2010

Importance Sampling of Word Patterns in DNA and Protein Sequences

Hock Peng Chan
National University of Singapore

Nancy R. Zhang
University of Pennsylvania

Louis H. Y. Chen
National University of Singapore

Recommended Citation

Chan, H., Zhang, N. R., & Chen, L. Y. (2010). Importance Sampling of Word Patterns in DNA and Protein Sequences. *Journal of Computational Biology, 17* (12), 1697-1709. http://dx.doi.org/10.1089/cmb.2008.0233

At the time of publication, author Nancy R. Zhang was affiliated with Stanford University. Currently, she is a faculty member at the Statistics Department at the University of Pennsylvania.

Importance Sampling of Word Patterns in DNA and Protein Sequences

Abstract
The use of Monte Carlo evaluation to compute p-values of pattern counting test statistics is especially attractive when an asymptotic theory is absent or when the search sequence or the word pattern is too short for an asymptotic formula to be accurate. The drawback of applying Monte Carlo simulations directly is its inefficiency when p-values are small, which precisely is the situation of importance. In this paper, we provide a general importance sampling algorithm for efficient Monte Carlo evaluation of small p-values of pattern counting test statistics and apply it on word patterns of biological interest, in particular palindromes and inverted repeats, patterns arising from position specific weight matrices, as well as co-occurrences of pairs of motifs. We also show that our importance sampling technique satisfies a log efficient criterion.

Keywords
importance sampling, biological sequence analysis, motif analysis

Disciplines
Biostatistics | Computational Biology | Genetics and Genomics | Statistics and Probability

Comments
At the time of publication, author Nancy R. Zhang was affiliated with Stanford University. Currently, she is a faculty member at the Statistics Department at the University of Pennsylvania.

Importance Sampling of Word Patterns in DNA and Protein Sequences

Hock Peng Chan*

Department of Statistics and Applied Probability, National University of Singapore, Singapore 119260, Republic of Singapore

stachp@nus.edu.sg

Nancy Ruonan Zhang*†

Department of Statistics, Stanford University, Stanford, CA 94305-4065, USA

nzhang@stanford.edu

Louis H.Y. Chen

Institute for Mathematical Sciences, National University of Singapore, Singapore 118402, Republic of Singapore

matchyl@nus.edu.sg

July 27, 2007

Abstract

The use of Monte Carlo evaluation to compute p-values of pattern counting test statistics is especially attractive when an asymptotic theory is absent or when the search sequence or the word pattern is too short for an asymptotic formula to be accurate. The drawback of applying Monte Carlo simulations directly is its inefficiency when p-values are small, which precisely is the situation of importance. In this paper, we provide a general importance sampling algorithm for efficient Monte Carlo evaluation of small p-values of pattern counting test statistics and apply it on word patterns of biological interest, in particular palindromes and inverted repeats, patterns arising from position specific weight matrices, as well as co-occurrences of pairs of motifs. We also show that our importance sampling technique satisfies a log efficient criterion.

Key words: importance sampling, biological sequence analysis, motif analysis.

* joint first authors.

† corresponding author.

1 Introduction

Searching for matches to a word pattern in a stretch of biological sequence has become a recurring theme in computational biology. The search sequence is usually a stretch of DNA or protein sequence. The word pattern usually represents a functional site, such as a transcription factor binding site in DNA or a ligand docking site in protein. The word pattern, which is often called a motif, has traditionally been specified in a variety of ways (e.g. as an exact pattern, consensus pattern, or by a position specific weight matrix). Recently, there has also been much interest in more complex word patterns, such as co-occurrence of several different motifs within a close range of each other.

Often, we are interested in testing for over-representation of the word pattern in a given sequence. Analytic approximations for the significance values of such tests have been developed for special types of word patterns, and are surveyed recently in Mitrophanov and Borodovsky (2006). However, these approximations rely on various assumptions, which often fail to hold in practice. Furthermore, some word patterns of more recent interest, such as co-occurrence patterns of multiple motifs, do not yet have analytic approximations, and one needs to resort to Monte Carlo simulation. We propose a general methodology based on importance sampling that can be easily adapted to a wide range of word patterns, and that achieves significant gains in efficiency as compared to simple Monte Carlo. To motivate the reader, we start by briefly sketching several types of word patterns, which will be treated in detail in the next sections.

Palindromic patterns and inverted repeats. A palindrome is a DNA sequence that is equal to its own reverse complement. In the search for origin of replication of viruses (Leung, Schachtel and Yu, 1994 and Leung, Choi, Xia and Chen, 2005), it is of interest to test for overrepresentation of palindromes in a given segment of DNA sequence. Given a DNA sequence of length n which we assume to be stationary Markov, what is the significance of finding c palindromes of a given length? We treat this problem in Section 3.

Position specific weight matrices (PSWMs). PSWMs are commonly used to search for fixed-length motifs where each position can take on multiple values. A PSWM on an alphabet $\mathcal{X}$ is a matrix of the form $W = \{w_i(j) : 1 \leq i \leq \ell, 1 \leq j \leq \#\mathcal{X}\}$, where ℓ is the length of the motif. For background on PSWMs, see Mitrophanov and Borodovsky (2006). The score H of a word $\mathbf{v} = v_1 \cdots v_\ell$ from $\mathcal{X}^\ell$ is simply the sum of the weights

corresponding to the letter at each position. Thus

$$H(\mathbf{v}) = \sum_{i=1}^{\ell} w_i(v_i). \tag{1.1}$$

Given a PSWM W and a sequence of length n, what is the significance of finding c hits to W with score greater than t? Popular analytic approximations to the p-value are based on assuming independence of overlapping positions in the sequence. We show in Section 4 that the accuracy of such approximations can vary widely depending on the matrix, and illustrate an importance sampling alternative.

Co-occurrence patterns. Instead of single motifs, one may be interested in testing for over-representation of motif "modules": multiple motifs that occur within a short distance of each other. Robin, Daudin, Sagot, and Schbath (2002) gave analytic approximations for some types of co-occurrence patterns. However, these approximations rely on a strict syntax for the components of the module, and are not computationally feasible when the allowed gap between the motifs is comparable in size to the search sequence. We will illustrate the application of importance sampling to patterns studied by Robin et al. (2002) and to co-occurrence patterns of PSWMs in Section 5.

Before proceeding, we provide a short preview of Monte Carlo simulations of rare word patterns. Let N be a random variable counting the number of times a word pattern occurrs in a random sequence $\mathbf{s}$, which we assume follows an underlying probability measure P. We are interested in the probability p of the event $\{N \geq c\}$, where c is a given positive integer. In direct Monte Carlo, we generate independent copies $\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(K)}$ from P for some $K > 0$ and estimate p by

$$\widehat{p}_{\mathrm{D}} := K^{-1} \sum_{k=1}^{K} \mathbf{1}_{\{N^{(k)} \geq c\}}, \tag{1.2}$$

where $\mathbf{1}$ denotes the indicator function. The estimator $\widehat{p}_{\mathrm{D}}$ is unbiased, that is, $E_P \widehat{p}_{\mathrm{D}} = p$. We can measure the accuracy of an unbiased estimator $\widehat{p}$ by its relative standard error (RSE) $p^{-1}\sqrt{\mathrm{Var}(\widehat{p})}$, that is the ratio between the standard error of the estimator and the underlying probability. In practice, one can pre-select a constant $\alpha > 0$ and choose K large enough such that the RSE does not exceed α. For direct Monte Carlo, the RSE is equal to $\sqrt{(1-p)/Kp}$. Hence for small p, we need to choose the number of simulation runs K approximately equal to $\alpha^{-2}p^{-1}$. For example, to achieve $\alpha = 0.1$ for $p = 10^{-4}$, we will need about 1 million simulation runs.

