.15	.33	.29	.08	.32	.78	.48	.08	1.0	0.0	0.0	0.0	.32	.18	.38	$P(a_i\|c_{i-1})$
.40	.26	.33	.26	.47	.29	.28	.47	.46	.04	.41	.80	0.0	0.0	0.0	0.0	.29	.29	.28	$P(c_i\|c_{i-1})$
.40	.09	.11	.20	.10	.07	.21	.05	.13	.17	.10	.05	0.0	0.0	0.0	0.0	.01	.17	.13	$P(g_i\|c_{i-1})$
.40	.30	.26	.29	.28	.31	.21	.40	.10	.01	0.0	.07	0.0	0.0	0.0	0.0	.38	.37	.22	$P(t_i\|c_{i-1})$
.17	.17	.45	.22	.24	.29	.29	.41	.21	.59	.19	.19	1.0	0.0	0.0	.28	.17	.09	.22	$P(a_i\|g_{i-1})$
.17	.35	.19	.37	.40	.36	.33	.30	.55	.03	.67	.35	0.0	0.0	0.0	.15	.35	.28	.47	$P(c_i\|g_{i-1})$
.17	.33	.15	.30	.21	.17	.29	.16	.21	.34	.06	.44	0.0	0.0	0.0	.48	.14	.39	.23	$P(g_i\|g_{i-1})$
.17	.15	.21	.11	.16	.17	.09	.14	.03	.03	.07	.01	0.0	0.0	0.0	.09	.34	.21	.07	$P(t_i\|g_{i-1})$
.19	.10	.11	.11	.07	.20	.05	.06	.14	.47	.30	.11	1.0	0.0	0.0	0.0	.04	.03	.13	$P(a_i\|t_{i-1})$
.19	.47	.37	.51	.48	.32	.20	.40	.59	.12	.20	.82	0.0	0.0	0.0	0.0	.44	.17	.46	$P(c_i\|t_{i-1})$
.19	.26	.24	.22	.23	.24	.60	.27	.20	.38	.10	.03	0.0	0.0	1.0	0.0	.30	.69	.25	$P(g_i\|t_{i-1})$
.19	.17	.28	.16	.23	.25	.15	.27	.06	.03	.40	.03	0.0	0.0	0.0	0.0	.22	.12	.16	$P(t_i\|t_{i-1})$

Table 4: Conditional probability matrix for vertebrate start sites. Each column after the first contains the probability of a base in that position given the base in the previous position, as indicated at the end of each row.

-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	
0.23	.25	.33	.22	.16	.29	.20	.25	.22	.66	.27	.15	1.0	0.0	0.0	.28	.24	.11	.26	$P(a)$
0.40	.32	.28	.35	.48	.31	.21	.33	.56	.05	.50	.58	0.0	0.0	0.0	.16	.29	.24	.40	$P(c)$
0.17	.25	.17	.25	.18	.16	.46	.21	.17	.27	.12	.22	0.0	0.0	1.0	.48	.20	.45	.21	$P(g)$
0.19	.19	.21	.18	.19	.24	.14	.21	.06	.02	.11	.05	0.0	1.0	0.0	.09	.26	.21	.12	$P(t)$

Table 5: Standard probability matrix for vertebrate start sites.

