JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 17, Number 4, 2010

Pp. 581–592
DOI: 10.1089/cmb.2009.0218

Research Articles

The Power of Detecting Enriched Patterns: An HMM Approach

*ZHIYUAN ZHAI,[1] *SHIH-YEN KU,[2] YIHUI LUAN,[1] GESINE REINERT,[3] MICHAEL S. WATERMAN,[2,4] and FENGZHU SUN[2,4]

ABSTRACT

The identification of binding sites of transcription factors (TF) and other regulatory regions, referred to as motifs, located in a set of molecular sequences is of fundamental importance in genomic research. Many computational and experimental approaches have been developed to locate motifs. The set of sequences of interest can be concatenated to form a long sequence of length n. One of the successful approaches for motif discovery is to identify statistically over- or under-represented patterns in this long sequence. A pattern refers to a fixed word W over the alphabet. In the example of interest, W is a word in the set of patterns of the motif. Despite extensive studies on motif discovery, no studies have been carried out on the power of detecting statistically over- or under-represented patterns Here we address the issue of how the known presence of random instances of a known motif affects the power of detecting patterns, such as patterns within the motif. Let $N_W(n)$ be the number of possibly overlapping occurrences of a pattern W in the sequence that contains instances of a known motif; such a sequence is modeled here by a Hidden Markov Model (HMM). First, efficient computational methods for calculating the mean and variance of $N_W(n)$ are developed. Second, efficient computational methods for calculating parameters involved in the normal approximation of $N_W(n)$ for frequent patterns and compound Poisson approximation of $N_W(n)$ for rare patterns are developed. Third, an easy to use web program is developed to calculate the power of detecting patterns and the program is used to study the power of detection in several interesting biological examples.

Key words: Hidden Markov model, motif, pattern recognition, statistical power.

1. INTRODUCTION

THE IDENTIFICATION OF BINDING SITES of transcription factors (TF) and other regulatory regions, referred to as motifs, of a set of molecular sequences is of fundamental importance in genomic research. A position weight matrix (PWM) that describes the alphabet distribution in each position of the motif is

[1]School of Mathematics, Shandong University, Jinan, Shandong, P.R. China.
[2]Molecular and Computational Biology Program, University of Southern California, Los Angeles, California.
[3]Department of Statistics, Oxford University, Oxford, United Kingdom.
[4]TNLIST/Department of Automation, Tsinghua University, Beijing, P.R. China.
*These two authors contribute equally to the article.

generally used to describe a motif. Many computational and experimental approaches have been developed to identify motifs. For example, Tompa et al. (2005) studied 13 motif finding algorithms, and brief descriptions of these algorithms were given. In addition to the algorithms studied in Tompa et al. (2005), many other motif finding algorithms are available (Bailey and Elkan, 1994, 1995; Lawrence et al., 1993; Liu, 1994; Liu et al., 2002).

One class of motif finding algorithms, for example, Weeder (Pavesi et al. 2004), is based on the idea of identifying over-/under-represented patterns. In this study, we clearly distinguish motifs from patterns. Motifs are described by PWM and are randomly inserted in the sequence. A pattern is defined as any specific word over the alphabet. Motifs and patterns each may also have different lengths. The comparative study of Tompa et al. (2005) indicated that the identification of over- and/or under-represented patterns in molecular sequences continues to play a key role in identifying motifs. The set of sequences of interest can be concatenated to form a long sequence of length n. For a pre-defined candidate pattern W (whose sequence length maybe different from the motif length), the number of occurrences $N_W(n)$ of W along the sequence and the corresponding p-value are calculated; in this article, we allow occurrences to overlap. The p-value is the probability that there are at least (or at most) $N_W(n)$ occurrences of W along random sequences of the same length for testing the hypothesis that the pattern is over- (or under-) represented. The random sequences should have similar statistical characteristics as the original sets of sequences.

Extensive studies have been carried out to estimate the p-value. One approach is to approximate the distribution of $N_W(n)$ using a normal approximation for short to moderate length patterns and Poisson or compound Poisson approximations for rare patterns (Godbole, 1991; Karlin et al., 1992; Nuel, 2004; Regnier, 2000; Reinert and Schbath, 1998; Reinert et al., 2000, 2005; Robin and Daudin, 1999; Vergne and Abadi, 2008). Robin and Schbath (2001) studied the accuracy and computation time of various approximations using simulations and suggested guidelines for choosing the different approximations. Huang (2002) gave an upper bound for the normal approximation of $N_W(n)$; Fu and Lou (2007) also presented an upper bound for the normal approximation of the distribution of the renewal count vector of word counts. Although these studies provided appropriate tools to approximate the p-values, these approximations are usually not accurate enough to rank the patterns. To overcome this problem, several computationally intensive methods have been developed to calculate the exact p-value for patterns using various computational methods (Nuel, 2006a,b, 2007; Ribeca and Raineri, 2008; Zhang et al., 2007). Boeva et al. (2007) developed a method to calculate the exact p-values for a group of patterns referred as cis-regulation modules (CRM). These developments made the p-value calculation of patterns feasible for relatively long sequences.

Despite the extensive studies of p-value calculation for the number of occurrences of one and/or multiple patterns under the independent identically distributed (i.i.d) or Markov models for the sequences, only simulation approaches have been used to evaluate the power of motif detecting methods. No systematic theoretical formulas are available for the power of detecting over-represented patterns when the sequence contain multiple incidences of motifs. The power of a test statistic is the probability that a pattern is declared as significant under the alternative hypothesis that random instances of a given motif are present in the sequences of interest. Note that we think of W being in the set of patterns generated by the PWM, most likely the dominant patterns. The power of a test statistic depends on several factors: (1) the distribution of the background sequence, referred to as the background model, (2) the PWM of the motif which describes the distribution of the alphabet at the different positions, referred to as the foreground model, (3) the density $1-\lambda$ of the motif along the sequence of interest, and (4) the length n of the sequence region of interest. We will study how the power depends on these four factors.

