Identification of trans-splicing sites in Leishmania major using probabilistic methods

Lukas Habegger
Identification of trans-splicing sites in *Leishmania major* using probabilistic methods

Approved:  

______________________________  
Gary R. Skuse, Ph.D.  
Director of Bioinformatics

______________________________  
Richard L. Doolittle, Ph.D.  
Head, Department of Biological Sciences

Submitted in partial fulfillment of the requirements for the Master of Science degree in Bioinformatics at the Rochester Institute of Technology.

Lukas Habegger  
May 2007
Thesis/Dissertation Author Permission Statement

Title of thesis or dissertation: Identification of trans-splicing sites in *Leishmania major* using probabilistic methods

Name of author: Lukas Habegger
Degree: Master of Science
Program: Bioinformatics
College: College of Science

I understand that I must submit a print copy of my thesis or dissertation to the RIT Archives, per current RIT guidelines for the completion of my degree. I hereby grant to the Rochester Institute of Technology and its agents the non-exclusive license to archive and make accessible my thesis or dissertation in whole or in part in all forms of media in perpetuity. I retain all other ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

**Print Reproduction Permission Granted:**
I, Lukas Habegger, hereby grant permission to the Rochester Institute Technology to reproduce my print thesis or dissertation in whole or in part. Any reproduction will not be for commercial use or profit.

Signature of Author: Lukas Habegger Date: ______

**Inclusion in the RIT Digital Media Library Electronic Thesis & Dissertation (ETD) Archive**

I, Lukas Habegger, additionally grant to the Rochester Institute of Technology Digital Media Library (RIT DML) the non-exclusive license to archive and provide electronic access to my thesis or dissertation in whole or in part in all forms of media in perpetuity. I understand that my work, in addition to its bibliographic record and abstract, will be available to the world-wide community of scholars and researchers through the RIT DML. I retain all other ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation. I am aware that the Rochester Institute of Technology does not require registration of copyright for ETDs.

I hereby certify that, if appropriate, I have obtained and attached written permission statements from the owners of each third party copyrighted matter to be included in my thesis or dissertation. I certify that the version I submitted is the same as that approved by my committee.

Signature of Author: Lukas Habegger Date: ______
THESIS ADVISORY COMMITTEE

Dr. Shuba Gopal
Assistant Professor
Department of Biological Sciences
College of Science
Rochester Institute of Technology

Dr. James Halavin
Professor
School of Mathematical Sciences
College of Science
Rochester Institute of Technology

Mr. Paul Tymann
Associate Professor and Department Chair
Department of Computer Science
College of Computing and Information Sciences
Rochester Institute of Technology

Dr. David Lawlor
Associate Professor
Department of Biological Sciences
College of Science
Rochester Institute of Technology
ABSTRACT

*Leishmania major*, a member of the Kinetoplastida family, is a primitive protozoan that causes a human disease, called leishmaniases, affecting numerous people worldwide. The identification of new drug targets to combat leishmaniases necessitates a thorough understanding of how genomic instructions are transformed into functional proteins. It requires not only the prediction and categorization of all the genes, but also a profound understanding of their regulation. Much of gene regulation may occur through a process known as *trans*-splicing. *Trans*-splicing, which is mechanistically similar to *cis*-splicing, is the process of cleaving a large polycistronic transcript into smaller monocistronic components.

The goal of this project was to establish a model to accurately predict sites where *trans*-splicing occurs. After carefully analyzing the data set, a second-order log odds ratio model was created. This method achieved an overall accuracy of 89% in predicting *trans*-splice sites.

Furthermore, this new method has been applied to a small data set with alternative *trans*-splice sites. Of the 70 EST-indicated alternative *trans*-splice sites 60 were identified as such. This represents the first computational method for the prediction of alternative splice sites. In addition, we have found the first real evidence for the branch-point signal which plays an essential role in the *trans*-splicing process.
# TABLE OF CONTENTS

**THESIS ADVISORY COMMITTEE**.................................................................ii

**ABSTRACT**........................................................................................................iii

1. **INTRODUCTION**.................................................................................................1 – 4

2. **MATERIALS & METHODS**..............................................................................5 – 12
   2.1 Data Set...........................................................................................................5
   2.2 Methods..........................................................................................................5 – 12
      2.2.1 Chi Square Test for Homogeneity..............................................................5 – 6
      2.2.2 Entropy Calculations..................................................................................6
      2.2.3 Calculating Transition Probabilities..........................................................7
      2.2.4 Chi Square Test to Determine the Order of a Markov Chain..................7 – 8
      2.2.5 Maximal Dependence Decomposition......................................................8 – 10
      2.2.6 Log Odds Ratio Calculations.....................................................................10 – 12
      2.2.7 Alternative Splicing Analysis....................................................................12

3. **RESULTS**.........................................................................................................13 – 24
   3.1 Nucleotide Frequencies..................................................................................13 – 14
   3.2 Chi-Square Test for Homogeneity....................................................................14
   3.3 Entropy............................................................................................................14 – 15
   3.4 Markov Chains...............................................................................................16 – 18
   3.5 Maximal Dependence Decomposition............................................................19 – 20
   3.6 Log Odds Results.............................................................................................20 – 21
   3.7 Analysis of the Inter-AG Distance...................................................................21 – 23
   3.8 Alternative Splicing.........................................................................................24

4. **DISCUSSION**.....................................................................................................25 – 27
   4.1 Model Selection...............................................................................................25
   4.2 Maximal Dependence Decomposition and the Importance of Position 26 ........26
   4.3 Alternative Splice Site Analysis.......................................................................27

5. **CONCLUSION**..................................................................................................28

6. **REFERENCES**..................................................................................................29 – 30

7. **APPENDICES**..................................................................................................31 – 39
1. INTRODUCTION

*Leishmania major* is an evolutionarily ancient protozoan that belongs to the Kinetoplastida family. It diverged from *Saccharomyces cerevisiae* and other eukaryotes about 800 million years ago (Fernandes, 1993). This organism is the causative agent of leishmaniasis, affecting approximately 2 million people in more than 88 countries. Recently, the genome of *Leishmania major* has been elucidated by shotgun sequencing. It consists of 36 chromosomes and 32,816,678 base pairs. Several computational methods were employed in order to investigate the genomic sequence. The analysis revealed 911 RNA genes, 39 pseudogenes, and 8272 protein-coding genes (Ivens, 2005).