When p is small, the event $\{N \geq c\}$ is rarely encountered and this is why direct Monte Carlo is inefficient. An alternative is to perform importance sampling via a change of measure by generating independent copies $\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(K)}$ from an alternative probability measure Q satisfying

$$P(B) > 0 \Rightarrow Q(B) > 0 \text{ for all } B \subset \{N \geq c\}. \tag{1.3}$$

Let the likelihood ratio $L = dQ/dP$. Then the importance sampling estimator

$$\widehat{p}_{\mathrm{I}} := K^{-1} \sum_{k=1}^{K} Y_k \text{ where } Y_k = L^{-1}(\mathbf{s}^{(k)}) \mathbf{1}_{\{N^{(k)} \geq c\}}, \tag{1.4}$$

is also unbiased for p and has a RSE bounded above by $p^{-1}\sqrt{E_Q Y_1^2/K}$. We will show how Q can be carefully chosen so that $E_Q Y_1^2$ is of the order p^2 and this will allow us to use a relatively small value of K even when p is small. The essential idea is to construct Q so that the event $\{N \geq c\}$ is encountered more frequently and L is uniformly large on $\{N \geq c\}$. Change of measure importance sampling procedures have been developed for sequential analysis cf. Siegmund (1976), bootstrapping cf. Johns (1988), Do and Hall (1991), communication systems cf. Cottrell, Fort and Malgouyres (1983), signal detection cf. Lai and Shan (1999), moderate deviations cf. Fuh and Hu (2004), and scan statistics cf. Chan and Zhang (2007). The widespread use of Monte Carlo methods for the evaluation of p-values of word pattern test statistics for DNA and protein sequences necessitates a similar improvement of accuracy and computational efficiency for this important task.

To avoid repetitive descriptions, we first lay out two general algorithms in Section 2, one for handling $c = 1$ and the other for $c > 1$. The specification of some functions and parameters when applying the algorithm to palindromic patterns, PSWMs and co-occurrence patterns are then given in Sections 3, 4 and 5 respectively. Numerical studies are also given in Sections 4 and 5. For the interested reader, we provide the statement and proof of log efficiency of our importance sampling algorithm in Section 6.

2 The importance sampling algorithms

Let $\#B$ denote the number of elements in a set B. By selecting randomly from a finite set B, we shall mean that each $b \in B$ has probability $(\#B)^{-1}$ of being selected. For any two sequences $\mathbf{v} = v_1 \cdots v_m$ and $\mathbf{w} = w_1 \cdots w_r$, the notation $\mathbf{vw}$ shall denote the concatenated sequence $v_1 \cdots v_m w_1 \cdots w_r$. Moreover, the lengths $\ell(\mathbf{v}) = m$ and $\ell(\mathbf{w}) = r$. Let $s_1 \cdots s_n$

denote a sequence of length n with each s_i taking values from an alphabet $\mathcal{X}$. For example, if we were looking at DNA sequences, then $\mathcal{X} = \{a, c, g, t\}$, the alphabet of four nucleotides, while for protein sequences $\mathcal{X}$ is the set of 20 amino acids. For ease of implementation to be made clear later on, we introduce a dummy variable s_0 and let $\mathbf{s} = s_0 \cdots s_n$. Let $\mathcal{V} \subset \cup_{j=1}^{\infty} \mathcal{X}^j$ denote a class of word patterns of biological interest. Let N be the number of times that a word pattern from $\mathcal{V}$ occurs in $s_1 \cdots s_n$ with no overlaps allowed among the word patterns. For example, if $\mathcal{V} = \{aag, ggt, agt\}$, then $N = 3$ for the sequence $aaggtaagagt$, with "aag" appearing twice and "agt" once. The pattern "ggt" overlaps with the first "aag" and is not counted. For a more precise definition, define recursively $\tau(0) = 0$ and

$$\tau(j+1) = \inf\{i + \ell - 1 : s_i \cdots s_{i+\ell-1} \in \mathcal{V} \text{ for some } i > \tau(j)\} \text{ for } j \geq 0. \tag{2.1}$$

Then $N = \sup\{j : \tau(j) \leq n\}$.

The underlying model for the generation of $\mathbf{s}$ is a Markovian chain with an irreducible and aperiodic transition matrix $\Sigma = (\sigma_{xy})_{x,y \in \mathcal{X}}$. Thus,

$$P\{s_{i+1} = y | s_i = x\} = \sigma_{xy} \text{ for all } i \geq 0. \tag{2.2}$$

Given s_j known, we say that $s_{j+1} \cdots s_{j+r}$ are generated from Σ if they are generated using (2.2) for $i = j, \ldots, j + r - 1$. Let π denote the stationary distribution of the Markov chain following (2.2). The underlying model assumes that s_0 follows the stationary distribution.

Let us first consider the importance sampling algorithm for $c = 1$. Define $\sigma(v_1 \cdots v_i) = \prod_{j=1}^{i-1} \sigma_{v_j v_{j+1}}$ and let $\mathcal{W} \subset \cup_{j=1}^{\infty} \mathcal{X}^j$. Let β be a positive function such that

$$q(\mathbf{v}) := \beta(\mathbf{v})\sigma(\mathbf{v}) \tag{2.3}$$

is a probability mass function on $\mathcal{W}$.

In our importance sampling algorithm, we generate $\mathbf{s}^{(k)} = (\mathbf{s} = s_0 \cdots s_n)$ using the following steps.

ALGORITHM A (for $c = 1$)

1. Generate $\mathbf{v}^* \in \mathcal{W}$ from the probability mass function q.

2. Select i_0 randomly from $\{1, \ldots, n - \ell(\mathbf{v}^*) + 1\}$.

3. Generate s_0 from π and $s_1 \cdots s_{i_0 - 1}$ from Σ. Let $s_{i_0} \cdots s_{i_0 + \ell(\mathbf{v}^*) - 1} = \mathbf{v}^*$ and generate $s_{i_0 + \ell(\mathbf{v}^*)} \cdots s_n$ from Σ.

The likelihood ratio function

$$L(\mathbf{s}) = \sum_{\mathbf{v} \in \mathcal{W}} \Big\{ \beta(\mathbf{v})(n - \ell(\mathbf{v}) + 1)^{-1} \sum_{i=1}^{n-\ell(\mathbf{v})+1} [\mathbf{1}_{\{s_i \cdots s_{i+\ell(\mathbf{v})-1} = \mathbf{v}\}} / \sigma(s_{i-1} s_i)] \Big\}. \tag{2.4}$$

If the dummy variable s_0 is not generated, then the likelihood ratio expression is like (2.4) with the inner sum replaced by

$$[\mathbf{1}_{\{s_1\cdots s_{\ell(\mathbf{v})}=\mathbf{v}\}}/\pi(s_1)] + \sum_{i=2}^{n-\ell(\mathbf{v})+1} [\mathbf{1}_{\{s_i\cdots s_{i+\ell(\mathbf{v})-1}=\mathbf{v}\}}/\sigma(s_{i-1}s_i)].$$

It is desirable to select $\mathcal{W} = \mathcal{V}$ as this will ensure that the event of interest $\{N \geq 1\}$ is observed in every simulation run. However the need to have an easily implementable algorithm may require us to choose $\mathcal{W}$ larger than $\mathcal{V}$, see for example Sections 4 and 5.

For $c > 1$, the random insertion of c word patterns from $\mathcal{W}$ during importance sampling seems natural but this may result in a hard to compute likelihood ratio function. Instead we construct an auxiliary hidden Markov model that determines when the word patterns should be inserted. The likelihood ratio function can then be computed recursively.

The detailed steps involved in the generation of each $\mathbf{s}^{(k)}(= \mathbf{s} = s_0 \cdots s_n)$ is then as follows. Let $\ell_{\min} = \min_{\mathbf{v}\in\mathcal{W}} \ell(\mathbf{v})$, $\ell_{\max} = \max_{\mathbf{v}\in\mathcal{W}} \ell(\mathbf{v})$ and $\mathcal{W}_k = \{\mathbf{v} \in \mathcal{W} : \ell(\mathbf{v}) \leq k\}$ for all $k \geq \ell_{\min}$. Then

$$q_k(\mathbf{v}) := q(\mathbf{v}) \Big/ \sum_{\mathbf{u}\in\mathcal{W}_k} q(\mathbf{u})$$

is a probability mass function on $\mathcal{W}_k$. Note that for all $k \geq \ell_{\max}$, $\mathcal{W}_k = \mathcal{W}$ and $q_k = q$. In the case where all the words in $\mathcal{W}$ are of fixed length, Step 3 of Algorithm B involves only the probability mass function q. Select some $0 < \rho < 1$ of order n^{-1}, for example cn^{-1}.