The study of power in detecting patterns has several important implications. We assume that the background and foreground models are given. First, for given λ and n, we can calculate the power of detecting a pattern. If the power is low, we should either increase the sequence length or enrich the density of the motif. Second, for a given sequence length n, we can estimate the minimum motif density that is needed to have a certain power of detection. Third, for given motif density $1-\lambda$, we can estimate the sequence length n needed to achieve a given power $1-\beta$.

The major new contributions in the paper are the following. First, a hidden Markov model (HMM) is developed to model sequences with random instances of a motif and the HMM is used to study the power of identifying enriched patterns. Although the HMM was implicitly used for modeling motif occurrences along genomic sequences previously (Liu, 1994), it has not been used in the evaluation of statistical significance of pattern occurrences. Second, for frequent patterns, efficient computational formulas and algorithms for calculating the mean and variance of $N_W(n)$ are derived, making the normal approximation of $N_W(n)$ feasible.

Third, for rare patterns, we also derive easy to calculate formulas and computational algorithms for calculating the parameters in a compound Poisson approximation for $N_W(n)$. Although previous studies have developed computational methods for calculating the first and second order moments, as well as the exact distribution of $N_W(n)$ (Kleffe and Langbecker, 1990; Kleffe and Borodovsky, 1992; Shan and Zheng, 2009) for Markov sequences, they are not practical when the order of the Markov chain is higher than 3 (Nuel, 2006b). It is well known that for a sequence $L_i, i = 1, 2, \cdots$ generated by a HMM model with hidden process $X_i, i = 1, 2, \cdots, (X_i, L_i), i = 1, 2, \cdots$ form a Markov process. However, it is computationally expensive and sometimes impossible to calculate the probabilities of events related to $L_1, L_2, \cdots$ directly using existing formulas related to Markov sequences. We take advantage of the special features of the HMM model we develop in this paper and design efficient algorithms to calculate quantities of interest related to $N_W(n)$. Fourth, the methods developed in this article can be used to study the distribution of any patterns when random instances of motifs are present. Fifth, we develop easy to use software to calculate the power of identifying enriched patterns and sample size needed to achieve a given power.

The organization of the article is as follows. In Section 2, we present our formulation of the problem, a HMM to model the sequences when instances of a motif are known to be present, as well as methods for calculating the mean and variance of the number of occurrences of any patterns. We also present efficient computational methods for calculating the parameters involved in a normal approximation of $N_W(n)$ for frequent patterns and for a compound Poisson approximation of $N_W(n)$ for rare patterns. In Section 3, we provide easy to use software package for calculating the power of detecting patterns when instances of a motif are present, simulation results to validate the approximations, and applications to the identification of several interesting biological patterns. The details of the simulation studies are given in the Supplementary material, and the proofs of the propositions are given in the Supplementary Material (see online Supplementary Material at www.liebertonline.com).

2. PROBLEM FORMULATION, NORMAL APPROXIMATION FOR FREQUENT PATTERNS, AND COMPOUND POISSON APPROXIMATION FOR RARE PATTERNS

2.1. Problem formulation

We model the sequence data using three components: the background model, the foreground model for the motif, and the distribution of motifs along the sequence of interest.

First, we model the background sequence $B_1B_2 \cdots B_n \cdots$ as independent identically distributed (IID) random variables taking L different states $(0, 1, \cdots, L-1)$ and $p_l = P(B_1 = l)$. If we are only interested in purine or pyrimidine in each position, we can let $L = 2$. For nucleotide sequences with state space (A, C, G, T), $L = 4$, and for amino acid sequences, $L = 20$.

Second, suppose that the motif is of length K. We use the product multinomial model to describe the motif as in Liu (1994). More specifically, at the k-th position of the motif, denote by $p_l^{(k)}, k = 1, 2, \cdots, K$, the probability that the base takes value l. We also assume that the motif positions are independent. The matrix $(p_l^{(k)})_{L\times K}$ is usually called the position weight matrix (PWM).

Third, we model the distribution of motifs as follows. With probability λ, a base with the background distribution is generated. With probability $1 - \lambda$, an instance of the motif of length K is inserted and the K bases are generated based on the PWM for the motif. Once an instance of the motif is generated, we move to the end of the instance of the motif to repeat this process again. This is done for convenience of the analysis. Note that the model described above was used to describe the sequences in Reinert et al. (2009).

Let W be any pattern whose length may be different from the length of the motif. Let $Y_1Y_2 \cdots$ be the sequence and $N_W(n)$ be the number of occurrences of pattern W within $Y_1Y_2 \ldots Y_n$. Our objective is to study the distribution of $N_W(n)$ for any pattern W as well as the joint distribution of $(N_W(n), W \in \mathcal{S})$, where $\mathcal{S}$ indicates any set of patterns.

2.2. Normal approximation for the numbers of occurrences of frequent patterns using hidden Markov models (HMM)

To study the limit distribution of $(N_W(n), W \in \mathcal{S})$, we reformulate the sequence using a HMM. We denote the underlying Markov chain (MC) by $Q_1Q_2 \cdots Q_n \cdots$ and the MC is described as follows. The

states of the MC are $\{F_0 = B, F_1, F_2, \cdots F_K\}$, where B indicates the background and F_k $(1 \le k \le K)$ indicates the k-th position of the motif foreground. Conditional on $Q_n = F_0$, $Q_{n+1} = F_0$ with probability λ and $Q_{n+1} = F_1$ with probability $1 - \lambda$. Conditional on $Q_n = F_k$, $1 \le k < K$, $Q_{n+1} = F_{k+1}$ with probability 1. Conditional on $Q_n = F_K$, $Q_{n+1} = F_0$ with probability λ and $Q_{n+1} = F_1$ with probability $1 - \lambda$.