A central interest in studying genomic sequences has been the identification of RNA splice sites in order to understand how genomic instructions are transformed into functional proteins. In addition to *cis*-splicing, which is the process of removing introns from the nascent mRNA, some eukaryotes also perform *trans*-splicing. This phenomenon has been widely observed in the Kinetoplastida family, but recent studies have indicated its presence in other eukaryotes as well (Blumenthal, 2002). *Trans*-splicing is the process of cleaving a large polycistronic transcript into smaller monocistronic components by adding a short spliced leader sequence to the 5’ untranslated region (UTR) of each mRNA and a polyadenylation tail at the 3’ end. A schematic representation of a polycistronic transcript and the *trans*-splicing process is shown in Figure 1.1.
Figure 1.1: Schematic representation of the trans-splicing process. Transcription yields a polycistronic transcript precursor which is subjected to trans-splicing. During trans-splicing a short splice leader sequence is added to each coding region at the 5’-end resulting in monocistronic mRNAs. Afterwards, a polyadenylation tail is appended at the 3’-end of each processed mRNA.

In the past years much effort has been focused on the identification of protein-coding genes and the detection of cis-splice sites in eukaryotic organisms. Although the development of sophisticated algorithms to discover trans-splice sites has been slow, a relatively simple but fairly accurate method has been proposed to address this issue. This method employs linear discriminant analysis (LDA), which is a statistical approach, to classify potential trans-splicing regions based on dinucleotide composition (Gopal, 2005). Studies have shown that trans-splicing regions show a bias in favor of pyrimidines (Lee, 1997 and Vanhamme, 1995). However, the presence of this bias is not sufficient to identify splice junctions. Careful analysis revealed that the distance between AG dinucleotides is a good indicator for the presence of a splice junction. Using the combination of LDA and the distance between AG dinucleotides this algorithm achieved
an overall accuracy of 82% and was capable of predicting 92% of all known trans-splicing sites (Gopal, 2005). Nonetheless, this investigation was based on a small data set consisting of 214 expressed-sequence-tag (EST)-mapped trans-splicing regions and 198 known, experimentally verified coding regions.

The recent completion of the genomic sequence of *Leishmania major* has provided the data necessary to build a more sophisticated model. Models that are based on probabilistic methods have proven to be successful in the identification of cis-splicing sites (Burge, 1997). Furthermore, evidence suggests that the two splicing processes use the same underlying mechanism since they share many of the spliceosomal components (Liang, 2003 and Palfi, 2000). Thus, the development and the application of such a method seems to be plausible.

By using a method that is based on a probabilistic approach we may improve the prediction of trans-splicing sites. In addition, this method might be used to draw inferences about the spliceosome. Having a more accurate algorithm to identify splice junctions will help to shed light into many aspects of the unknown biology of trypanosomes. Until now gene prediction in trypanosomes was based upon prokaryotic methods since these organisms exhibit features of genomic organization that resemble prokaryotes. However, the prediction of genes in trypanosomes using these methods proved to be unreliable (Gopal, 2003).

In order to identify new drug targets that cure leishmaniasis it is essential to better understand the biology of *Leishmania major*. This involves finding all the genes and understanding their regulation. Previous studies have shown that the genes in this organism are regulated via trans-splicing. Thus, in order to understand gene regulation it
is necessary to elucidate the signals related to trans-splicing. These signals are located in the upstream region of coding genes while they are absent in pseudogenes and non-coding open reading frames. Therefore, the identification of trans-splicing sites may reveal the loci of previously unknown genes. Another benefit of the proposed method is that it does not depend on previously identified genes.

Furthermore, the proposed method could be used to detect alternative splice sites. Alternative splicing has been observed in many organisms (Brett, 2002). There has been some speculation about the presence of multiple trans-splice sites for certain transcripts in Leishmania major (Gopal, 2005 and Vassella, 1994). The development of this method may help to identify alternative trans-splice sites with some reliability. Since there is presently no such approach for trans-splicing, this method might help to address this issue.

The development of this novel probabilistic method can be subdivided into three major parts. First, the basic properties of the genomic sequences need to be analyzed in order to obtain a better understanding of the nature of the signals involved in trans-splicing. The second step involves building a probabilistic model to identify trans-splicing regions. Finally, given a potential trans-splicing region, the model has to predict the actual splice junction.
2. MATERIALS & METHODS

2.1 Data Set

The data set used for this project contains 611 unique regions of Leishmania major, each consisting of 400 nucleotides upstream of the splice junction. These sequences were obtained by EST mappings to the Leishmania major genome (Griggs, 2007). All the chromosomes are represented in this data set. A table that summarizes the counts across the chromosomes is shown in Appendix A.

2.2 Methods

In the preliminary stage of this project the sequences in the data set were analyzed by using exploratory statistical methods in order to examine the nature of the trans-splicing signals. First, the nucleotide relative frequencies in each position of the 611 sequences were plotted. A similar graph of the purine and pyrimidine relative frequencies was generated as well. Secondly, a Chi-Square test of the nucleotide counts in 36 chromosomes was conducted in order to examine whether sequence conservation exists in the upstream region across the chromosomes. The last step of the preliminary analysis was concerned with single and dinucleotide distributions along the 400 positions using entropy as a metric.

2.2.1 Chi Square Test for Homogeneity

A Chi Square test for homogeneity was conducted to determine if the distribution of the four nucleotides is analogous among the 36 chromosomes. The test statistic was
computed as shown in Equation 2.1 (Devore, 2004). It should be noted that \( I \) represents the number of chromosomes while \( J \) indicates the number of nucleotides. \( N \) stands for the total number of nucleotides.

\[
\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \left( \frac{n_{ij} - \hat{e}_{ij}}{\hat{e}_{ij}} \right)^2
\]

\[
\hat{e}_{ij} = \frac{n_i \times n_j}{N}
\]

\( I = 36 \quad (2.1) \)

\( J = 4 \)

\( DF = (36-1) \times (4-1) = 105 \)

**Equation 2.1:** Calculation of the Chi Square statistic in the test for homogeneity

### 2.2.2 Entropy Calculations

The entropy of a certain probability distribution is a measure that relates to the amount of uncertainty. It is defined in Equation 2.2. Clearly, if the probability of an outcome is one then the entropy will be zero since there is no uncertainty associated with the event. On the other hand, if all the outcomes are equally likely then the entropy is maximized at \( \log_n (n) \) (Kullback, 1951).