ALGORITHM B (for $c > 1$)

1. Let $i = j = 1$ and generate s_0 from the stationary distribution π.

2. Let $T(j)$ be a geometric random variable satisfying

$$P\{T(j) = t\} = \rho(1-\rho)^t \quad \text{for } t = 0, 1, \ldots. \tag{2.5}$$

3. If $i+T(j) \leq n-\ell_{\min}+1$, generate $s_i \cdots s_{i+T(j)-1}$ from Σ, $\mathbf{v}^* \in \mathcal{W}_{n-(i+T(j))+1}$ from the probability mass function $q_{n-(i+T(j))+1}$ and let $s_{i+T(j)} \cdots s_{i+T(j)+\ell(\mathbf{v}^*)-1} = \mathbf{v}^*$. Otherwise, generate $s_i \cdots s_n$ from Σ and stop.

4. Increment i by $T(j) + \ell(\mathbf{v}^*)$ and j by 1. Go to step 2.

Let $L_0(s_0) = 1$ and

$$\begin{aligned} L_i(s_0 \cdots s_i) &= (1-\rho)^{\mathbf{1}_{\{i\leq n-\ell_{\min}+1\}}} L_{i-1}(s_0 \cdots s_{i-1}) \\ &+\rho \sum_{\mathbf{v}\in\mathcal{W}} L_{i-\ell(\mathbf{v})}(s_0 \cdots s_{i-\ell(\mathbf{v})}) q_{n-(i-\ell(\mathbf{v}))+1}(\mathbf{v}) \mathbf{1}_{\{s_{i-\ell(\mathbf{v})+1}\cdots s_i=\mathbf{v}\}}/\sigma(s_{i-\ell(\mathbf{v})}\mathbf{v}). \end{aligned} \tag{2.6}$$

Then $L(\mathbf{s}) = L_n(\mathbf{s})$.

In most practical applications, $\mathcal{V}$ can be very large but the identification of whether $\mathbf{v}$ belongs to $\mathcal{V}$ should take only $\ell(\mathbf{v})$ time. If β is chosen such that the selection of $\mathbf{v}^* \in \mathcal{W}_k$ is of order k, then the computational time of our importance sampling algorithm is of the same order as direct Monte Carlo. In contrast, recursive computation to evaluate p directly via suffix trees, see for example Gusfield (1997), grows in general with $\#\mathcal{V}$. Moreover, our algorithm entertains modifications of $\mathcal{V}$ easily. Further details on the implementation of Algorithms A and B in concrete examples are given in the next three sections.

3 Palindromic patterns and inverted repeats

In Masse, Karlin, Schachtel and Mocarski (1992), clusters of palindromic patterns were found near origin of replications of viruses. Subsequently, analytical approximations were developed to determine significance of observed clusters, see for example Leung, Schachtel and Yu (1994) and Leung, Choi, Xia and Chen (2005). Let $\mathcal{X} = \{a, c, g, t\}$, the alphabet of four nucleotides. These nucleotides can be divided into two complementary base pairs with a and t forming a pair and c and g forming the second pair. We denote this relation by writing $a^c = t$, $t^c = a$, $c^c = g$ and $g^c = c$. A palindromic pattern of length $\ell = 2m$ is a DNA segment that can be expressed in the form $v_1 \cdots v_m v_m^c \cdots v_1^c$. For example, $\mathbf{v} = acgcgt$ is a palindromic pattern. Note that the complement of $\mathbf{v}$, that is the word obtained by replacing each letter of $\mathbf{v}$ by its complement, is $tgcgca$, which is just $\mathbf{v}$ read backwards. This interesting property explains the use of the terminology "palindromic pattern".

Let $m \geq 1$ and $\mathcal{V}$ the class of all palindromic patterns of length $2m$. Thus $\#\mathcal{V} = 4^m$. In our importance sampling algorithm, we shall choose $\mathcal{W} = \mathcal{V}$. The generation of $\mathbf{v}^* \in \mathcal{V}$ in step 1 of Algorithm A and step 3 of Algorithm B requires us to first derive a new transition matrix from Σ. Note first that each transition $x \to y$ on the left side of palindrome $\mathbf{v}$ has a complementary transition $y^c \to x^c$ on the right side. Let $\Xi = (\xi_{xy})_{x,y\in\mathcal{X}}$ with

$$\xi_{xy} = \sigma(xy)\sigma(y^c x^c) \text{ for all } x, y \in \mathcal{X}. \tag{3.1}$$

Let $\lambda(< 1)$ be the largest eigenvalue of Ξ and r the corresponding non-negative real eigenvector. Hence $\Xi r = \lambda r$. This ensures that

$$Z := (\zeta_{xy})_{x,y\in\mathcal{X}}, \text{ where } \zeta_{xy} = \lambda^{-1}\xi_{xy} r(y)/r(x) \tag{3.2}$$

is a probability transition matrix. The generation of $\mathbf{v}^* \in \mathcal{V}$ can be executed in the following manner.

(a) Select v_1 randomly from $\mathcal{X}$.

(b) Generate $v_2 \cdots v_m$ from the transition matrix Z and let $\mathbf{v}^* = v_1 \cdots v_m v_m^c \cdots v_1^c$.

For the above selection process, the probability mass function

$$\begin{aligned} q(\mathbf{v}) &= \sigma(v_1 \cdots v_m)\sigma(v_m^c \cdots v_1^c) r(v_m)/[4\lambda^{m-1} r(v_1)] \\ &= \sigma(\mathbf{v}) r(v_m)/[4\lambda^{m-1}\sigma(v_m v_m^c) r(v_1)]. \end{aligned} \tag{3.3}$$

Inverted repeats are derived from palindromic patterns by inserting an arbitrary DNA segment in the middle of the pattern. The class of word patterns can be expressed in the form

$$\mathcal{V} = \{v_1 \cdots v_m \mathbf{z} v_m^c \cdots v_1^c : \ell(\mathbf{z}) \le d_2\} \tag{3.4}$$

for some $d_2 \ge 0$. For $2m \le k \le d_2 + 2m$, $\mathcal{W}_k = \{v_1 \cdots v_m \mathbf{z} v_m^c \cdots v_1^c : \ell(z) \le k - 2m\}$. To simulate positive or large counts of words from $\mathcal{W}_k$ efficiently, we modify the algorithm above to obtain the following.

(a) Select v_1 from $\mathcal{X}$ and d randomly from $\{0, \ldots, k - 2m\}$.

(b) Generate $v_1 \cdots v_m$ from the transition matrix Z. Let $z_0 = v_m$ and generate $\mathbf{z} = z_1 \cdots z_d$ from the underlying transition matrix Σ. Define $\mathbf{v}^* = v_1 \cdots v_m \mathbf{z} v_m^c \cdots v_1^c$.

For this algorithm,

$$\begin{aligned} q_k(\mathbf{v}^*) &= \sigma(v_1 \cdots v_m)\sigma(v_m^c \cdots v_1^c)\sigma(v_m \mathbf{z}) r(v_m)/[4(k - 2m + 1)\lambda^{m-1} r(v_1)] \\ &= \sigma(\mathbf{v}^*) r(v_m)/[4(k - 2m + 1)\lambda^{m-1} r(v_1)\sigma(z_{\ell(\mathbf{z})} v_m^c)], \end{aligned} \tag{3.5}$$

with $q = q_{d_2+2m}$.