-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10	+11	+12	+13	+14	+15	+16	+17	
.35	.60	.07	0.0	0.0	0.0	.64	.06	.20	.24	.19	.26	.16	.29	.23	.25	.28	.24	.26	.25	.24	$P(a_i\|a_{i-1})$
.35	.09	.02	0.0	0.0	0.0	.10	.03	.11	.22	.28	.24	.19	.18	.20	.20	.25	.24	.21	.24	.18	$P(c_i\|a_{i-1})$
.35	.18	.86	1.0	0.0	0.0	.13	.89	.39	.37	.23	.30	.38	.33	.37	.30	.31	.31	.30	.27	.34	$P(g_i\|a_{i-1})$
.35	.14	.06	0.0	0.0	0.0	.13	.03	.30	.17	.29	.20	.27	.19	.20	.25	.16	.21	.23	.24	.23	$P(t_i\|a_{i-1})$
.35	.69	.17	0.0	0.0	0.0	.70	.19	.25	.35	.20	.27	.27	.30	.22	.26	.24	.22	.21	.29	.27	$P(a_i\|c_{i-1})$
.35	.11	.06	0.0	0.0	0.0	.05	.21	.27	.26	.38	.31	.33	.33	.34	.35	.33	.36	.37	.30	.31	$P(c_i\|c_{i-1})$
.35	.07	.61	1.0	0.0	0.0	.07	.41	.09	.13	.06	.11	.11	.11	.10	.11	.10	.09	.10	.11	.09	$P(g_i\|c_{i-1})$
.35	.13	.16	0.0	0.0	0.0	.18	.20	.39	.26	.37	.31	.29	.27	.33	.28	.33	.33	.32	.31	.34	$P(t_i\|c_{i-1})$
.19	.65	.11	0.0	0.0	0.0	.83	.05	.15	.28	.21	.18	.21	.20	.24	.19	.24	.17	.20	.19	.20	$P(a_i\|g_{i-1})$
.19	.15	.01	0.0	0.0	0.0	.06	.05	.15	.29	.26	.30	.24	.23	.26	.25	.21	.30	.25	.22	.20	$P(c_i\|g_{i-1})$
.19	.11	.80	1.0	0.0	0.0	.09	.87	.15	.28	.34	.37	.39	.43	.32	.42	.36	.35	.39	.43	.39	$P(g_i\|g_{i-1})$
.19	.09	.08	0.0	1.0	0.0	.03	.03	.55	.15	.20	.14	.15	.14	.18	.15	.19	.17	.15	.16	.21	$P(t_i\|g_{i-1})$
.11	.16	.02	0.0	0.0	.51	.19	.05	.11	.24	.15	.15	.15	.18	.16	.10	.13	.15	.16	.15	.15	$P(a_i\|t_{i-1})$
.11	.24	.03	0.0	0.0	.03	.08	.11	.12	.19	.30	.28	.21	.18	.25	.24	.25	.22	.26	.21	.23	$P(c_i\|t_{i-1})$
.11	.31	.86	1.0	0.0	.43	.63	.77	.43	.36	.28	.31	.37	.40	.25	.32	.27	.30	.31	.32	.34	$P(g_i\|t_{i-1})$
.11	.29	.08	0.0	0.0	.03	.10	.06	.33	.20	.27	.25	.26	.24	.34	.35	.35	.33	.26	.32	.29	$P(t_i\|t_{i-1})$

Table 6: Conditional probability matrix for vertebrate donor sites.

-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10	+11	+12	+13	+14	+15	+16	+17	
0.35	.59	.08	0.0	0.0	.51	.71	.06	.15	.27	.19	.21	.20	.24	.22	.20	.22	.19	.20	.22	.21	$P(a)$
0.35	.13	.02	0.0	0.0	.03	.08	.05	.16	.23	.30	.29	.25	.23	.26	.26	.26	.28	.28	.24	.23	$P(c)$
0.19	.14	.82	1.0	0.0	.43	.12	.84	.17	.31	.23	.26	.30	.32	.27	.29	.26	.26	.27	.28	.29	$P(g)$
0.11	.14	.08	0.0	1.0	.03	.09	.05	.52	.20	.27	.24	.24	.21	.25	.25	.26	.26	.24	.26	.26	$P(t)$

Table 7: Standard probability matrix for vertebrate donor sites.

-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	
.09	.18	.22	.10	.13	.14	.14	.11	.09	.18	.58	.06	1.0	0.0	0.0	$P(a_i\|a_{i-1})$
.09	.27	.33	.29	.36	.40	.35	.41	.40	.34	.28	.76	0.0	0.0	0.0	$P(c_i\|a_{i-1})$
.09	.03	.02	.04	.02	.02	.01	.02	0.0	0.0	.04	.01	0.0	1.0	0.0	$P(g_i\|a_{i-1})$
.09	.52	.43	.56	.50	.44	.50	.47	.51	.48	.09	.18	0.0	0.0	0.0	$P(t_i\|a_{i-1})$
.34	.09	.09	.09	.10	.12	.13	.09	.08	.10	.36	.04	1.0	0.0	0.0	$P(a_i\|c_{i-1})$
.34	.32	.32	.32	.37	.41	.38	.42	.51	.48	.31	.68	0.0	0.0	0.0	$P(c_i\|c_{i-1})$
.34	.06	.03	.03	.04	.02	.07	.03	.02	.02	.09	0.0	0.0	0.0	0.0	$P(g_i\|c_{i-1})$
.34	.52	.57	.56	.50	.45	.42	.47	.40	.41	.23	.27	0.0	0.0	0.0	$P(t_i\|c_{i-1})$
.13	.08	.09	.03	.10	.06	.06	.04	.06	.07	.24	.02	1.0	0.0	.25	$P(a_i\|g_{i-1})$
.13	.27	.33	.24	.32	.33	.38	.42	.30	.27	.21	.85	0.0	0.0	.16	$P(c_i\|g_{i-1})$
.13	.11	.12	.15	.16	.10	.12	.14	.08	.19	.45	0.0	0.0	0.0	.50	$P(g_i\|g_{i-1})$
.13	.54	.46	.58	.42	.50	.43	.40	.55	.47	.10	.13	0.0	0.0	.09	$P(t_i\|g_{i-1})$
.44	.06	.06	.03	.05	.07	.06	.06	.05	.06	.16	.04	1.0	0.0	0.0	$P(a_i\|t_{i-1})$
.44	.32	.30	.32	.40	.32	.41	.40	.41	.28	.26	.68	0.0	0.0	0.0	$P(c_i\|t_{i-1})$
.44	.18	.18	.15	.14	.20	.17	.12	.09	.08	.31	0.0	0.0	0.0	0.0	$P(g_i\|t_{i-1})$
.44	.44	.46	.50	.42	.41	.36	.42	.45	.58	.26	.28	0.0	0.0	0.0	$P(t_i\|t_{i-1})$