Let \( X \) be a discrete random variable with \( n \) possible outcomes, then the entropy \( H(X) \) is defined as:

\[
H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 \left( p(x_i) \right)
\]

where

\( p(x_i) \) is the probability of the \( i^{th} \) outcome of \( X \)

**Equation 2.2:** Definition of entropy. The entropy of a certain probability distribution measures to the amount of uncertainty.
2.2.3 Calculating Transition Probabilities

A program was written in C to estimate the transition probabilities for a given position within the set of sequences. It should be noted that transition probabilities are conditional probabilities which are defined using the following formula:

Formulas to calculate transition probabilities:
First-Order:
\[ P(X_{n+1} = i | X_n = j) = \frac{P(X_{n+1} = i, X_n = j)}{P(X_n = j)} \]

Second-Order:
\[ P(X_{n+2} = i | X_{n+1} = j, X_n = k) = \frac{P(X_{n+2} = i, X_{n+1} = j, X_n = k)}{P(X_{n+1} = j, X_n = k)} \] (2.3)

where:
\[ X_n \] denotes a nucleotide at position \( n \)
where \( i, j, \) and \( k \in \{ A, C, G, T \} \)

Equation 2.3: Formulas to calculate the first-order and second-order transition probabilities.

It is evident from Equation 2.3 that the first-order transition probabilities are conditioned on the previous nucleotide while second-order transition probabilities depend on the previous two nucleotides.

2.2.4 Chi Square Test to Determine the Order of a Markov Chain

Chi Square tests can be used to investigate the dependence structure of Markov Chain. To test for first-order dependence, a chi square test statistic is calculated for all adjacent positions in the set of sequences by making a 4 by 4 table representing the counts of the different dinucleotide combinations. The rows of that table indicate the nucleotides at position \( i \) while the columns indicate the nucleotides at position \( i+1 \). From this table, the chi square statistic is calculated by summing the square difference between the observed and expected count, divided by the expected count. The expected
count for a particular entry in this table is obtained by taking the product of the row total and column total, which is then divided by the total count of the table. According to statistical theory, under the null hypothesis, this sum follows asymptotically a chi-square distribution with 9 degrees of freedom (Devore, 2004).

If the null hypothesis of independence is rejected, a chi square test for second-order dependence can conducted. Instead of constructing a 4 by 4 table for all adjacent positions, a 16 by 4 table is established. The sixteen rows indicate all the possible dinucleotides at positions \(i\) and \(i+1\), respectively. The chi square test statistic is calculated in the same fashion as described previously. In this case the test statistic follows a chi square distribution with 45 degrees of freedom. The associated critical value is 61.66 at the 5% level (Devore, 2004). A more general notation for this Chi Square test and all the required formulas are described in Appendix B.

### 2.2.5 Maximal Dependence Decomposition

Maximal dependence decomposition is a method used to analyze the dependency structure in a set of \(n\) sequences each of length \(k\). A program was written in C to perform this analysis. The procedure consists of two major steps. First, a \(k \times k\) matrix is generated. The entries of this matrix at row \(i\) and column \(j\) are obtained by conducting a chi square test for independence using the nucleotide counts at the fixed positions \(i\) and \(j\). Thus, for each entry in the matrix a 4 by 4 table has to be constructed which is then used to calculate the test statistic. After calculating all entries of the \(k \times k\) matrix the row sums can be determined. The position with the greatest row sum exerts the most
influence over the other $k-1$ positions. In the second step, the data set is split according to the nucleotide present in the most influential position. Thus, four new subsets are generated each with either an $A, T, G,$ or $C$ at that position. Subsequently, the first step of this procedure is repeated. Each subset can be further subdivided as long as there is sufficient data. Figures 2.1 and 2.2 summarize the major steps of this method. With this technique it is possible to obtain conditional weight matrices which can be used to dissect and analyze the dependence structure of $n$ sequences (Burge, 1998 and Ewens, 2001).

**Figure 2.1:** Overview of the Maximal Dependence Decomposition process. First, the original data set is used to build the Chi Square matrix (400 by 400). Next, the sum of each row in the matrix is determined and the position with the greatest row sum is identified. Subsequently, the data set is subdivided into four subsets according to the nucleotide found in the previously identified (most influential) position. This process is repeated for each subset as long as there is sufficient data.
Figure 2.2: Example that shows how an entry in the 400 by 400 Chi Square matrix is calculated. Each entry in this matrix requires a 4 by 4 table that is used to calculate the Chi Square statistic. The rows of the 4 by 4 table represent the nucleotides observed in position 3 while the columns correspond to the nucleotides found in position 4. The Chi Square value for this example is 50.19.

2.2.6 Log Odds Ratio Calculations

Before conducting the analysis the original data set, consisting of 611 sequences, was randomly split into a training as well as a testing data set in order to perform cross-validation. The training data set includes 549 sequences or 90 percent of all sequences while the testing data set contains the remaining 62 sequences.

The first 100 nucleotides upstream of the splice junction of each sequence were used as the “plus” model. The region located 300 to 400 base pairs upstream of the splice junction served as the “minus” model as shown in Figure 2.3. It should be noted that both
of these regions can be analyzed in the 5’ to 3’ as well as in 3’ to 5’ direction yielding different results.

Figure 2.3: Schematic representation that shows where the regions for the “plus” and “minus” model are located.

For each region both first-order and second-order transition probabilities were obtained. Next, for each transition probability a log odds ratio was calculated as shown in Equation 2.4 (Durbin, 1998).

\[
\log_2 \left( \frac{P(X_i^+)}{P(X_i^-)} \right)
\]

where
+ indicates the plus model
- indicates the minus model

Case 1:
\( i = 1, 2, 3, ..., 16 \) representing all the first-order transition probabilities

Case 2:
\( i = 1, 2, 3, ..., 64 \) representing all the second-order transition probabilities

Equation 2.4: Formula to calculate the first-order and second-order log odds ratios.