The emission probabilities are given as follows. Given $Q_n = F_k$, $0 \le k \le K$, Q_n emits $Y_n = l \in (0, 1, \cdots, L-1)$ with probability $p_l^{(k)}$. For simplicity, we denote $p_l^{(0)} = p_l$. The emission probability matrix is $E = (p_l^{(k)})_{(K+1)\times L}$. Note that E is just the PWM plus the row corresponding to the background. Let $Y_n, n = 1, 2, \cdots$ be the realized sequence of the hidden Markov model. The HMM model generates sequences with the same probability distribution as those described in the Subsection 2.1.

The following two propositions follows immediately from the basic theory of HMM (Rabiner, 1989).

Proposition 2.1. *If $0 < \lambda < 1$, then $Q_1 Q_2 \cdots Q_n \cdots$ forms a irreducible Markov process with stationary distribution $\pi = \frac{1}{\lambda + K(1-\lambda)}(\lambda, 1-\lambda, 1-\lambda, \cdots, 1-\lambda)$.*

Proposition 2.2. *For any pattern $W = W_1 W_2 \cdots W_w$ of length w, let $N_W(n)$ be the number of occurrences of W within $Y_1 \cdots Y_n$. Assume that the Markov chain $Q_1 \cdots Q_n$ starts in the stationary distribution. Then*

$$E(N_W(n)) = (n - w + 1)P(W), \tag{1}$$

where $P(W) = \sum_{k=0}^{K} \alpha_w^{(W)}(k)$ and $\alpha_i^{(W)}(k) = P(Y_1 Y_2 \cdots Y_i = W_1 W_2 \cdots W_i, Q_i = F_k), i = 1, 2, \cdots, w; k = 0, 1, \cdots, K$ can be calculated recursively using the following equations.

$$\alpha_{i+1}^{(W)}(0) = (\alpha_i^{(W)}(0) + \alpha_i^{(W)}(K))\lambda p_{W_{i+1}}, \tag{2}$$

$$\alpha_{i+1}^{(W)}(1) = (\alpha_i^{(W)}(0) + \alpha_i^{(W)}(K))(1-\lambda)p_{W_{i+1}}^{(1)}, \tag{3}$$

$$\alpha_{i+1}^{(W)}(k) = \alpha_i^{(W)}(k-1)p_{W_{i+1}}^{(k)}, \quad k = 2, 3, \cdots, K, \tag{4}$$

and

$$\alpha_1(0) = \frac{\lambda p_{W_1}}{\lambda + K(1-\lambda)}, \quad \alpha_1(k) = \frac{(1-\lambda)p_{W_1}^k}{\lambda + K(1-\lambda)}, \quad k = 1, 2, \cdots, K.$$

The following proposition shows that the covariance of the numbers of occurrences of any two patterns changes linearly with respect to the length of the sequence of interest n. Its proof is similar to that of Theorem 12.2 in Waterman (1995), except that we change the Markov model to the HMM for the sequences. Thus, the proof is omitted here.

Proposition 2.3. *Let $U = U_1 U_2 \cdots U_u$ and $V = V_1 V_2 \cdots V_v$ with lengths u and v, respectively, with $u \le v$. Let $N_U(n)$ and $N_V(n)$ be the numbers of occurrences of the corresponding patterns. Then*

$$\begin{aligned}
\lim_{n\to\infty} \mathrm{Cov}(N_U(n), N_V(n))/n &= (\beta_{UV}(0)P(V_0 U) - P(U)P(V)) \\
&+ \sum_{j=1}^{u-1} (\beta_{UV}(j)P(U_j V) - P(U)P(V)) \\
&+ \sum_{j=1}^{v-1} (\beta_{VU}(j)P(V_j U) - P(U)P(V)) \\
&+ \sum_{j=0}^{\infty}\sum_{k=0}^{K}\sum_{m=0}^{K} \alpha_u^{(U)}(m)\left(t_{mk}^{(j+1)} - \pi_k\right)P_k(V) \\
&+ \sum_{j=0}^{\infty}\sum_{k=0}^{K}\sum_{m=0}^{K} \alpha_v^{(V)}(m)\left(t_{mk}^{(j+1)} - \pi_k\right)P_k(U),
\end{aligned}$$

where $U_j V = U V_{u-j+1} \cdots V_v$ and we define the overlap bit $\beta_{U,V}(j) = 1$ if $U_{j+1} = V_1, U_{j+2} = V_2, \cdots, U_u = V_{u-j}, 0 \le j \le U - 1$ and $\beta_{U,V}(j) = 0$ otherwise. The concatenated sequence $V_j U$ and $\beta_{V,U}(j) = 0$ are similarly defined. We also abbreviate $P_k(U) = P(U_1 U_2 \cdots U_u | Q_1 = F_k)$, and $P_k(V)$ is similarly defined. Moreover, $t_{mk}^{(j+1)}$ is the $j+1$-step transition probability from m to k for the underlying Markov Chain $Q_1 Q_2 \cdots$.

We shall see that all the quantities in Proposition 3 can be calculated efficiently using the same forward procedure for HMMs as for Proposition 2.2.

The following proposition shows that $(N_W(n), W \in S)$, where S is a set of patterns, is approximately normal. This fact follows directly from Theorem 12.5 in Waterman (1995). An upper bound on the distance for the approximation is available in Huang (2002).