Table 8: Conditional probability matrix for vertebrate acceptor sites.

-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	
0.09	.08	.09	.06	.07	.09	.09	.07	.06	.08	.27	.04	1.0	0.0	.25	$P(a)$
0.34	.31	.31	.31	.38	.36	.39	.41	.44	.37	.28	.74	0.0	0.0	.16	$P(c)$
0.13	.12	.11	.10	.10	.11	.11	.08	.05	.05	.22	0.0	0.0	1.0	.50	$P(g)$
0.44	.49	.49	.53	.45	.44	.40	.44	.44	.49	.23	.22	0.0	0.0	.09	$P(t)$

Table 9: Standard probability matrix for vertebrate acceptor sites.

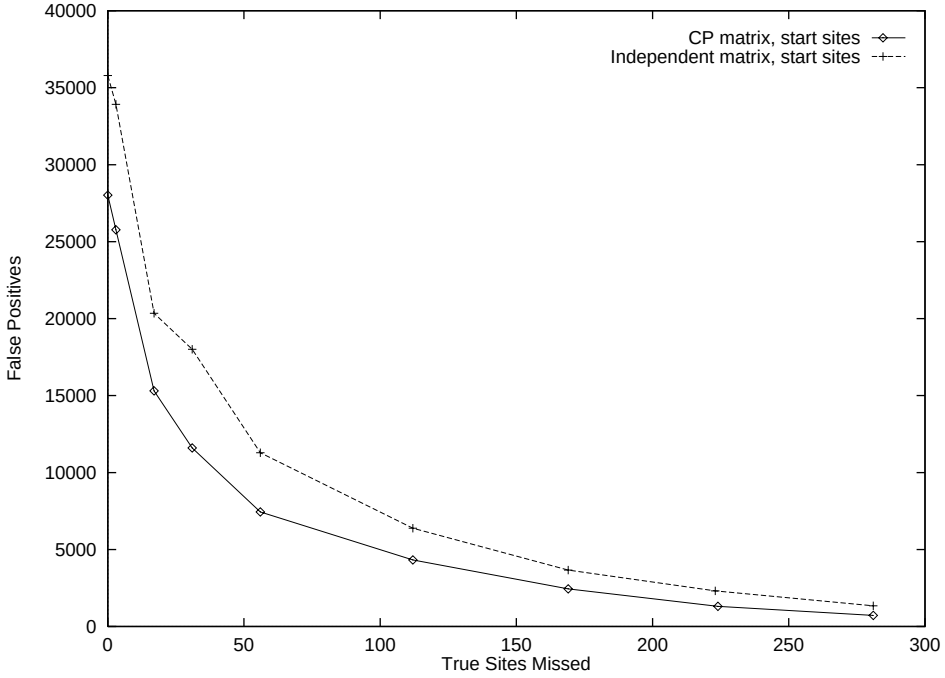

Figure 1: Comparison of the conditional probability (CP) matrix and independent probability matrix for detection of start sites. The vertical axis shows the number of sites incorrectly labeled as start sites (false positives) out of 2.8 million sites. The horizontal axis shows how many true sites were missed out of 562 total. The CP method has fewer false positives for every threshold setting.