For any given sequence, these log odds ratios can be used to estimate whether the sequence originated from the “plus” or the “minus” model. This is done by scoring the sequence. The score is obtained by scanning along the sequence while adding up the log odds ratios. If the final score is greater than zero, it is likely that the sequence originated from the plus model, indicating the presence of splice region.
After obtaining these log odds ratios from the training data set, the testing data set was used to determine the adequacy of the model. The “plus” and “minus” region of each sequence in the testing data set was scored. Subsequently, these scores were assessed for their correctness. For example, the “plus” regions were expected to yield positive scores. If these scores were indeed positive, they were counted as a “true positives.” Otherwise, they were counted as a “false negatives”. A similar analysis was conducted for the “minus” regions of the testing data set. After categorizing the scores it was possible to determine the specificity, sensitivity, and accuracy of the proposed method. The formulas for these calculations are shown in Equation 2.5.

\[
\begin{align*}
TP &= \text{True Positive} \\
FP &= \text{False Positive} \\
TN &= \text{True Negative} \\
FN &= \text{False Negative} \\
\text{Specificity} &= \frac{TN}{TN + FP} \\
\text{Sensitivity} &= \frac{TP}{TP + FN} \\
\text{Accuracy} &= \frac{TP + TN}{TP + FP + TN + FN}
\end{align*}
\]

**Equation 2.5**: Formulas to determine the specificity, sensitivity, and accuracy of the proposed method.

### 2.2.7 Alternative Splicing Analysis

A data set consisting of 34 sequences with EST-indicated alternative splice sites from *Leishmania major* was used to perform this analysis (Griggs, 2007). A program was written in C to analyze each AG in those sequences. For every AG the first 100 nucleotides upstream were scored using the second-order log odds model. The objective was to see if this method can identify alternative splice sites.
3. RESULTS

3.1 Nucleotide Frequencies

Figure 3.1 shows the purine and pyrimidine relative frequencies of the 400 nucleotides upstream of the splice junction using all 611 sequences in the data set. The frequencies of the individual nucleotides are shown in Appendix C. The actual splice site is located on the left hand side of the graphs. It is evident from these graphs that the relative frequencies change as we near the splice junction.

Figure 3.1: Purine and pyrimidine relative frequencies of the first 400 base pairs upstream before the splice junction of the 611 sequences.
The pyrimidine relative frequency begins to increase considerably starting approximately 230 base pairs upstream of the splice junction. It reaches a peak roughly 50 base pairs upstream of the splice junction before it starts to decrease again. It appears very likely that the signals for trans-splicing are contained within the first 200 base pairs upstream of the splice junction.

3.2 Chi-Square Test for Homogeneity

The calculated Chi-Square value was statistically significant with a p-value of 0.000. Hence, there is evidence to conclude that the nucleotide distribution among the 36 chromosomes is not the same. This finding supports previous evidence that there is minor sequence conservation in the upstream region of a splice junction (Gopal, 2005).

3.3 Entropy

Figure 3.2 illustrates the entropy plot of the 400 base pairs upstream of the splice junction using single nucleotide relative frequencies. The region, which is located between 200 and 400 base pairs upstream of the splice junction, exhibits an entropy that is almost maximal reaching a value of two \([\log_2(4) = 2]\). This indicates that the distribution of the nucleotide frequencies is approximately uniform. Thus, the level of uncertainty is maximized.
Entropy of the first 400 base pairs Upstream of the Splice Junction using Single Nucleotide Frequencies (N = 611)

![Graph showing entropy changes](image)

**Figure 3.2:** Entropy of the first 400 base pairs upstream of the splice junction using single nucleotide frequencies.

Starting approximately 230 base pairs away from the splice junction the entropy begins to decrease indicating that nucleotide frequencies are not uniform, which is in agreement with Figure 3.1. This result could be attributed to the presence of the pyrimidine rich region, known as pyrimidine trap, which is located in close proximity to the splice junction.

The entropy plots utilizing dinucleotides and trinucleotides are shown in Appendix D. It should be noted that both of these graphs reveal a very similar pattern, comparable to the one seen in Figure 3.2. This indicates that there is a preferred set of nucleotides used upstream of the splice junction which can be observed at the single, di, and trinucleotide level.
3.4 Markov Chains

After investigating the DNA sequences at the single nucleotide level the analysis was expanded using Markov Chains. A Markov Chain can be regarded as a special case of a stochastic process, which is a collection of random variables defined on an index set \( K \) denoted \( \{ X(t), t \in K \} \). In this stochastic process \( X \) can take on a certain set of values, which is known as the state space \( S \). Applying this concept to DNA sequences, the state space is defined as \( S = \{ A, T, G, C \} \) while \( K \) is the set of nonnegative integers relating to the position within the sequence (\( K = 1, 2, 3, ..., 400 \)). In essence, a DNA sequence can be regarded as a stochastic process of letters which are defined by \( S \) and index set \( K \).

The main feature that distinguishes a Markov Chain from other stochastic processes is the fact that, given the present state of the process, the previous and future states are independent of one another. The transition probability matrix governs how the process changes from the current state to the next. The entries of this matrix are conditional probabilities defined as \( P(X_{n+m} = j | X_n = i) = p_{ij}^m(n) \). This is called the m-step transition probability of the Markov Chain at time \( n \). From this definition it is evident that the transition probability depends on \( i, j, m, \) and \( n \). However, in a special case where the transition probabilities are independent of \( n \), the Markov Chain is said to be homogeneous. The interpretation of such a process at the biological sequence level would be that the same Markov Chain governs the composition of the DNA sequence independent of the location of that sequence.

When analyzing Markov Chains it is often important to determine the proper order of the Markov Chain. The order is a way of characterizing the dependence structure of a Markov Chain. For example, using a second-order Markov chain for biological sequences
would mean that the probability of observing the next nucleotide would be dependent on the previous two nucleotides. On the other hand, if the Markov Chain is said to be of zero-order that would indicate that the next nucleotide is independent of the current one.

The chi square test statistics for testing first-order dependence are shown in Figure 3.3. At a cutoff level of 5%, the critical chi square value with 9 degrees of freedom is 16.92. It is evident from this graph that most positions show first-order dependence. In fact, 87.22 percent of all the positions were statistically significant. Furthermore, it appears that the positions in close proximity to the splice junction have particularly high chi square values.