Proposition 2.4. *Let $P(U)$ be the quantity as shown in Proposition 2.2 with W changed to U, and $C=(C_{U,V})_{U,V\in S}$ where $C_{U,V}=\lim_{n\to\infty}\mathrm{Cov}(N_U(n),N_V(n))/n$ as given in Proposition 2.3. Assume that $0<\lambda<1$ and that C is non-degenerate. Then $(N_W(n)-nP(W))_{W\in S}/\sqrt{n}$ converges in distribution to a multinormal distribution with mean 0 and covariance matrix C.*

Proposition 2.3 shows that we need to use $d_{mk}=\sum_{j=0}^{\infty}(t_{mk}^{(j+1)}-\pi_k)$ when we calculate the covariance of the numbers of occurrences of patterns. The following proposition gives a simple way to calculate d_{mk}. In the following, for any set A, $I(A)$ is the indicator function, which equals 1 when A is true, and 0 otherwise.

Proposition 2.5. *Let $A_j=t_{00}^{(j)}$ and $S_i=\sum_{j=1}^{i}(A_j-\pi_0)$, where $\pi_0=\frac{\lambda}{\lambda+K(1-\lambda)}$. Then*

1. For $1\le i\le K+1$, S_i is given by

$$\frac{\lambda}{1-\lambda}(1-\lambda^i)-i\pi_0.$$

2. S_i follows the recursive equation,

$$S_i=\lambda S_{i-1}+(1-\lambda)S_{i-K},\quad i>K+1.$$

3. Let $\gamma=\lim_{i\to\infty}Si$ and $d_{mk}=\sum_{j=0}^{\infty}(t_{mk}^{(j+1)}-\pi_k)$, $m,k=0,1,\cdots,K$. Then for any $1\le k,\ m\le K$,

$$\begin{aligned}
d_{k0}&=\gamma-(K-k)\pi_0,\\
d_{mk}&=\frac{(1-\lambda)}{\lambda}(\gamma-(K-m+k-1)\pi_0)+I(k\ge m+1),\\
d_{0k}&=d_{Kk}.
\end{aligned}$$

From Lemma 12.1 in Waterman (1995), it is known that there exists constants $C\ge 0$, $0\le\rho<1$ such that $|A_j-\pi_0|\le C\rho^j$. Therefore, the limit γ exists and for all N,

$$|S_N-\gamma|\le\frac{C\rho^{N+1}}{(1-\rho)}.$$

Hence, γ can be approximated efficiently.

In the Supplementary Material, we present detailed proofs of the propositions and closed-form formulas for the mean and variance of the numbers of occurrences of patterns for the special case of $K=2$ (see online Supplementary Material at www.liebertonline.com).

2.3. Compound Poisson approximation for the number of occurrences of rare patterns

The normal approximation given in Subsection 2.2 is only appropriate for frequent patterns. For rare patterns, a compound Poisson approximation for $N_W(n)$ is more suitable (Reinert and Schbath, 1998; Reinert et al., 2005; Schbath, 1995, 2000). In this subsection, we provide efficient formulas for calculating the parameters used in a compound Poisson approximation for $N_W(n)$.

2.3.1. Poisson approximation for the number of clumps. The basic idea of the compound Poisson approximation is to divide the occurrences of pattern W into *clumps*, where a *clump* of pattern W is a maximum set of overlapping occurrences of W. The number of clumps can be approximated by a Poisson distribution. We introduce the following notation for a word $W=W_1W_2\cdots W_w$:

- $W^{(p)}=W_1W_2\cdots W_p$: the p-th prefix of the word W;
- $\mathcal{P}(W)=\{p\in\{1,2,\cdots,w-1\}: w_i=w_{i+p},\quad \forall i=1,2,\cdots w-p\}$: referred to as the *periods* of W;
- *Principal periods* $P'(W)$ are those periods that cannot be written as multiples of other periods.

Based on Theorem 6.6.4 of Reinert et al. (2005), we have the following proposition.

Proposition 2.6. *Under the model in Subsection 2.1, the number of clumps $NC_n(W)$ for a pattern W of length w within $Y_1Y_2\cdots Y_n$ can be approximated by a Poisson distribution with mean $\Lambda(W)=(n-w+1)\tilde{\mu}(W)$, where $\tilde{\mu}(W)$ is the probability that a clump of W occurs at a particular position (Reinert and Schbath, 1998; Reinert et al., 2005; Schbath, 1995), and*

$$\tilde{\mu}(W)=P(W)-\sum_{p\in\mathcal{P}'(W)}P(W^{(p)}W).$$

In our case, $P(W)$ and $P(W^{(p)}W)$ can be calculated using the recursive formula in Proposition 2.2.

2.3.2. The approximate distribution of the number of occurrences of pattern W in a clump. Next, we calculate the distribution of the number of occurrences of pattern W in each clump. Let $\tilde{\mu}_j(W)$ be the probability that a clump with exactly j overlapping occurrences of W starts at a pre-specified position in the sequence. It was shown in Reinert and Schbath (1998) and Schbath (1995) that

$$\tilde{\mu}_j(W)=P(C_j)-2P(C_{j+1})+P(C_{j+2}),\quad j=1,2,\cdots,$$

where

$$C_j=\{W^{(p_1)}W^{(p_2)}\cdots W^{(p_{j-1})}W,\quad p_1,p_2,\cdots,p_{j-1}\ \textit{are principal periods of}\ W\}$$

is the set of words which consist of exactly j overlapping occurrences of W. For each combination of $p_1,p_2,\cdots,p_{j-1}$ of principal periods of W, the value of $P(W^{(p_1)}W^{(p_2)}\cdots W^{(p_{j-1})}W)$ can be calculated using the recursive equations in Proposition 2.2.