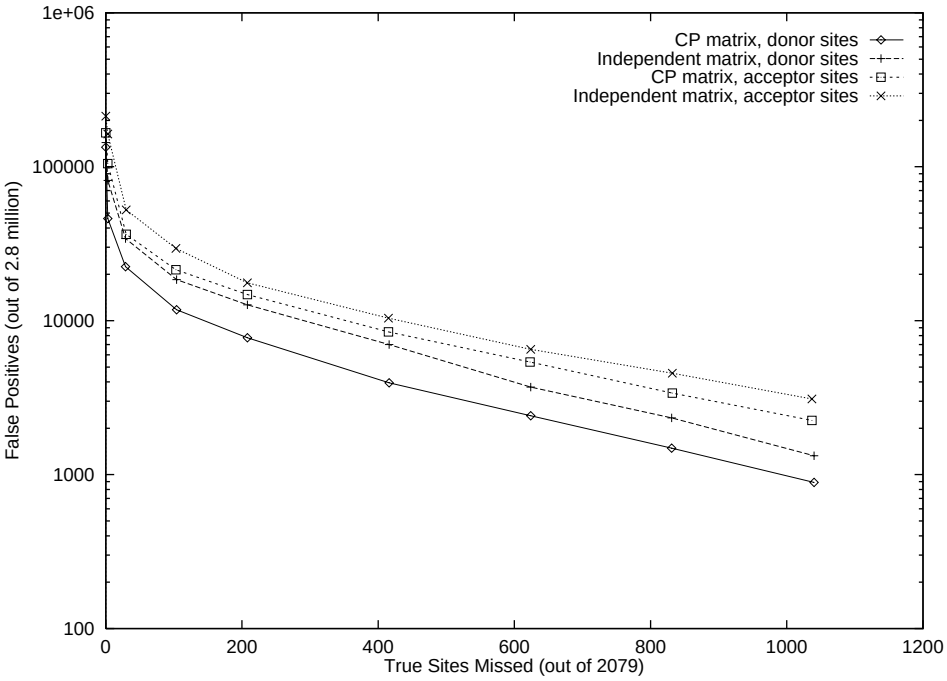

Figure 2: Comparison of the conditional probability (CP) matrix and independent probability matrix for detection of donor and acceptor sites. The horizontal axis, which is shown on a log scale for clarity, shows how many true sites were missed. The CP method has fewer false positives for every threshold setting, for both the donor and acceptor site matrices.

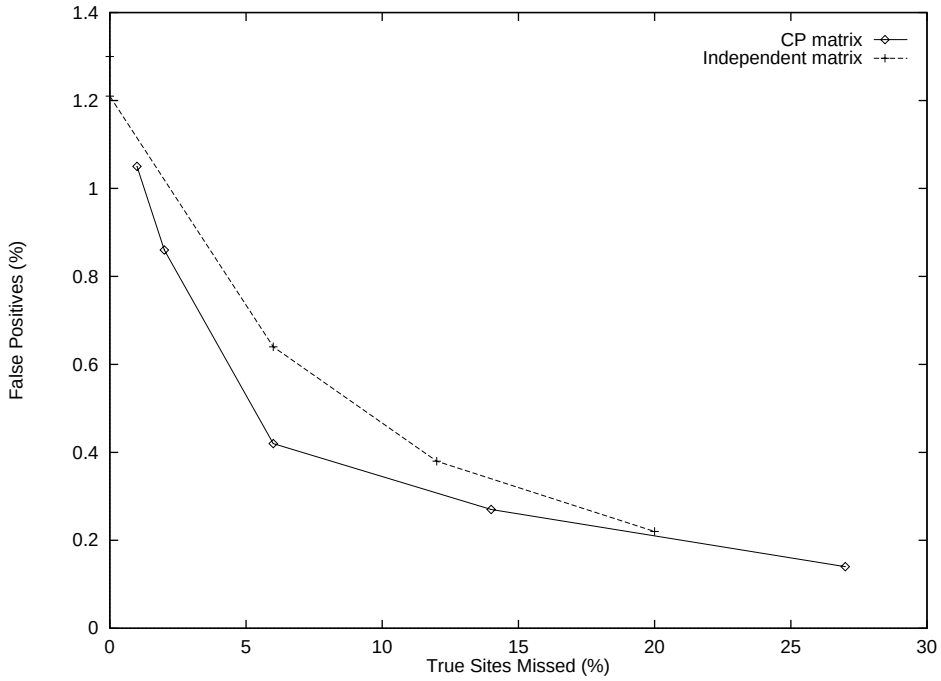

Figure 3: Comparison of the conditional probability (CP) matrix and independent probability matrix for detection of start sites on a separate test set of 114 sequences.