4 Position specific weight matrices (PSWMs)

The score of a sequence by a PSWM is introduced in (1.1) in Section 1. We are interested here in the motif class $\mathcal{V} = \{\mathbf{v} : H(\mathbf{v}) > t\}$ for a given weight matrix W and threshold $t > 0$. Unlike in the previous section, we choose, for our importance sampling algorithm, $\mathcal{W} = \mathcal{X}^\ell$. Let $\theta > 0$. We would like to generate $\mathbf{v}^*$ from the probability mass function

$$q(\mathbf{v}) = e^{\theta H(\mathbf{v})}\sigma(\mathbf{v})/\Lambda(\theta), \tag{4.1}$$

where $\Lambda(\theta) = \sum_{\mathbf{v}\in\mathcal{X}^\ell} e^{\theta H(\mathbf{v})}\sigma(\mathbf{v})$. Consider the backward recursive relations

$$\begin{aligned} \Lambda_\ell(\theta, x) &= e^{\theta w_\ell(x)}, \\ \Lambda_i(\theta, x) &= e^{\theta w_i(x)} \sum_{y\in\mathcal{X}} \sigma(xy)\Lambda_{i+1}(\theta, y) \text{ for all } x \in \mathcal{X} \text{ and } i = 1, \ldots, \ell - 1. \end{aligned} \tag{4.2}$$

Then $\Lambda(\theta) = \sum_{x\in\mathcal{X}} \Lambda_1(\theta, x)$. The generation of $\mathbf{v}^* = v_1^* \cdots v_\ell^* \in \mathcal{X}^\ell$ from q can then be done via the Markovian relations

$$\begin{aligned} P\{v_1^* = x\} &= \Lambda_1(\theta, x)/\Lambda(\theta), \\ P\{v_{i+1}^* = y | v_i^* = x\} &= e^{\theta w_i(x)}\sigma(xy)\Lambda_{i+1}(\theta, y)/\Lambda_i(\theta, x) \text{ for } i = 1, \ldots, \ell - 1. \end{aligned} \tag{4.3}$$

Let $\Lambda_i'(\theta, x) = \frac{d}{d\theta}\Lambda_i(\theta, x)$, $\Lambda_i''(\theta, x) = \frac{d^2}{d\theta^2}\Lambda_i(\theta, x)$ and similarly for $\Lambda'(\theta)$ and $\Lambda''(\theta)$. For optimal performance, we should select $\theta = \theta^*$ to satisfy

$$E_{\theta^*} H(\mathbf{v}) = \Lambda'(\theta^*)/\Lambda(\theta^*) = t. \tag{4.4}$$

The root of (4.4) can be easily obtained by using the Newton-Ralphson method. Start with an initial guess $\theta = \theta_0$, for example $\theta_0 = 0$. We then compute $\Lambda'(\theta)$ via the recursive relations

$$\begin{aligned} \Lambda_\ell'(\theta, x) &= w_\ell(x)e^{\theta w_\ell(x)} \\ \Lambda_i'(\theta, x) &= e^{\theta w_i(x)} \sum_{y\in\mathcal{X}} \sigma(xy)[\Lambda_{i+1}'(\theta, y) + w_i(x)\Lambda_{i+1}(\theta, y)], \end{aligned} \tag{4.5}$$

and $\Lambda''(\theta)$ via

$$\begin{aligned} \Lambda_\ell''(\theta, x) &= w_\ell^2(x)e^{\theta w_\ell(x)}, \\ \Lambda_i''(\theta, x) &= e^{\theta w_i(x)} \sum_{y\in\mathcal{X}} \sigma(xy)[\Lambda_{i+1}''(\theta, y) + 2w_i(x)\Lambda_{i+1}'(\theta, y) + w_i^2(x)\Lambda_{i+1}(\theta, y)]. \end{aligned} \tag{4.6}$$

Then $\Lambda'(\theta) = \sum_{x\in\mathcal{X}} \Lambda_1'(\theta, x)$ and $\Lambda''(\theta) = \sum_{x\in\mathcal{X}} \Lambda_1''(\theta, x)$. By the Taylor expansion

$$\frac{\Lambda'(\theta^*)}{\Lambda(\theta^*)} = \frac{\Lambda'(\theta)}{\Lambda(\theta)} + (\theta^* - \theta)\Big\{\frac{\Lambda''(\theta)}{\Lambda(\theta)} - \Big[\frac{\Lambda'(\theta)}{\Lambda(\theta)}\Big]^2\Big\} + O((\theta^* - \theta)^2)$$

and (4.4), our next guess of θ^* is

$$\theta_{\text{new}} = \theta + [t - \Lambda'(\theta)/\Lambda(\theta)]/\{[\Lambda''(\theta)/\Lambda(\theta)] - [\Lambda'(\theta)/\Lambda(\theta)]^2\}.$$

Then, recompute $\Lambda(\theta), \Lambda'(\theta)$ and $\Lambda''(\theta)$ using (4.2), (4.5) and (4.6) with $\theta = \theta_{\text{new}}$ and iterate until convergence.

Example 1. It is often of interest to score a short sequence, which is usually the promoter sequence of some gene, for a hit to a PSWM. Analytic approximations for p-values

of PSWM hits usually assumes independence of adjacent overlapping positions. For example, a popular method is to compute $p_t = P(H(v) \geq t)$ using the method in Huang et al. (2004) and then use the p-value approximation

$$1 - \sum_{i=0}^{c-1} \binom{n-l+1}{i} (1-p_t)^{n-l+1-i} p_t^i.$$

We demonstrate our method, and explore the accuracy of analytic approximations, using two hypothetical PSWMs, assumed to be defined on the alphabet $\{a, c, g, t\}$:

$$W_{\text{rep}} = \begin{pmatrix} 2 & \dots & 2 \\ 1 & \dots & 1 \\ 1 & \dots & 1 \\ 1 & \dots & 1 \end{pmatrix}_{4\times 12},$$

which is a highly repetitive, and

$$W_{\text{norep}} = \begin{pmatrix} 2 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 & 1 & 1 & 2 & 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 & 2 & 1 & 1 & 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 2 \end{pmatrix},$$

which only repeats at the ninth position. We define a hit to either W_1 or W_2 to be a score greater than 20, and we are interested in computing the p-value of $c = 1$ hit in a sequence of $n = 200$, which is roughly the length of a promoter sequence in bacteria. Table 1 compares the results from importance sampling and direct Monte Carlo, using 1000 iterations each. We see that the importance sampling method gives a much smaller RSE, achieving a 25-100 fold increase in efficiency. The tables also show the analytic approximation to be grossly wrong for W_{rep} but acceptable for W_{norep}. In reality, the accuracy of analytic approximations should fall somewhere in between the results for these two matrices.

Example 2. Databases such as TRANSFAC, JASPAR, and SCPD curate PSWMs for families of transcription factors, which can range between 4 to more than 20 bases in length. Here we use the PSWM for the trancription factor SWI5 in yeast, obtained from SCPD (Zhu and Zhang, 1999) as an illustration. For the cut-off we choose $t = 50$. It is customary to use the 700 base pairs preceding the transcription start site as the promoter sequence in yeast. The background transition matrix Σ is estimated using all of the 700 base pair promoter sequences extracted from the yeast genome. Table 2 shows the p-value estimates obtained from direct Monte Carlo, importance sampling, and analytic approximation for $c = 1, 2$. For

$c = 1$, importance sampling almost quadruple the gain in efficiency over direct Monte Carlo for 1000 iterations. For $c = 2$, the efficiency of importance sampling is roughly 100 times that of direct Monte Carlo 5000 iterations. We also see that when $c = 2$, the estimate obtained from analytic approximation is not accurate.

5 Structured Motifs and Pairwise Co-occurrences

In a more detailed analysis of cis-regulation of gene transcription, one can search promoter sequences for cis-regulatory modules (CRM) instead of single motifs. A CRM can be defined as a collection of fixed length motifs that are located in a fixed order in proximity to each other. They are signals for cooperative binding of transcription factors, and are important to the study of combinatorial regulation of genes. CRMs have recently been used successfully to gain a deeper understanding of gene regulation cf. Chiang, Moses, Kellis, Lander and Eisen (2003), Zhang, Wildermuth and Speed (2007), and Zhou and Wong (2004). We focus here on the simplest type of CRM: A pair of fixed length motifs separated by a gap sequence of variable length.