2.3.3. Calculating the approximate distribution of the number of occurrences of pattern W using Panjer recursion. With the above preparations, we can calculate the approximate distribution of $N_W(n)$ for rare patterns using Panjer recursion (Panjer, 1981). From Proposition 2.6, the number of clumps $NC_n(W)$ for the pattern W can be approximated by a Poisson random variable with mean $\Lambda(W)$. The probability that a clump with exactly j overlapping occurrences of W starts at a pre-specified position in the sequence is $\tilde{\mu}_j=\tilde{\mu}_j(W)$. Note that the total number of occurrences of pattern $Y_1Y_2\cdots Y_n$ can be written as,

$$N_n(W)=\sum_{s=1}^{NC_n(W)}Z_s,$$

where Z_s is the number of occurrences of pattern W in the s-th clump. Thus, we have the following proposition for a compound Poisson approximation of $N_W(n)$ for rare patterns. An upper bound for the distance between the true distribution and the approximation was given in Reinert and Schbath (1998), Reinert et al. (2005), and Schbath (1995).

Proposition 2.7. *Under the model described in Subsection 2.1, the number of occurrences of a rare pattern W, $N_W(n)$, can be approximated by a compound Poisson random variable*

$$\tilde{N}(W)=\sum_{s=1}^{\widetilde{NC}_n(W)}\tilde{Z}_s,$$

where $\widetilde{NC}(W)$ is a Poisson random variable with mean $\Lambda(W)=(n-w+1)\tilde{\mu}(W)$ and

$$P\{Z_s=j\}=\tilde{\mu},j=1,2,\cdots.$$

The probability of $g_j=P(\tilde{N}(W)=j)$ can be calculated recursively by

$$g_j=\frac{\Lambda(W)}{j}\sum_{j=1}^{j}i\tilde{\mu}_ig_{j-i},\quad j=1,2,\cdots$$

with the initial value of $g_0=e^{-\Lambda(W)}$; see Equation (2.3) in Willmot and Panjer (1987).

3. AN ONLINE PROGRAM FOR CALCULATING POWER, NUMERICAL STUDIES, AND APPLICATIONS

With the general formulation of motif occurrences, the theoretical formulas for the mean and covariance of the number of occurrences of patterns involved in the normal approximation of $N_W(n)$ for frequent patterns, and the equations for the parameters involved in the compound Poisson approximation of rare patterns, we develop a program that can calculate the power of detecting patterns when motif instances are present. Simulation studies are carried out to study the validity of our theoretical results and identify parameter regions where the theoretical results work the best. Finally, we give some real biological examples on identifying frequent and rare patterns.

3.1. A program for calculating the power of detecting patterns when motif instances are present

Although we do not have a closed formula to calculate the power of pattern detection for the general model, based on the results developed above, we provide an online computer algorithm to approximate the power, which is implemented at http://meta.cmb.usc.edu/motif_power/. The inputs of the program are:

1. The background nucleotide or amino acid frequencies of the sequence under study $p_l, l=0, 1, \cdots L-1$;
2. The nucleotide or amino acid frequencies at each position of the motif (PWM), $p_l^{(k)}, l=0, 1, \cdots L-1, k=1, 2, \cdots, K$;
3. The length of the sequence under study, n;
4. The probability of the background, λ;
5. A given pattern, W;
6. Type I error.

For each set of parameters, the program will output the mean and variance of the number of occurrences of the pattern W. Based on these quantities, the program gives suggestions whether normal approximation or compound Poisson approximation is more appropriate for calculating the power of detection. In our program, we suggest using normal approximation when the mean number of occurrences is greater than 500 and using compound Poisson approximation otherwise. This suggestion is the same as in Robin and Schbath (2001). The program outputs the power of detecting the pattern W as over-represented. In addition, we also provide an option of calculating the sequence length needed to achieve a pre-specified power. In this case, the users need to input the pre-specified power.

3.2. Simulation studies

We carry out simulations to evaluate the validity of the theoretical results and our program. First, we consider the situation that there are $L=2$ states, 0 or 1, at each position. Since the probability of motifs, $1-\lambda$, is generally small, we choose $\lambda=0.9, 1.0$ in our simulations. While the case $\lambda=1$ is excluded in Propositions 1 and 2, analogous results to Propositions 3, 6 and 7 are well-known for this simple case that the sequence is composed of independent identically distributed letters (Reinert et al., 2005). We let the background probability of choosing 1, $p=P(B_1=1)$, to be 0.1, 0.5, and 0.7, respectively, to see the effect of p on our theoretical results. The inserted motif is "11." In the second simulation, we consider $L=4$ states at each position intended to simulate nucleotide sequences. We denote the four states to be (A, C, G, T) corresponding to the four nucleotide bases. The following three situations are considered: (1) uniform $P(l)=1/4$, $l=A, C, G, T$; (2) GC rich: $P(A)=P(T)=0.15$, $P(C)=P(G)=0.35$; and (3) GC poor: $P(A)=P(T)=0.35$, $P(C)=P(G)=0.15$. We consider relative short patterns of length $K=4$, (2a) "ACGT" with no principal periods, and (2b) "CGCG" with principal period 2. The third simulation is almost the same as the second simulation except that we consider relatively long motifs: (3a) "ACGTATC" with no principal periods and (3b) "AAGAAGAA" with principal periods {3, 7}. In all our simulations, the pattern of interest is the same as the inserted motif. The values of sequence length n and the motif intensity $1-\lambda$ are given in Supplementary Material (see online Supplementary Material at www.liebertonline.com).

We evaluate the theoretical results of Section 3 using three different criteria. The first criterion is to compare the theoretical variance of $N_W(n)$ given in Proposition 2.3 with the simulated variance. It is shown that they are close in all the simulations, indicating the general applicability of the results in Proposition 2.3.

However, we caution that theoretical upper bounds for the approximations are needed to determine the applicability of Proposition 2.3. The second criterion is to compare the normal approximation and the compound Poisson approximation with the simulated distribution of $N_W(n)$. It is shown that normal approximation is approximate in the first simulation when the pattern of interest is relatively common. Compound Poisson approximation is appropriate under the other cases of the simulations. Third, we compare the theoretical power using the corresponding approximate distributions with the simulated power. The simulation results are presented in Supplementary Material.