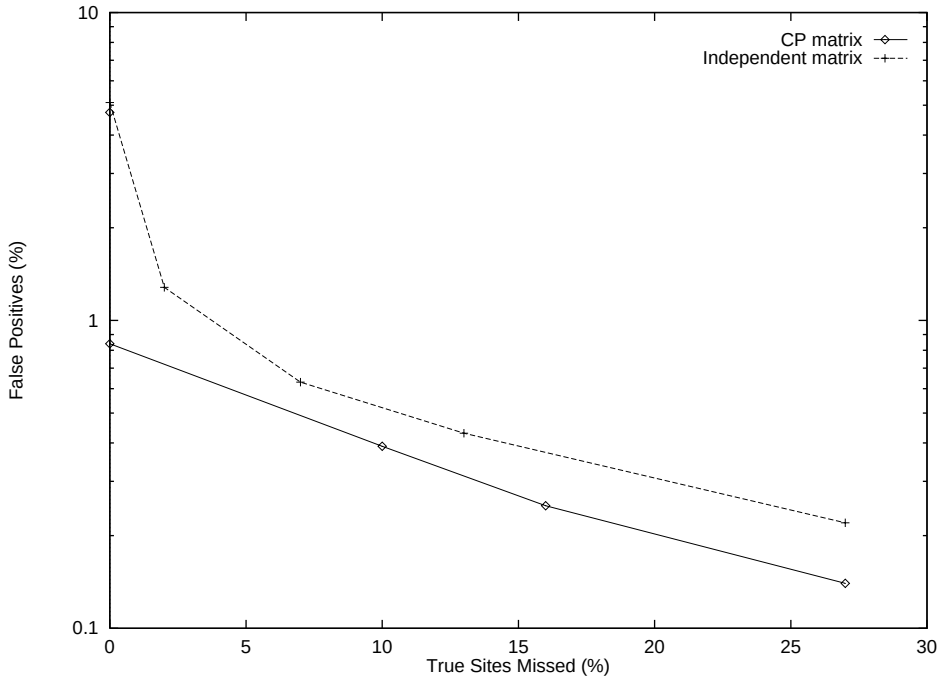

Figure 4: Comparison of the conditional probability (CP) matrix and independent probability matrix for detection of donor sites on a separate test set of 114 sequences with 385 donor sites.

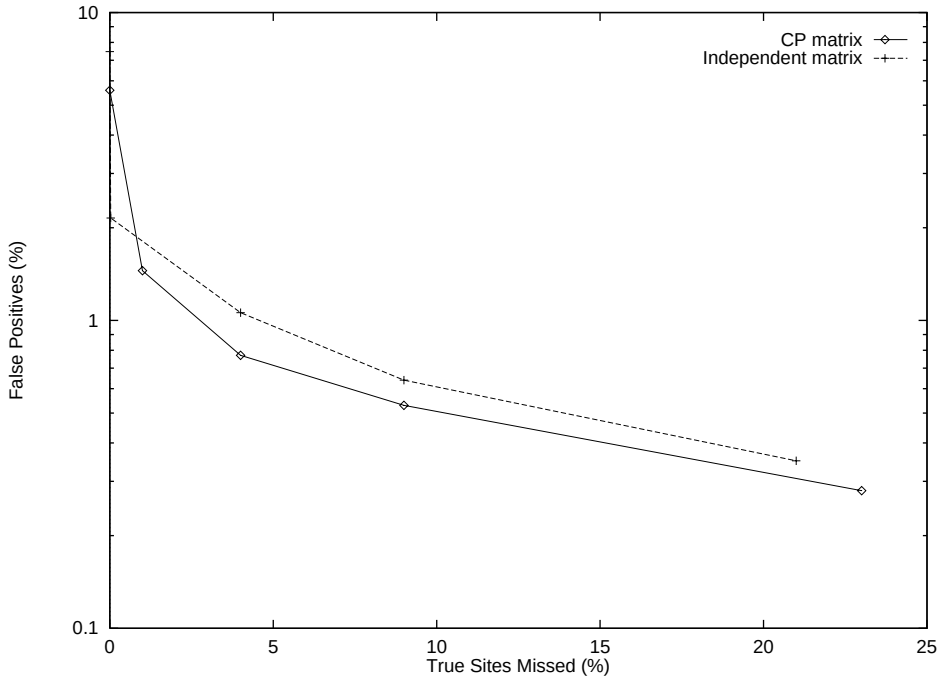

Figure 5: Comparison of the conditional probability (CP) matrix and independent probability matrix for detection of acceptor sites on a separate test set of 114 sequences with 385 acceptor sites.