5.1 Structured Motifs

We give two examples of such co-occurring motif pairs. In the first example, we consider what is sometimes called "structured motifs", where the motifs are essentially fixed word patterns $\mathbf{x}$ and $\mathbf{y}$ with an allowance for the mutation of up to one letter in $\mathbf{xy}$. More precisely, the motif class is

(5.1)

$$\mathcal{V} = \Big\{\mathbf{x}'\mathbf{z}\mathbf{y}' : d_1 \le \ell(\mathbf{z}) \le d_2, \ell(\mathbf{x}') = \ell(\mathbf{x}), \ell(\mathbf{y}') = \ell(\mathbf{y}), \sum_{i=1}^{\ell(\mathbf{x})} \mathbf{1}_{\{x_i \neq x_i'\}} + \sum_{i=1}^{\ell(\mathbf{y})} \mathbf{1}_{\{y_i' \neq y_i\}} \le 1\Big\}.$$

Robin, Daudin, Sagot and Schbath (2002) gave the background and some analytical approximations for these patterns. For our importance sampling algorithm, we shall let $\mathcal{W} = \mathcal{V}$. The selection of $\mathbf{v}^* = \mathbf{x}'\mathbf{z}\mathbf{y}' \in \mathcal{V}$ with probability mass function q_k for $d_1 + \ell(\mathbf{xy}) \le k \le d_2 + \ell(\mathbf{xy})$ shall proceed in the following manner.

(a) Select d randomly from $\{d_1, \ldots, k - \ell(\mathbf{xy})\}$ and r randomly from $\{0, \ldots, \ell(\mathbf{xy})\}$.

(b) If $1 \le r \le \ell(\mathbf{x})$, select x_r' randomly from $\mathcal{X} \setminus \{x_r\}$ and let $x_t' = x_t$ for all $t \neq r$, $\mathbf{y}' = \mathbf{y}$. If $r > \ell(\mathbf{x})$, select $y_{r-\ell(\mathbf{x})}'$ randomly from $\mathcal{X} \setminus \{y_{r-\ell(\mathbf{x})}\}$ and let $\mathbf{x}' = \mathbf{x}$, $y_t' = y_t$ for all $t \neq r - \ell(\mathbf{x})$. If $r = 0$, let $\mathbf{x}' = \mathbf{x}$ and $\mathbf{y}' = \mathbf{y}$.

(c) Let $z_0 = x'_{\ell(\mathbf{x})}$ and generate $\mathbf{z} = z_1 \cdots z_d$ from the underlying transition matrix Σ.

Then

$$\begin{aligned} q_k(\mathbf{v}^*) &= \sigma(x'_{\ell(\mathbf{x})}\mathbf{z})/[3^{\mathbf{1}_{\{r>0\}}}(k-\ell(\mathbf{xy})-d_1+1)(\ell(\mathbf{xy})+1)] \\ &= [3^{\mathbf{1}_{\{r>0\}}}(k-\ell(\mathbf{xy})-d_1+1)(\ell(\mathbf{xy})+1)\sigma(\mathbf{x}')\sigma(z_d\mathbf{y}')]^{-1}\sigma(\mathbf{v}^*), \end{aligned} \tag{5.2}$$

with $q = q_{d_2+\ell(\mathbf{xy})}$.

Example 3. We perform a simulation study of eight structural motifs selected for their high frequency of occurrences in part of the *Bacillus subtilis* DNA dataset. We consider $[d_1, d_2] = [16, 18], [10, 20]$ and $[5, 50]$ with length of DNA segment $n = 100$, threshold level $c = 1$ and transition matrix

$$\Sigma = \begin{pmatrix} 0.35 & 0.16 & 0.18 & 0.31 \\ 0.33 & 0.20 & 0.15 & 0.32 \\ 0.32 & 0.22 & 0.19 & 0.27 \\ 0.25 & 0.20 & 0.19 & 0.35 \end{pmatrix}$$

on the state space $\mathcal{X} = \{a, c, g, t\}$. In addition to our importance sampling estimate, we also obtain analytical estimates from Robin et al. (2002) and direct Monte Carlo estimates. The analytical estimates are computed via recursive methods with computation time that grows exponentially with $d_2 - d_1$.

To illustrate the flexibility of our importance sampling technique to deal with more complex situations, we also use it to obtain a combined p-value for all eight motifs. This method can be used in general for the estimation of $p := P\{\max_{1\le m\le M}(N^{(m)} - c_m) \ge 0\}$, where c_m are positive integers and $N^{(m)}$ is the total word count from a class of words $\mathcal{V}^{(m)}$. Let $q^{(m)}$ be a probability mass function that is efficient for simulating words from $\mathcal{V}^{(m)}$. The kth simulation run, $1 \le k \le K$, consists of the following steps.

1. Select m_k randomly from $\{1, \ldots, M\}$.

2. Generate the sequence $\mathbf{s}^{(k)}$ using either Algorithm A or B, whichever is appropriate, with $q = q^{(m_k)}$ and for Algorithm B, $\rho = \rho^{(m_k)}$.

3. Compute the likelihood ratio $L^{(m_k)}$ via either (2.4) or (2.6) with $\beta = \beta^{(m_k)}$, $q_k = q_k^{(m_k)}$ and $\rho = \rho^{(m_k)}$.

Then

$$\widehat{p} = K^{-1}\sum_{k=1}^{K}[L^{(m_k)}(\mathbf{s}^{(k)})]^{-1}(M/\#\{m : N^{(m)}(\mathbf{s}^{(k)}) \ge c_m\})\mathbf{1}_{\{N^{(m_k)}(\mathbf{s}^{(k)})\ge c_{m_k}\}} \tag{5.3}$$

is unbiased for p. The key feature in (5.3) is the correction term $\#\{m : N^{(m)}(\mathbf{s}^{(k)}) \geq c_m\}$. Without this term, $\widehat{p}$ is an unbiased estimator for the Bonferroni upper bound $\sum_{m=1}^{M} P\{N^{(m)} \geq c_m\}$. The correction term adjusts the estimator downwards proportionately in the sample space where more than one threshold c_m are exceeded.

The results of our simulation study are summarized in Table 3. The variance reduction, that is the ratio of the standard error squared, is substantial when the importance sampling technique is used. In fact, the direct Monte Carlo estimate is often unreliable. Such savings in computation time is valuable both to the end user and also to the researcher trying to test the reliability of his or her analytical estimates on small p-values. We observe for example, that the numerical estimates given in Robin et al. (2002) tends to underestimate the true underlying probability but are relatively accurate. However, the computational needs of these estimates grows exponentially with $d_2 - d_1$ and hence are computed and displayed only for $[d_1, d_2] = [16, 18]$.

5.2 Pairwise Co-occurences of PSWM Hits

We next look at CRMs involving long-range interactions between transcription factor binding sites represented by PSWMs, as described, for example, in Chiang et al. (2003) and Zhang et al. (2007). In Example 3, we showed that the necessity of importance sampling methods become clear when the range of the allowed gap $d_2 - d_1$ is large. We will explore this situation further in the next example with word patterns specified by PSWMs $W_1 = \{w_{1i}(j) : 1 \leq i \leq \ell_1, 1 \leq j \leq \mathcal{X}\}$ and $W_2 = \{w_{2i}(j) : 1 \leq i \leq \ell_2, 1 \leq j \leq \mathcal{X}\}$. Let

$$\begin{aligned} \mathcal{V} &= \{\mathbf{xzy} : d_1 \leq \ell(\mathbf{z}) \leq d_2, H_1(\mathbf{x}) > t, H_2(\mathbf{y}) > u\}, \\ &\text{where } H_1(\mathbf{x}) = \sum_{i=1}^{\ell_1} w_{1i}(x_i) \text{ and } H_2(\mathbf{y}) = \sum_{i=1}^{\ell_2} w_{2i}(y_i). \end{aligned} \tag{5.4}$$

For our importance sampling algorithm, first find the roots θ_1^* and θ_2^* of the equations $E_{\theta_1^*} H_1(\mathbf{x}) = t$ and $E_{\theta_2^*} H_2(\mathbf{y}) = u$ using the Newton-Ralphson method as given in Section 4. Let $\Lambda_i^{(1)}$ and $\Lambda^{(1)}$ be the moment generating functions corresponding to W_1 and $\Lambda_i^{(2)}$, $\Lambda^{(2)}$ corresponding to W_2, see (4.2). To select $\mathbf{v}^*$ from q_k for $d_1 + \ell_1 + \ell_2 \leq k \leq d_2 + \ell_1 + \ell_2$, we do the following.

(a) Select d randomly from $\{d_1, \ldots, k - \ell_1 - \ell_2\}$.