![Chi Square Test For First Order Dependence](image)

**Figure 3.3:** Chi square test for first order dependence.

Since the null hypothesis of independence was rejected, a chi square test for second-order dependence was conducted. The results for testing second-order
dependence are shown in Figure 3.4. It should be noted that the overall pattern in this plot is analogous to the one observed in Figure 3.3. Further analysis revealed that 92.71 percent of the tested positions were statistically significant at the 5% cutoff level. Next, a chi square test for third-order dependence should be conducted. However, the data set is too small to perform a higher order chi square test.

![Chi Square Test For Second Order Dependence](Image)

**Figure 3.4:** Chi square test for second order dependence.
3.5 Maximal Dependence Decomposition

Maximal dependence decomposition is an approach used to analyze the dependency structure in a set of \( n \) sequences each of length \( k \). The objective of this method is to find the positions that exert the most influence over the other \( k - 1 \) positions (Burge and Ewens). Figure 3.5 shows the row sums obtained from the 400 by 400 maximal dependence decomposition matrix. This graph clearly indicates that position 26 upstream of the splice junction (the splice junction AG would be positions 1 and 2, respectively) is most influential. Furthermore, it seems that the positions in close proximity to the splice junction are quite influential. Subsequently, the data set was subdivided into four subsets according to nucleotide present in position 26. The results are shown in Figure 3.6.

![Chi Square Statistics using Maximal Dependence Decomposition (Row Sums)](image)

**Figure 3.5**: Plot of the row sums of the chi square statistics using maximal dependence decomposition.
<table>
<thead>
<tr>
<th>Subset at Position 26</th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Sequences</td>
<td>82</td>
<td>199</td>
<td>104</td>
<td>226</td>
</tr>
<tr>
<td>Next Most Influential Position</td>
<td>23</td>
<td>25</td>
<td>278</td>
<td>149</td>
</tr>
</tbody>
</table>

**Figure 3.6:** Summary of the subsets obtained after splitting the original data set on position 26. It should be noted that if the nucleotide at position 26 is an adenine or a thymine then next most influential position is located nearby. On the other hand, if the nucleotide at position 26 is a guanine or a cytosine then the next influential position is farther upstream. The graphs of the Chi Square statistics for each subset are shown in Appendix E.

This table shows that if the nucleotide at position 26 is an adenine or a thymine then the next influential position is located near the most influential position. Conversely, if the nucleotide at position 26 is a guanine or a cytosine then the next influential position is farther upstream. This finding suggests that position 26 may be an important signal, perhaps the branch-point. The significance of this is discussed further in the Discussion section.

### 3.6 Log Odds Results

Probabilistic methods based on Hidden Markov Models have been widely used to identify *cis*-splicing sites in many eukaryotes (Burge, 1997). However, the nature of the *trans*-splicing signal is less well understood and the absence of a clear delineation between splice and non-splice regions makes this method not well suited (see Discussion). Instead a simpler method that is based on log odd scores was used. This method has previously been employed to identify CpG – islands in DNA sequences (Durbin, 1998).

The goal of this method was to detect potential splice regions. A ten-fold cross-validation was used to assess the adequacy of this method as shown in Figure 3.7. This
Table shows the results of the log odds analysis using four distinctive models. These results were obtained by taking the mean of the 10 independent runs. The outcomes of each individual run are shown in Appendix F. The models differ in the direction they were built and whether they used first-order or second-order Markov Chains. It is evident that the second-order model that was constructed in the 3' to 5' direction yields the highest average specificity, sensitivity, and accuracy. One explanation for this finding might be that the spliceosome, which operates in the 5' to 3' direction, evaluates a potential splice site by weighting the nucleotides ahead more than previously observed nucleotides.

<table>
<thead>
<tr>
<th>Direction</th>
<th>Order</th>
<th>Runs</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>3' to 5'</td>
<td>first</td>
<td>10</td>
<td>83.39</td>
<td>87.58</td>
<td>85.49</td>
</tr>
<tr>
<td>3' to 5'</td>
<td>second</td>
<td>10</td>
<td>88.39</td>
<td>90.49</td>
<td>89.44</td>
</tr>
<tr>
<td>5' to 3'</td>
<td>first</td>
<td>10</td>
<td>80.65</td>
<td>89.52</td>
<td>85.08</td>
</tr>
<tr>
<td>5' to 3'</td>
<td>second</td>
<td>10</td>
<td>80.97</td>
<td>88.87</td>
<td>84.92</td>
</tr>
</tbody>
</table>

*Figure 3.7: Results of the log odds analysis using four distinct models.*

### 3.7 Analysis of the Inter-AG Distance

Previously, it has been shown that the distance between the AG upstream of the splice junction and the actual splice site can be a useful metric in predicting the actual splice site (Gopal, 2005). Figure 3.8 shows a histogram of the total AG counts per sequence. It is evident from this plot that the number of AGs per sequence varies widely. The mean of this distribution is 17.47 while the standard deviation is 5.98. Furthermore, Figure 3.9 shows a histogram of the number of AGs (including the splice junction AG) found in the first 100 nucleotides upstream of the splice junction. Roughly 42 percent of the sequences contain no additional AGs in the first 100 nucleotides upstream of the splice junction while 27 percent possess one AG in that region.
Figure 3.8: Histogram of the total number of AGs per sequence.

Figure 3.9: Histogram of the number of AGs per sequence in the first 100 nucleotides upstream of the splice junction.
Figure 3.10 shows a histogram of the distance between the splice junction AG and the next upstream AG. This plot shows that there are quite a large number of sequences that have another AG in close proximity to the actual splice junction. In fact, roughly 19 percent of all the sequences have an additional AG in the first 30 nucleotides upstream of the splice junction. This finding was somewhat unexpected since it was assumed that the presence of two purines that close to the splice junction would have a negative impact on the splicing process (Siegel, 2005). Typically, the immediate upstream region is very pyrimidine rich as indicated by the presence of the pyrimidine tract.