3.3. Biological examples

3.3.1. The power of detecting CpG enriched regions using normal approximation for $N_W(n)$. CpG dinucleotides play important roles in many biological processes including CpG methylation which can cause high mutation rate and modify chromatin structure (Nguyen et al., 2002; Takai and Jones, 2002). It has also been shown that methylation at non-promoter CpG islands (genomic regions with significantly enriched CpG dinuclotides) is associated with aging and cancer in human (Nguyen et al., 2002). Moreover, CpG islands have been shown to be associated with the origin of replication and to be frequently located at or near the transcription start sites of genes (Antequera and Bird, 1999). Therefore, the identification of CpG islands is an important problem.

Gentles and Karlin (1999) defined dinucleotide relative abundance as

$$\rho_{XY} = f_{XY}/(f_X f_Y),$$

where f_{XpY} is the dinucleotide frequency of X Y and f_X is the nucleotide frequency for X. Then XpY enriched genomic regions are defined as those satisfying $f_{XpY}/(f_X f_Y) \geq 1.23$ (Gentles and Karlin, 1999). Since our mathematical model assumes an independent identically distributed model for the background sequence and does not consider other biological mechanisms such as methylation, we only consider organisms that do not show strong evidence of methylation (average $\rho \geq 0.90$) including *C. elegans*, *D. melanogaster*, and *E. coli* (Takai and Jones, 2002). The G + C content in these organisms are 0.36, 0.43, and 0.51, respectively. In this example, both the motif and the pattern are "CG."

For each organism, we estimate the corresponding motif density $1-\lambda$ so that $P_{CG} = 1.23$ using the equation

$$1.23 f_C f_G = P_\lambda(CG),$$

where $P_\lambda(CG)$ is defined in Proposition 2.2. The resulting estimated value of the motif density $1-\lambda$, the average distance between the random instances of the inserted motif "CG," and the sequence length needed to achieve 80% power (type I error 0.025 or 0.05) based on the normal approximation for $N_{CG}(n)$ are given in Table 1. The table shows that as the G + C content increases, the motif density $1-\lambda$ needed to achieve the same relative dinucleotide abundance also increases and the sequence length needed to detect the enrichment of CpG islands decreases. Figure 1 (left) shows the power of detecting CpG islands with relative dinucleotide abundance of 1.23 versus the length of sequences for *C. elegans*, *D. melanogaster*, and *E. coli*, respectively (type I error 0.025), using the normal approximation. It can be seen that the power increases as the G + C content increases.

3.3.2. The power of detecting relative long patterns using the compound Poisson approximation. We then present three examples of detecting relative long patterns: a binding site of a transcription factor SP1, a zinc finger protein motif C2H2, and a structural motif.

A. The DNA binding site of transcription factor SP1 in human sequences: The human transcription factor SP1 binds to GC box promoter elements and then activates mRNA synthesis, for details of the biological properties of SP1, see UniProt at www.uniprot.org/uniprot/P08047. Thiesen and Bach (1990) studied the DNA binding site of SP1 and identified its PWM (Supplementary Material) (see online Supplementary Material at www.liebertonline.com). The DNA binding site is characterized as a GC rich domain (TRANSFAC 2.1.). The background frequencies for *A, C, G, T* are calculated from the human genome sequences. We consider the dominant pattern, "M = GGGGCGGGGT," of this motif.

B. The C2H2 zinc finger motif: Zinc finger proteins have several conserved sequences that could be classified into 8 fold groups: C2H2 like, Cag knuckle, Treble clef, Zinc ribbon, Zn/Cys6, TAZ2 domain like, Zinc binding loops, and Metallothionein (Krishna et al., 2003). The Human transcription factor SP1 has three

TABLE 1. G+C CONTENT, ESTIMATED (CG) MOTIF DENSITY, DISTANCE BETWEEN CONSECUTIVE MOTIFS, AND SEQUENCE LENGTHS NEEDED TO HAVE 80% POWER (TYPE I ERROR 2.5%, FOURTH COLUMN; TYPE I ERROR 5%, FIFTH COLUMN) TO DETECT THE PATTERN CpG FOR *C. ELEGANS*, *D. MELANOGASTER*, AND *E. COLI*

Organism	*G+C content*	*Motif density* $(1-\lambda)$	*Distance*	*Sequence length* $(\alpha=0.025)$	*Sequence length* $(\alpha=0.05)$
C. elegans	0.36	0.008	125	4719	3743
D. melanogaster	0.43	0.012	83	3147	2495
E. coli	0.51	0.016	63	2654	2101

contiguous Zinc finger domains, known as the "C2H2" domain. We use MEME with default parameter values (Bailey and Elkan, 1994) to find a motif for the SCOP family g.37.1.1, which is described as "Classic zinc finger, C2H2". There are 121 protein chains within the family according to version 1.71 of the SCOP database. The PWM of the motif with the smallest e-value is given in the Supplementary Material, and the dominant pattern is "M=CEICDRRFSRSDHLTRHIRIH."

C. A structural alphabet motif: Several investigators encode the protein structure into structural alphabets and design computational structural alignment tools based on the alphabet sequences. For example, PBE (Tyagi et al., 2006), SARST (Lo et al., 2007), and SAFAST (Ku and Hu, 2008) defined alphabet sets by using the dihedral angle(ψ, ϕ). 3d-BLAST (Yang and Tung, 2006) and YAKUSA (Carpentier et al., 2005) defined the alphabet sets using the other dihedral angles(κ, α). These studies showed that the structural alphabet could represent the structural information. SAFAST (Ku and Hu, 2008) also provides an approach to find the protein structural motifs using MEME (Bailey and Elkan, 1994). Here we choose the protein structural alphabets of g.37.1.1 (C2H2 Zinc finger) from SAFAST (Ku and Hu 2008). MEME is then used to find the structural motif. The resulting PWM is given in the Supplementary Material, and the dominant pattern is "M=GHNACARQ."