(b) Select $\mathbf{x} \in \mathcal{X}^{\ell_1}$ with probability $e^{\theta_1^* H_1(\mathbf{x})} \sigma(\mathbf{x}) / \Lambda^{(1)}(\theta_1^*)$, generate $\mathbf{z} = z_1 \cdots z_d$ from the underlying transition matrix Σ with $z_0 = x_{\ell_1}$, and select $\mathbf{y} \in \mathcal{X}^{\ell_2}$ with probability

$e^{\theta_2^* H_2(\mathbf{y})}\sigma(\mathbf{y})/\Lambda^{(2)}(\theta_2^*)$. Then

$$q_k(\mathbf{xzy}) = e^{\theta_1^* H_1(\mathbf{x})+\theta_2^* H_2(\mathbf{y})}\sigma(\mathbf{xzy})/[(k-\ell_1-\ell_2-d_1+1)\Lambda^{(1)}(\theta_1^*)\Lambda^{(2)}(\theta_2^*)\sigma(z_d y_1)], \tag{5.5}$$

with $q = q_{d_2+\ell_1+\ell_2}$.

Example 4. Let W_1 and W_2 be the PSWMs for transcription factors SFF and MCM1 respectively in yeast, taken from SCPD (Zhu and Zhang, 1999). We let $t = 48$, $u = 110$. Both transcription factors are regulators of the cell cycle, and are known to cooperate in the regulation of downstream genes such as CLB1, CLB2, BUD4, and SWI5 (Spellman et al., 1998). As in Example 3, we let the promoter sequence be of length $n = 700$. We set the range of the allowed gap to be $d_1 = 0$, $d_2 = 100$. 5000 iterations of direct Monte Carlo give a p-value estimate of $(2.4 \pm 0.8) \times 10^{-3}$ with RSE of 0.33. 5000 importance sampling simulations give a p-value estimate of $(3.4 \pm 0.2) \times 10^{-3}$, with RSE of 0.075. We see that in this situation, importance sampling achieves substantial variance reduction over direct Monte Carlo.

6 Log Efficiency Theory

We start off with a discussion of the log efficient criterion widely adopted in the importance sampling literature, then show that Algorithms A and B, as applied in Sections 3-5, indeed satisfy this criterion. In an ideal situation, the importance measure Q satisfies

$$\sqrt{E_Q Y_1^2} = O(p) \text{ as } p \to 0. \tag{6.1}$$

Then the constraint on the RSE, $p^{-1}\sqrt{\mathrm{Var}_Q(\widehat{p}_{\mathrm{I}})/K} \le \alpha$, can be satisfied with a uniformly bounded K as $p \to 0$. If instead of (6.1), the importance measure Q satisfies

$$\sqrt{E_Q Y_1^2} = O(p^{1-\epsilon}) \text{ as } p \to 0 \text{ for all } \epsilon > 0, \tag{6.2}$$

then we say that the measure Q is log efficient, cf. Sadowsky and Bucklew (1990) and Dupuis and Wang (2005). Under (6.2), the RSE constraint can be satisfied with $\log K = o(|\log p|)$. If $Q = P$, that is when direct Monte Carlo is used, then $Y_1 \in \{0, 1\}$ so that $E_Q Y_1^2 = p$ and clearly (6.2) is not satisfied.

Remark 1 below provides a simple heuristic to ensure log efficiency of Algorithms A and B, namely to select probability mass function q [see (2.3)] such that $\mathcal{W} = \mathcal{V}$ and β is of the same order over all $\mathbf{v} \in \mathcal{V}$. Sometimes for ease of implementation, we may have to select $\mathcal{W}$ larger than $\mathcal{V}$, as in Sections 4 and 5.2, but log efficiency can still be attained if β is relatively small on $(\mathcal{W} \setminus \mathcal{V})$.

We preface our main results with the following preliminary lemma. We will assume throughout that Σ is fixed with $\sigma_0 := \min_{x,y\in\mathcal{X}} \sigma(xy)$ positive. There is also no loss of generality in assuming that $n \geq \ell_{\max}$. Let

$$\beta_0 = \inf_{\mathbf{v}\in\mathcal{V}} \beta(\mathbf{v}) \text{ and } \tau = \sum_{\mathbf{v}\in\mathcal{V}} \sigma(\mathbf{v}). \tag{6.3}$$

Also let $\lfloor\cdot\rfloor$ denote the greatest integer function.

Lemma 1. *Let $c = 1$ and $\gamma = \lfloor n/\ell_{\max}\rfloor \sigma_0\tau/\ell_{\max}$. Then*

$$p \geq \begin{cases} \frac{1}{4} & \text{if } \gamma > \frac{1}{2}, \\ \frac{\gamma}{2} & \text{if } \gamma \leq \frac{1}{2}. \end{cases} \tag{6.4}$$

Proof. Let $p_0 = \inf_{s_0\in\mathcal{X}} P_{s_0}\{s_1\cdots s_\ell \in \mathcal{V} \text{ for some } \ell\}$. Then

$$\begin{aligned} p_0 &\geq \inf_{s_0\in\mathcal{X}} \max_{\ell\leq\ell_{\max}} P_{s_0}\{s_1\cdots s_\ell \in \mathcal{V}\} \\ &\geq \inf_{s_0\in\mathcal{X}} \Big[\sum_{\ell=1}^{\ell_{\max}} P_{s_0}\{s_1\cdots s_\ell \in \mathcal{V}\}\Big]\Big/\ell_{\max} \geq \sigma_0\tau/\ell_{\max}. \end{aligned} \tag{6.5}$$

By partitioning $\{1,\ldots,\ell_{\max}\lfloor n/\ell_{\max}\rfloor\}$ into subsets $\{1,\ldots,\ell_{\max}\}$, $\{\ell_{\max}+1,\ldots,2\ell_{\max}\},\ldots$, we obtain the inequalities

$$p \geq \lfloor n/\ell_{\max}\rfloor \sup_{p_1\leq p_0} p_1(1-p_1)^{\lfloor n/\ell_{\max}\rfloor} \geq \sup_{p_1\leq p_0} (p_1\lfloor n/\ell_{\max}\rfloor)(1-p_1\lfloor n/\ell_{\max}\rfloor)$$

and (6.4) follows from (6.5). □

Theorem 1. *Let*

$$\tau \to 0,\ \tau \geq \beta_0^{-1+o(1)} \text{ and } \log n = o(\log\beta_0) \text{ as } p \to 0. \tag{6.6}$$

Then Algorithm A *satisfies the log efficient criterion* (6.2).

Proof. By (2.4) and (6.6),

$$\sqrt{E_Q Y_1^2} \leq \sup_{\mathbf{s}} L^{-1}(\mathbf{s}) \leq n\beta_0^{-1} = n\tau^{1+o(1)}. \tag{6.7}$$

By (6.6), $\tau \to 0$ and $\ell_{\max} \leq n = \tau^{-o(1)}$ as $p \to 0$ and (6.2) follows from (6.7) and Lemma 1. □

Remark 1. Let $\beta_1 = \sup_{\mathbf{v}\in\mathcal{V}} \beta(\mathbf{v})$. If $\mathcal{V} = \mathcal{W}$, then by (2.3), $\beta_0\tau \leq 1 \leq \beta_1\tau$ so that $\beta_1^{-1} \leq \tau \leq \beta_0^{-1}$. The conditions $\tau \to 0$ and $\tau = \beta_0^{-1+o(1)}$ then holds if (β_1/β_0) is uniformly bounded and $\beta_0 \to \infty$ as $p \to 0$.

Example 5. In both Sections 3 (palindromes and inverted repeats) and 5.1 (structural motifs), we indeed selected $\mathcal{W} = \mathcal{V}$. For the simulation of palindromes using (3.3),

$$\beta_0 \geq C_0/[4\lambda^{m-1}] \text{ and } \beta_1 \leq C_0^{-1}/[4\sigma_0\lambda^{m-1}], \text{ where } C_0 = \inf_{x\in\mathcal{X}} r(x)/\sup_{x\in\mathcal{X}} r(x) \tag{6.8}$$

and $0 < \lambda < 1$. By Remark 1, (6.6) is satisfied if $\log n = o(m)$. The ratio (β_1/β_0) has a similar form in the simulation of inverted repeats using (3.5) and (6.6) is also satisfied when $\log n = o(m)$.