![Histogram of the Distances between the Splice Junction AG and the next Upstream AG](image)

**Figure 3.10:** Histogram of the distance between the splice junction AG and the next upstream AG.
3.8 Alternative Splicing

Due to the presence of some AGs in close proximity to the actual splice junction the possibility of alternative trans-splicing was investigated. The second-order log odds model (3’ to 5’ direction) was used to score all the AGs in a set of 34 sequences. In these sequences there was a total of 452 AGs while 70 of them were EST-indicated alternative trans-splice sites. Two different methods were employed for this analysis. The first method scored the 100 nucleotides upstream of the AG under consideration. Using this method 60 of the 70 alternative splice sites were identified as splice sites. The second approach scored the upstream region of the AG in consideration up to the next AG. For this method an inter-AG distance threshold of 30 nucleotides was used in order to obtain a meaningful log odds score. This procedure only identified 32 of the 70 alternative splice sites. One caveat of this method was that many AGs were not scored due to small inter-AG distance.
4. DISCUSSION

4.1 Model Selection

The ultimate goal of this project was to identify trans-splicing sites using probabilistic methods. Careful analysis of the nucleotides in the upstream region of the splice junction revealed that there is a slow transition in the nucleotide frequencies from purines to pyrimidines. This gradual shift starts roughly 200 nucleotides upstream of the splice junction. These findings combined with the results from the entropy analysis indicate that the distribution of single and dinucleotides depends on the position within the sequence. The fact that the shift in nucleotide distributions is dependent on the position within the sequence violates the assumption of homogeneity which is used for Hidden Markov Models. In practice, this means that there must be a clear demarcation between the two states (splice versus nonsplice region), which is not the case in this data set.

The results of the Chi Square tests to determine the order of the Markov Chain as well as the maximal dependency decomposition revealed that local dependencies are an inherent part of the trans-splicing signal. Thus, the use of a Markov Chain, which captures local nucleotide dependencies, is appropriate to model the signal. The log odds metric, which is a ratio of the two Markov Chains representing the “plus” and “minus” model, can be utilized to discriminate between the splice and nonsplice regions. As the results show, this method is very powerful yielding an overall accuracy of more than 89 percent.
4.2 Maximal Dependence Decomposition and the Importance of Position 26

The maximal dependence decomposition revealed that position 26 exerts most influence over all other positions in the 400 nucleotide upstream region. This position is located in the vicinity of the suspected branch-point which plays an essential role in the trans-splicing process. Studies have shown this branch-point consists of a unique adenosine (A) that is part of a highly conserved consensus sequence. In *Saccharomyces cerevisiae* this consensus sequence (TACTAAC) is located between 18 and 37 nucleotides upstream of the splice junction and is stringently conserved. In higher eukaryotes the consensus sequence (YNCTRA\text{Y}) is slightly more flexible (Lücke, 2005). Recently, a study was conducted to elucidate the branch-point signal in trypanosomes using 16 sequences. However, no consensus sequence could be found and so far the branch-point in these organisms has remained a mystery (Lücke, 2005).

Nonetheless, it could be that an adenine in position 26, with the next most influential position at 23, actually represents the branch-point. Similarly, a thymine in position 26 with its next influential position at 25 could be an indicator of the branch-point assuming that there is some similarity to the mammalian and yeast consensus sequence (Lücke, 2005). The interpretation of a guanine or cytosine in that position is not as clear and needs to be investigated further.
4.3 Alternative Splice Site Analysis

In order to test the applicability of the second-order log odds model, a set of sequences with EST-indicated alternative trans-splice sites was used. It should be noted that this data set was very small. Furthermore, more experimental data is needed to truly assess the performance of this prediction. While there is evidence that the EST-indicated sites are actually involved with trans-splicing, nothing can be said about other AGs in close proximity. Thus, experimental data is essential so that these sites can be classified as splice or non-splice sites. Given this information, the sensitivity, specificity, and accuracy of this prediction can be calculated.
5. CONCLUSION

The accurate identification of trans-splicing sites in *Leishmania major* is essential in order to understand the unusual biology of this organism. With the development of the second-order log odds model we can now accurately predict almost 90 percent of these sites. In addition, this method can be employed to detect alternative splice sites. Furthermore, this study may have revealed the first real evidence of the branch-point signal which plays an indispensable role in the trans-splicing process.

While there are still many unanswered questions that need to be further investigated, the development of this new method will significantly contribute to surmount these challenges by helping to accurately predict the splice sites and by providing a framework to test hypotheses about the mechanism of trans-splicing.
6. REFERENCES