Figure 1 (right) gives the power of detecting the corresponding patterns as a function of motif density $1-\lambda$ for the three cases respectively (type I error 0.025). For cases A and C, $n=10^4$ and for case B, $n=10^8$. The motif density that can be detected with the corresponding lengths can be identified from the figure.

4. CONCLUSION

In this article, we study the power of detecting certain patterns when random motif instances are present in a sequence of interest. Theoretical approximate formulas for the mean and covariance of the number of occurrences of patterns are obtained and simulation studies show that these formulas approximate their corresponding true values extremely well with a relative error less than 3% (data not shown). Simulations also show that when the mean number of occurrences of the pattern relatively large (e.g., $\geq$10) and sequence

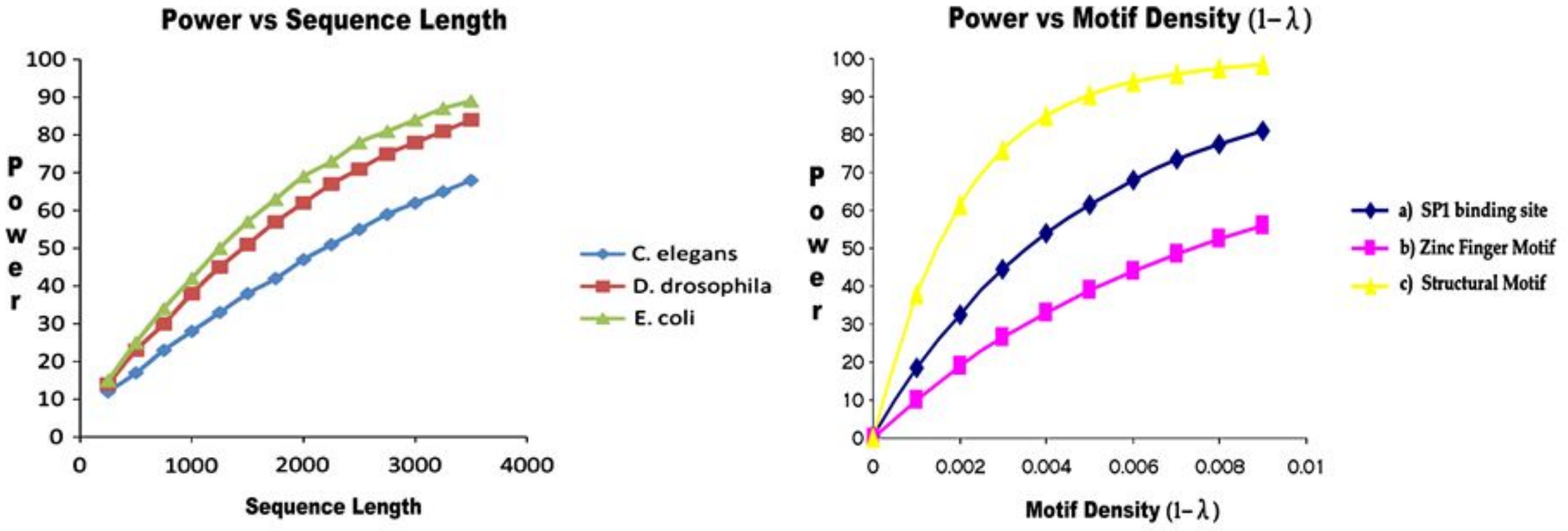

FIG. 1. (**Left**) The power of detecting CpG island with relative dinucleotide abundance of 1.23 versus the sequence length for *C. elegans*, *D. melanogaster*, and *E. coli*, using the normal approximation. (**Right**) The power of detecting (a) the SP1 binding site ($n=10^4$), (b) the zinc finger motif ($n=10^8$), and (c) the structural motif ($n=10^4$) versus the motif density $1-\lambda$ using a compound Poisson approximation.

length is also large (e.g., $\geq$1000), our normal approximation for the number of pattern occurrences works well. However, our normal approximation does not work well when the number of occurrences of the pattern is relative small as shown in previous studies. For patterns with relatively low number of occurrences (e.g., the mean occurrence is less than 10), we develop formulas and computer algorithms to calculate the parameters involved in compound Poisson approximation, and it is shown that the approximation works well. According to the results derived from this study, we develop an easy to use web-based program for calculating the power of detecting certain patterns when random motif instances are present. As far as we are aware, no such programs are currently available.

In this study, we make several simplified assumptions, including (1) the background sequences are independent and identically distributed, (2) the motifs are randomly distributed along the sequence of interest, and (3) the motif is modeled by a multiplicative multinomial model. All these assumptions can be violated in reality. A commonly used sequence background model is the Markov chain (MC). Motifs can also be modeled as MC. The framework developed in this paper can be relatively easily extended to such situations without too much difficulty by considering (Q_i, Y_i) as a MC with LK states. Let T be the transition matrix of this MC. The calculation of $T^j, j = 1, 2, \cdots$ can be computationally expensive when LK is large. How to efficiently calculate these matrixes is a problem of future research. In this article, we assume that motif instances do not overlap to simplify our analysis. However, motif instances can overlap and there are potentially multiple motifs in real sequence data. How these complications affect the power of detecting over- and under-represented patterns need to be studied in the future.

In practical applications, the motif density $1 - \lambda$ is usually unknown. One natural question is how we estimate λ. Motif finding programs sometimes give an estimated value of λ. If not, one approach is through moment estimation. We can first calculate the number of occurrences of the fixed pattern of interest in the sequences. Using the formula in Proposition 2.2, we can estimate λ by equating the theoretical expected count with the observed counts. When new sequences potentially containing the motif are available, we can calculate the power of detecting the motif using the estimated λ.