For structural motifs, see (5.2),

(6.9)

$$\beta_0 \geq \sigma_0 5[3(d_2-d_1+1)(\ell(\mathbf{xy})+1)\sigma(\mathbf{y})\sigma(\mathbf{x})]^{-1} \text{ and } \beta_1 \leq [(d_2-d_1+1)(\ell(\mathbf{xy})+1)\sigma_0 5\sigma(\mathbf{y})\sigma(\mathbf{x})]^{-1}.$$

Again by Remark 1, (6.6) is satisfied when $\log n = o(\ell(\mathbf{xy}))$.

Example 6. For simulation of high-scoring motifs with respect to PSWMs via (4.1), the condition $\mathcal{V} = \mathcal{W}$ is no longer satisfied and (6.6) is verified directly. Let us consider the class of all weight matrices W whose entries are uniformly positive and bounded above (i.e. $a \leq w_i(j) \leq b$ for all i, j for some $0 < a < b < \infty$) and with $E_0 H(\mathbf{v}) + \delta_0\ell < t < \sum_{i=1}^{\ell} \max_{1\leq j\leq\mathcal{X}}[w_i(j) - \delta_0]$ for some $\delta_0 > 0$. Let θ^* be the solution to (4.4) and $\widetilde{\theta}$ the corresponding solution when t is replaced by for $t + \delta\ell$. Then $\sup_{W,t} |\theta^* - \widetilde{\theta}| \to 0$ as $\delta \to 0$.

Let $\lambda(\theta) = \log \Lambda(\theta)$. It follows from martingale theory that $\mathrm{Var}_{\tilde{\theta}}(H(\mathbf{v})) \leq C\ell$ for some uniform constant $C > 0$ and hence by Chebyshev's inequality,

$$\begin{aligned}
\tau &= \sum_{s_1\in\mathcal{X}} P_{s_1}\{H(\mathbf{v}) > t\} \geq \sum_{s_1\in\mathcal{X}} P_{s_1}\{t < H(\mathbf{v}) < t + 2\delta\ell\} \\
&\geq e^{-\tilde{\theta}(t+2\delta\ell)} \sum_{\mathbf{v}\in\mathcal{W}} [e^{\tilde{\theta}H(\mathbf{v})}\sigma(\mathbf{v})\mathbf{1}_{\{t<H(\mathbf{v})<t+2\delta\ell\}}] \\
&\geq e^{-\tilde{\theta}(t+2\delta\ell)}\Lambda(\widetilde{\theta})[1 - 2(\delta\ell)^{-2}\mathrm{Var}_{\tilde{\theta}}(H(\mathbf{v}))] \sim e^{-\tilde{\theta}(t+2\delta\ell)}\Lambda(\widetilde{\theta})
\end{aligned}$$

as $\ell \to \infty$. Since $\beta_0 \geq e^{\theta^* t}/\Lambda(\theta^*)$, the condition $\tau \geq \beta_0^{-1+o(1)}$ can be seen to hold by letting $\delta \to 0$. Moreover by large deviations theory, $|\log \tau|$ is of order ℓ and hence (6.6) holds when $\log n = o(\ell)$. Similar lines of reasoning will show that for pairwise co-occurrences, the conditions of Theorem 1 are satisfied when $\log n = o(\ell_1 + \ell_2)$.

In Algorithm B, we perform importance sampling for $c > 1$. We will show in Theorem 2 that the conditions required for log efficiency of Algorithm A will also ensure log efficiency of Algorithm B.

Theorem 2. *Let $c > 1$. Under* (6.6), Algorithm B *satisfies the log optimal criterion* (6.2) *if $\rho \sim c_1 n^{-1}$ for some $c_1 > 0$ as $p \to 0$.*

PROOF. Partition $\{1,\ldots,\ell_{\max}\lfloor n/\ell_{\max}\rfloor\}$ into subsets $\{1,\ldots,\ell_{\max}\}$, $\{\ell_{\max}+1,\ldots,2\ell_{\max}\}$, $\ldots$. Define p_0 and γ as in Lemma 1 and its proof. Then it follows from the arguments in the proof of Lemma 1 that

$$\begin{aligned} p &\geq \binom{\lfloor n/\ell_{\max}\rfloor}{c} \sup_{p_1\leq p_0} p_1^c(1-p_1)^{\lfloor n/\ell_{\max}\rfloor} \\ &\geq (c!)^{-1}\Big(1-\frac{c}{\lfloor n/\ell_{\max}\rfloor}\Big)^c \sup_{p_1\leq p_0} (p_1\lfloor n/\ell_{\max}\rfloor)^c(1-p_1\lfloor n/\ell_{\max}\rfloor) \\ &\geq \begin{cases} (c!)^{-1}(1-c/\lfloor n/\ell_{\max}\rfloor)^c/2^{c+1} & \text{if } \gamma>1/2, \\ (c!)^{-1}(1-c/\lfloor n/\ell_{\max}\rfloor)^c\gamma^c/2 & \text{if } \gamma\leq 1/2. \end{cases} \end{aligned} \tag{6.10}$$

By (2.6),

$$\sqrt{E_Q Y_1^2} \leq \sup_{\mathbf{s}} L^{-1}(\mathbf{s}) \leq [(1-\rho)^{n+\ell_{\max}}(\rho\beta_0)^c]^{-1}$$

and (6.2) follows from (6.10) and (6.6). □

7 Concluding Remarks

The importance sampling algorithms in Section 2 are stated in general terms to accomodate an arbitrary set of word patterns. We have illustrated these algorithms on three types of patterns: palindromes and inverted repeats, position specific weight matrices, and co-occurring motif pairs. It is straightforward to extend the algorithms to some word patterns not covered in the examples. For example, in Section 5.2, if the order of appearance of the two motifs is arbitrary, one can simply add a step for sampling the order, and include the necessary normalizing term in the likelihood ratio. In Example 3, we have also shown how to modify the algorithms to compute the p-value for the maximum count over a set of word patterns.

As we gain biological understanding, the models we formulate for DNA and protein functional sites gain in complexity. Over the years, they have evolved from deterministic words to consensus sequences to PSWMs to motif modules. As probabilistic models for promoter architecture become more complex and context specific, importance sampling methods are a favorable alternative to direct Monte Carlo in the absence of accurate analytic formulas.

8 Acknowledgements

This research was partially supported by Grant C-389-000-010-101 at the National University of Singapore.

References

[1] Chan, H.P. and Zhang, N.R. (2007) Scan statistics with weighted observations. *Jour. Amer. Statist. Assoc.*, 102:595-602.

[2] Chiang, D.Y., Moses, A.M., Kellis, M., Lander, E. and Eisen, M. (2003) Phylogenetically and spatially conserved word pairs associated with gene-expression changes in yeasts. *Genome Biology*, 4:R43.

[3] Cottrell, M., Fort, J.C. and Malgouyres, G. (1983) Large deviations and rare events in the study of stochastic algorithms. *IEEE Trans. Automat. Contr.*, 28:907-920.

[4] Do, K.A. and Hall, P. (1992) Distribution estimating using concomitant of order statistics, with applications to Monte Carlo simulation for the bootstrap, *JRSS 'B'*, 54:595-607.

[5] Dupuis, P. and Wang, H. (2005) Dynamic importance sampling for uniformly recurrent Markov chains. *Ann. Appl. Probab.*, 15:1-38.

[6] Fuh, C.D. and Hu, I. (2004) Efficient importance sampling for events of moderate deviations with applications, *Biometrika*, 91:471-490.

[7] Gusfield, D. (1997) *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*, Cambridge University Press, London.

[8] Huang, H., Kao, M., Zhou, X., Liu, J., and Wong, W. (2004) Determination of local statistical significance of patterns in Markov sequences with applications to promoter element identification. *J. Comput. Biology*, 11:1-14.

[9] John, M.V. (1988) Importance sampling for bootstrap confidence intervals, *Jour. Amer. Statist. Assoc.*, 83:709-714.

[10] Lai, T.L. and Shan, J.Z. (1999) Efficient recursive algorithms for detection of abrupt changes insignals and control systems. *IEEE Trans. Automat. Contr.*, 44:952-966.