7. APPENDICES
### Appendix A

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Number of Sequences</th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0.200</td>
<td>0.264</td>
<td>0.195</td>
<td>0.341</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.150</td>
<td>0.255</td>
<td>0.283</td>
<td>0.313</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.185</td>
<td>0.283</td>
<td>0.243</td>
<td>0.290</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>0.183</td>
<td>0.238</td>
<td>0.225</td>
<td>0.354</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>0.173</td>
<td>0.268</td>
<td>0.225</td>
<td>0.335</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>0.183</td>
<td>0.261</td>
<td>0.224</td>
<td>0.332</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>0.214</td>
<td>0.239</td>
<td>0.210</td>
<td>0.336</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>0.157</td>
<td>0.220</td>
<td>0.252</td>
<td>0.371</td>
</tr>
<tr>
<td>9</td>
<td>13</td>
<td>0.178</td>
<td>0.242</td>
<td>0.266</td>
<td>0.314</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
<td>0.209</td>
<td>0.275</td>
<td>0.225</td>
<td>0.291</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>0.173</td>
<td>0.253</td>
<td>0.257</td>
<td>0.319</td>
</tr>
<tr>
<td>12</td>
<td>7</td>
<td>0.198</td>
<td>0.285</td>
<td>0.242</td>
<td>0.276</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>0.204</td>
<td>0.252</td>
<td>0.229</td>
<td>0.315</td>
</tr>
<tr>
<td>14</td>
<td>10</td>
<td>0.195</td>
<td>0.249</td>
<td>0.254</td>
<td>0.302</td>
</tr>
<tr>
<td>15</td>
<td>10</td>
<td>0.198</td>
<td>0.217</td>
<td>0.234</td>
<td>0.352</td>
</tr>
<tr>
<td>16</td>
<td>12</td>
<td>0.180</td>
<td>0.250</td>
<td>0.248</td>
<td>0.322</td>
</tr>
<tr>
<td>17</td>
<td>14</td>
<td>0.174</td>
<td>0.246</td>
<td>0.240</td>
<td>0.340</td>
</tr>
<tr>
<td>18</td>
<td>13</td>
<td>0.180</td>
<td>0.238</td>
<td>0.258</td>
<td>0.324</td>
</tr>
<tr>
<td>19</td>
<td>16</td>
<td>0.192</td>
<td>0.259</td>
<td>0.235</td>
<td>0.314</td>
</tr>
<tr>
<td>20</td>
<td>8</td>
<td>0.168</td>
<td>0.264</td>
<td>0.225</td>
<td>0.343</td>
</tr>
<tr>
<td>21</td>
<td>20</td>
<td>0.208</td>
<td>0.262</td>
<td>0.229</td>
<td>0.303</td>
</tr>
<tr>
<td>22</td>
<td>10</td>
<td>0.189</td>
<td>0.252</td>
<td>0.257</td>
<td>0.302</td>
</tr>
<tr>
<td>23</td>
<td>12</td>
<td>0.195</td>
<td>0.255</td>
<td>0.240</td>
<td>0.310</td>
</tr>
<tr>
<td>24</td>
<td>14</td>
<td>0.187</td>
<td>0.251</td>
<td>0.247</td>
<td>0.316</td>
</tr>
<tr>
<td>25</td>
<td>20</td>
<td>0.189</td>
<td>0.271</td>
<td>0.227</td>
<td>0.313</td>
</tr>
<tr>
<td>26</td>
<td>21</td>
<td>0.186</td>
<td>0.256</td>
<td>0.239</td>
<td>0.319</td>
</tr>
<tr>
<td>27</td>
<td>21</td>
<td>0.180</td>
<td>0.278</td>
<td>0.230</td>
<td>0.313</td>
</tr>
<tr>
<td>28</td>
<td>21</td>
<td>0.186</td>
<td>0.257</td>
<td>0.243</td>
<td>0.314</td>
</tr>
<tr>
<td>29</td>
<td>21</td>
<td>0.183</td>
<td>0.267</td>
<td>0.238</td>
<td>0.312</td>
</tr>
<tr>
<td>30</td>
<td>38</td>
<td>0.189</td>
<td>0.278</td>
<td>0.232</td>
<td>0.302</td>
</tr>
<tr>
<td>31</td>
<td>35</td>
<td>0.181</td>
<td>0.259</td>
<td>0.238</td>
<td>0.322</td>
</tr>
<tr>
<td>32</td>
<td>36</td>
<td>0.178</td>
<td>0.277</td>
<td>0.234</td>
<td>0.311</td>
</tr>
<tr>
<td>33</td>
<td>27</td>
<td>0.189</td>
<td>0.270</td>
<td>0.235</td>
<td>0.306</td>
</tr>
<tr>
<td>34</td>
<td>32</td>
<td>0.185</td>
<td>0.257</td>
<td>0.238</td>
<td>0.320</td>
</tr>
<tr>
<td>35</td>
<td>49</td>
<td>0.186</td>
<td>0.261</td>
<td>0.235</td>
<td>0.318</td>
</tr>
<tr>
<td>36</td>
<td>62</td>
<td>0.186</td>
<td>0.260</td>
<td>0.243</td>
<td>0.310</td>
</tr>
<tr>
<td></td>
<td>611</td>
<td>0.186</td>
<td>0.257</td>
<td>0.238</td>
<td>0.319</td>
</tr>
</tbody>
</table>

*Figure A.1:* Summary of the nucleotide proportions by chromosomes
Appendix B

Let
\[ X_{lt} = \begin{cases} 1, & \text{if position } t \text{ of sequence } l \text{ is nucleotide } j \\ 0, & \text{otherwise} \end{cases} \]

where
\( t = 1, 2, ..., T \)
\( l = 1, 2, ..., n \)
\( i = 1, 2, 3, 4 \) (A, T, G, T)

Define
\[ \sum_{l=1}^{n} \sum_{t=1}^{T} X_{lt} = T \]
\[ \sum_{l=1}^{n} X_{lt} = n \]
\[ \sum_{l=1}^{n} \sum_{t=1}^{T} X_{lt} = nT = X \]
\[ n_j(t) = \sum_{l=1}^{n} X_{lt} = X_{st} \]
\[ n = \sum_{t=1}^{T} n_j(t) = X_{t} \Rightarrow \text{total count of nucleotide } i \text{ in } n \text{ sequences} \]
\[ n_j(t) = \sum_{l=1}^{n} X_{lt} = X_{st} \Rightarrow \text{total count of nucleotides in position } t \]
\[ n_j(t, t+k) = \sum_{l=1}^{n} X_{lt} X_{lt+k} \Rightarrow \text{number of times we observe nucleotide } i \text{ in position } t \]

and nucleotide \( j \) in position \( t+k \) in \( n \) sequences
\( t = 1, 2, ..., T - k \) and \( k = 1, 2, ..., T - t \)
\[ \sum_{l=1}^{n} \sum_{j=1}^{4} n_j(t, t+k) = n - k \]
\[ n_j(t, t+k) = \sum_{l=1}^{n} n_j(t, t+k) \]
\[ n_j(t, t+k) = \sum_{l=1}^{n} n_j(t, t+k) \]
\[ \chi^2(t, k) = \sum_{l=1}^{n} \sum_{j=1}^{4} \left( \frac{n_j(t, t+k) - n_j(t, t+k)n_j(t, t+k)}{n - k} \right)^2 \]
\[ \chi^2(k, t) = \chi^2(t, k) \]

Figure B.1: General notation to determine the order of a Markov Chain using a Chi Square test
Appendix C

Single Nucleotide Relative Frequencies of the first 400 bp Upstream before the Splice Junction (N = 611)

Figure C.1: Single nucleotide relative frequencies of the first 400 base pairs upstream before the splice junction of the 611 sequences.
Appendix D

Entropy of the first 400 base pairs Upstream of the Splice Junction using Dinucleotide Frequencies (N = 611)

Figure D.1: Entropy of the first 400 base pairs upstream of the splice junction using dinucleotide frequencies.

Entropy of the first 400 base pairs Upstream of the Splice Junction using Trinucleotide Frequencies (N = 611)

Figure D.2: Entropy of the first 400 base pairs upstream of the splice junction using trinucleotide frequencies.
Appendix E

MDD Subset with an Adenine at Position 26

Figure E.1: Graph that shows row sums of Maximal Dependence Decomposition of the subset (n = 82) that has an adenine in position 26. The next most influential position is 23.

MDD Subset with a Thymine at Position 26

Figure E.2: Graph that shows row sums of Maximal Dependence Decomposition of the subset (n = 199) that has a thymine in position 26. The next most influential position is 25.
Figure E.3: Graph that shows row sums of Maximal Dependence Decomposition of the subset \( n = 104 \) that has a guanine in position 26. The next most influential position is 278.