[11] Leung M.Y., Choi K.P., Xia A. and Chen, L.H.Y. (2005) Nonrandom clusters of palindromes in herpesvirus genomes. *J. Comput. Biology*, 12:331-354.

[12] Leung M.Y., Schachtel G.A. and Yu H.S.(1994) Scan statistics and DNA sequence analysis: The search for an origin of replication in a virus. *Nonlinear World*.

[13] Masse, M.J.O., Karlin, S., Schachtel, G.A. and Mocarski, E.S. (1992) Human cytomegalovirus origin of DNA replication (oriLyt) resides within a highly complex repetitive region. *Proceedings of the National Academy of Sciences* 89:5246-5250.

[14] Mitrophanov, A.Y. and Borodovsky, M. (2006) Statistical significance in biological sequence analysis. *Briefings in Bioinformatics*, 7:2-24.

[15] Robin, S., Daudin, J., Richard, H., Sagot, M. and Schbath, S. (2002) Occurrence probability of structured motifs in random sequences. *J. Comput. Biology*, 9:761-773.

[16] Sadowsky, J.S. and Bucklew, J.A. (1990) On large deviations theory and asymptotically efficient Monte Carlo estimation, *IEEE Trans. Info. Theory* 36:579-588.

[17] Siegmund, D. (1976) Importance sampling in the Monte Carlo study of sequential test. *Ann. Statist.*, 4:673-684.

[18] Spellman P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B., (1998). Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Molecular Biology of the Cell* 9:3273-3297.

[19] Zhang, N.R., Wildermuth, M.C. and Speed, T.P. (2007) Transcription Factor Binding Site Prediction with Multivariate Gene Expression Data, *Submitted*.

[20] Zhou, Q. and Wong, W. (2004) CisModule: De novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proceedings of the National Academy of Sciences* 101:12114-112119.

[21] Zhu, J. and Zhang, M.Q. (1999). SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics*, Vol 15, 607-611.

W	Method	$\hat{p}$	RSE
W_{rep}	Direct	$(1 \pm 1) \times 10^{-3}$	1
	IS	$(3.4 \pm 0.3) \times 10^{-3}$	0.09
	Analytic	7.2×10^{-3}	–
W_{norep}	Direct	$4.1 \pm 1.9 \times 10^{-3}$	0.50
	IS	$7 \pm 0.5 \times 10^{-3}$	0.08
	Analytic	7×10^{-3}	–

Table 1: Comparison of direct Monte carlo, importance sampling, and analytic estimates on repetitive and non-repetitive PSWMs in Example 1. For both direct Monte Carlo and importance sampling, 1000 simulation runs were used and the results are displayed in the form estimate±standard error.

c	Method	$\hat{p}$	RSE
1	Direct	$(3.1 \pm 0.5) \times 10^{-2}$	0.20
1	IS	$(2.4 \pm 0.2) \times 10^{-2}$	0.07
1	Analytic	2.3×10^{-2}	–
2	Direct	$(4.0 \pm 2.8) \times 10^{-4}$	0.70
2	IS	$(1.9 \pm 0.1) \times 10^{-4}$	0.07
2	Analytic	2.6×10^{-4}	–

Table 2: Comparison of direct Monte carlo, importance sampling, and analytic estimates for SWI5 in Example 2. For both direct Monte Carlo and importance sampling, 1000 simulation runs were used for $c = 1$ and 5000 simulation runs were used for $c = 2$, and the results are displayed in the form estimate±standard error.

d_1	d_2	$\mathbf{x}$	$\mathbf{y}$	Direct	IS	Analytic
16	18	*gttgaca*	*atataat*	$(2 \pm 1) \times 10^{-4}$	$(1.038 \pm 0.006) \times 10^{-4}$	1.01×10^{-4}
		gttgaca	*tataata*	0	$(9.00 \pm 0.05) \times 10^{-5}$	8.82×10^{-5}
		tgttgac	*tataata*	$(20 \pm 10) \times 10^{-5}$	$(9.39 \pm 0.05) \times 10^{-5}$	9.20×10^{-5}
		ttgaca	*ttataat*	$(9 \pm 3) \times 10^{-4}$	$(6.65 \pm 0.03) \times 10^{-4}$	6.55×10^{-4}
		ttgacaa	*tacaat*	$(4 \pm 2) \times 10^{-4}$	$(4.64 \pm 0.02) \times 10^{-4}$	4.57×10^{-4}
		ttgacaa	*tataata*	$(2 \pm 1) \times 10^{-4}$	$(1.798 \pm 0.009) \times 10^{-4}$	1.78×10^{-4}
		ttgacag	*tataat*	$(5 \pm 2) \times 10^{-4}$	$(3.62 \pm 0.02) \times 10^{-4}$	3.59×10^{-4}
		ttgacg	*tataat*	$(10 \times 3) \times 10^{-4}$	$(9.90 \pm 0.06) \times 10^{-4}$	9.76×10^{-4}
		combined	p-value	$(2.0 \pm 0.4) \times 10^{-3}$	$(2.96 \pm 0.03) \times 10^{-3}$	
10	20	*gttgaca*	*atataat*	$(3 \pm 2) \times 10^{-4}$	$(3.89 \pm 0.02) \times 10^{-4}$	
		gttgaca	*tataata*	$(2 \pm 1) \times 10^{-4}$	$(3.36 \pm 0.02) \times 10^{-4}$	
		tgttgac	*tataata*	$(3 \pm 2) \times 10^{-4}$	$(3.41 \pm 0.02) \times 10^{-4}$	
		ttgaca	*ttataat*	$(2.6 \pm 0.5) \times 10^{-3}$	$(2.48 \pm 0.01) \times 10^{-3}$	
		ttgacaa	*tacaat*	$(1.5 \pm 0.4) \times 10^{-3}$	$(1.73 \pm 0.01) \times 10^{-3}$	
		ttgacaa	*tataata*	$(2 \pm 1) \times 10^{-4}$	$(6.71 \pm 0.03) \times 10^{-4}$	
		ttgacag	*tataat*	$(0.9 \pm 0.3) \times 10^{-3}$	$(1.331 \pm 0.007) \times 10^{-3}$	
		ttgacg	*tataat*	$(3.9 \pm 0.6) \times 10^{-3}$	$(3.60 \pm 0.02) \times 10^{-3}$	
		combined	p-value	$(0.80 \pm 0.09) \times 10^{-2}$	$(1.06 \pm 0.01) \times 10^{-2}$	
5	50	*gttgaca*	*atataat*	$(1 \pm 0.3) \times 10^{-3}$	$(1.265 \pm 0.008) \times 10^{-3}$	
		gttgaca	*tataata*	$(0.4 \pm 0.2) \times 10^{-3}$	$(1.103 \pm 0.007) \times 10^{-3}$	
		tgttgac	*tataata*	$(1.8 \pm 0.4) \times 10^{-3}$	$(1.150 \pm 0.007) \times 10^{-3}$	
		ttgaca	*ttataat*	$(7.4 \pm 0.9) \times 10^{-3}$	$(7.88 \pm 0.05) \times 10^{-3}$	
		ttgacaa	*tacaat*	$(5.0 \pm 0.7) \times 10^{-3}$	$(5.50 \pm 0.04) \times 10^{-3}$	
		ttgacaa	*tataata*	$(1.5 \pm 0.4) \times 10^{-3}$	$(2.21 \pm 0.01) \times 10^{-3}$	
		ttgacag	*tataat*	$(3.1 \pm 0.6) \times 10^{-3}$	$(4.23 \pm 0.03) \times 10^{-3}$	
		ttgacg	*tataat*	$(0.9 \pm 0.1) \times 10^{-2}$	$(1.126 \pm 0.008) \times 10^{-2}$	
		combined	p-value	$(2.7 \pm 0.2) \times 10^{-2}$	$(3.30 \pm 0.04) \times 10^{-2}$	

Table 3: Comparison of direct Monte Carlo, importance sampling and analytical estimates for sets of structured motifs in Example 3. For both direct Monte Carlo and importance sampling, 10,000 simulation runs were used and the results are displayed in the form estimate±standard error.