Figure E.4: Graph that shows row sums of Maximal Dependence Decomposition of the subset \( n = 204 \) that has a cytosine in position 26. The next most influential position is 149.
### Appendix F

<table>
<thead>
<tr>
<th>Run</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>91.94</td>
<td>87.1</td>
<td>89.52</td>
</tr>
<tr>
<td>2</td>
<td>79.03</td>
<td>90.32</td>
<td>84.68</td>
</tr>
<tr>
<td>3</td>
<td>77.42</td>
<td>90.32</td>
<td>83.87</td>
</tr>
<tr>
<td>4</td>
<td>82.26</td>
<td>85.48</td>
<td>83.87</td>
</tr>
<tr>
<td>5</td>
<td>83.87</td>
<td>90.32</td>
<td>87.1</td>
</tr>
<tr>
<td>6</td>
<td>67.74</td>
<td>93.55</td>
<td>80.65</td>
</tr>
<tr>
<td>7</td>
<td>80.65</td>
<td>82.26</td>
<td>81.45</td>
</tr>
<tr>
<td>8</td>
<td>85.48</td>
<td>90.32</td>
<td>87.9</td>
</tr>
<tr>
<td>9</td>
<td>75.81</td>
<td>93.55</td>
<td>84.68</td>
</tr>
<tr>
<td>10</td>
<td>82.26</td>
<td>91.94</td>
<td>87.1</td>
</tr>
</tbody>
</table>

**Mean** 80.65 89.52 85.08

**Standard Deviation** 6.41 3.59 2.83

**Figure F.1**: Log odds results of ten runs using the first-order model in the 5’ to 3’ direction.

<table>
<thead>
<tr>
<th>Run</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80.65</td>
<td>90.32</td>
<td>85.48</td>
</tr>
<tr>
<td>2</td>
<td>77.42</td>
<td>90.32</td>
<td>83.87</td>
</tr>
<tr>
<td>3</td>
<td>85.48</td>
<td>93.55</td>
<td>89.52</td>
</tr>
<tr>
<td>4</td>
<td>90.32</td>
<td>90.32</td>
<td>90.32</td>
</tr>
<tr>
<td>5</td>
<td>79.03</td>
<td>85.48</td>
<td>82.26</td>
</tr>
<tr>
<td>6</td>
<td>77.42</td>
<td>91.94</td>
<td>84.68</td>
</tr>
<tr>
<td>7</td>
<td>82.26</td>
<td>83.87</td>
<td>83.06</td>
</tr>
<tr>
<td>8</td>
<td>82.26</td>
<td>88.71</td>
<td>85.48</td>
</tr>
<tr>
<td>9</td>
<td>77.42</td>
<td>90.32</td>
<td>83.87</td>
</tr>
<tr>
<td>10</td>
<td>77.42</td>
<td>83.87</td>
<td>80.65</td>
</tr>
</tbody>
</table>

**Mean** 80.97 88.87 84.92

**Standard Deviation** 4.29 3.35 3.02

**Figure F.2**: Log odds results of ten runs using the second-order model in the 5’ to 3’ direction.
<table>
<thead>
<tr>
<th>Run</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>90.32</td>
<td>96.77</td>
<td>93.55</td>
</tr>
<tr>
<td>2</td>
<td>80.65</td>
<td>83.87</td>
<td>82.26</td>
</tr>
<tr>
<td>3</td>
<td>80.65</td>
<td>88.71</td>
<td>84.68</td>
</tr>
<tr>
<td>4</td>
<td>74.19</td>
<td>88.71</td>
<td>81.45</td>
</tr>
<tr>
<td>5</td>
<td>83.87</td>
<td>79.03</td>
<td>81.45</td>
</tr>
<tr>
<td>6</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
</tr>
<tr>
<td>7</td>
<td>87.1</td>
<td>85.48</td>
<td>86.29</td>
</tr>
<tr>
<td>8</td>
<td>83.87</td>
<td>93.55</td>
<td>88.71</td>
</tr>
<tr>
<td>9</td>
<td>85.48</td>
<td>88.71</td>
<td>87.1</td>
</tr>
<tr>
<td>10</td>
<td>80.65</td>
<td>83.87</td>
<td>82.26</td>
</tr>
<tr>
<td>Mean</td>
<td>83.39</td>
<td>87.58</td>
<td>85.49</td>
</tr>
<tr>
<td></td>
<td>Standard Deviation</td>
<td>4.56</td>
<td>5.05</td>
</tr>
</tbody>
</table>

**Figure F.3:** Log odds results of ten runs using the first-order model in the 3' to 5' direction.

<table>
<thead>
<tr>
<th>Run</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>90.32</td>
<td>91.94</td>
<td>91.13</td>
</tr>
<tr>
<td>2</td>
<td>88.71</td>
<td>91.94</td>
<td>90.32</td>
</tr>
<tr>
<td>3</td>
<td>91.94</td>
<td>93.55</td>
<td>92.74</td>
</tr>
<tr>
<td>4</td>
<td>80.65</td>
<td>91.94</td>
<td>86.29</td>
</tr>
<tr>
<td>5</td>
<td>85.48</td>
<td>88.71</td>
<td>87.1</td>
</tr>
<tr>
<td>6</td>
<td>91.94</td>
<td>91.94</td>
<td>91.94</td>
</tr>
<tr>
<td>7</td>
<td>88.71</td>
<td>80.65</td>
<td>84.68</td>
</tr>
<tr>
<td>8</td>
<td>87.1</td>
<td>93.55</td>
<td>90.32</td>
</tr>
<tr>
<td>9</td>
<td>87.1</td>
<td>90.32</td>
<td>88.71</td>
</tr>
<tr>
<td>10</td>
<td>91.94</td>
<td>90.32</td>
<td>91.13</td>
</tr>
<tr>
<td>Mean</td>
<td>88.39</td>
<td>90.49</td>
<td>89.44</td>
</tr>
<tr>
<td></td>
<td>Standard Deviation</td>
<td>3.55</td>
<td>3.76</td>
</tr>
</tbody>
</table>

**Figure F.4:** Log odds results of ten runs using the second-order model in the 3' to 5' direction.