In this article, we consider overlapping counts of the patterns of interest. Another interesting quantity is non-overlapping counts of the patterns (also called renewal counts), that is, when the pattern occurs at a position, we move to the end of the pattern to restart the counting process. Thus, the counted patterns do not overlap. The statistical properties of the renewal pattern counts when motifs are present is a topic of further study.

In summary, we provide the theoretical foundations and an easy to use web-based program for calculating the power of detecting certain patterns in sequences. This program provides a useful tool for researchers to see if the available sequences suffice for detecting the patterns of interests.

ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China (grant 10671110 to Z.Y.Z. and Y.H.L.), by EPSRC (grant GR/R52183/01), by BBSRC and EPSRC through OCISB (to G.R.), by NIH P50 HG 002790 and NIH R21AG032743 (to M.S.W. and F.Z.S.), and by NSFC (grants 60928007 and 60805010 to F.Z.S.).

DISCLOSURE STATEMENT

The authors do not have conflicts of interests.

REFERENCES

Antequera, F., and Bird, A. 1999. CpG islands as genomic footprints of promoters that are associated with replication origins. *Curr. Biol.* 9, 661667.

Bailey, T.L., and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. 2nd Int. Conf. Intell. Syst. Mol. Biol.* 28–36.

Bailey, T.L., and Elkan, C. 1995. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Mach. Learn.* 21, 51–80.

Boeva, V., Clément, J., Régnier M., et al. 2007. Exact p-value calculation for heterotypic clusters of regulatory motifs and its application in computational annotation of cis-regulatory modules. *Algorithms Mol. Biol.* 2, 13.

Carpentier, M., Brouillet, S., and Pothier, J. 2005. YAKUSA: a fast structural database scanning method. *Proteins* 61, 137–151.

Ewens, W.J., and Grant, G. 2005. *Statistical Methods in Bioinformatics: An Introduction*, 2nd ed. Springer-Verlag, New York.

Fu, J.C., and Lou, W.Y.W. 2007. On the normal approximation for the distribution of the number of simple or compound patterns in a random sequence of multi-state trials. *Methodol. Comput. Appl. Probabil.* 9, 195–205.

Gentles, A.J., and Karlin, S. 2001. Genome-scale compositional comparisons in eukaryotes. *Genome Res.* 11, 540546.

Godbole, A.P. 1991. Poisson approximations for runs and patterns of rare events. *Adv. Appl. Probabil.* 23, 851–865.

Huang, H. 2002. Error bounds on multivariate normal approximations for word count statistics. *Adv. Appl. Probabil.* 34, 559–586.

Karlin, S., Burge, C., and Campbell, A.M. 1992. Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic Acids Res.* 20, 1363–1370.

Kleffe, J., and Borodovsky, M. 1992. First and second moment of count of words in random texts generated by Markov chains. *Comput. Appl. Biosci.* 8, 433–441.

Kleffe, J., and Langbecker, U. 1990. Exact computation of pattern probabilities in random sequences generated by Markov chains. *Comput. Appl. Biosci.* 6, 347–353.

Krishna, S., Majumdar, I., and Grishin, N. V. 2003. Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res.* 31, 532–550.

Ku, S.Y., and Hu, Y.J. 2008. Protein structure search and local structure characterization. *BMC Bioinform.* 9, 349.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., et al. 1993. Detecting subtle sequence signals—a Gibbs sampling strategy for multiple alignment. *Science* 262, 208–214.

Liu, J.S. 1994. The collapsed Gibbs sampler in Bayesian computations with applications to a gene-regulation problem. *J. Am. Statist. Assoc.* 89, 958–966.

Liu, X.L., Brutlag, D.L., and Liu, J.S. 2002. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.* 20, 835–839.

Lo, W.C., Chang, C.H., Huang, P.J., et al. 2007. Protein structural similarity search by Ramachandran codes. *BMC Bioinform.* 8, 307.

Nguyen, C., Liang, G.M., Nguyen, T.T., et al. 2001. Susceptibility of nonpromoter CpG islands to *de novo* methylation in normal and neoplastic cells. *J. Natl. Cancer Inst.* 93, 1465–1472.

Nuel, G. 2004. LD-SPatt: large deviations statistics for patterns on Markov chains. *J. Comp. Biol.* 11, 1023–1033.

Nuel, G. 2006a. Effective P-value computations using finite Markov chain imbedding (FMCI): application to local score and to pattern statistics. *Algorithms Mol. Biol.* 1, 5.

Nuel, G. 2006b. Numerical solutions for pattern statistics on Markov chains. *Statist. Appl. Genet. Mol. Biol.* 5, 26.

Nuel, G. 2007. Pattern Markov chains: optimal Markov chain embedding through deterministic finite automata. *J. Appl. Probabil.* 45, 226–243.

Panjer, H.H. 1981. Recursive evaluation of a family of compound distributions. *Astin Bull.* 12, 22–26.

Pavesi, G., Mereghetti, P., Mauri, G., et al. 2004. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.* 32, W199–W203.

Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications inspeech recognition. *Proc. IEEE* 77, 257–286.

Regnier, M. 2000. A unified approach to word occurrence probabilities. *Discr. Appl. Math.* 104, 259–280.

Reinert, G., and Schbath, S. 1998. Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. *J. Comput. Biol.* 5, 223–253.

Reinert, G., Schbath, S., and Waterman, M.S. 2000. Probabilistic and statistical properties of words: an overview. *J. Comput. Biol.* 7, 1–46.

Reinert, G., Schbath, S., and Waterman, M.S. 2005. Statistics on words with applications to biological sequences, 251–328. In Berstel, J., and Perrin, D., eds., *Lothaire: Applied Combinatorics on Words.* Cambridge University Press, New York.

Reinert, G., Chew, D., Sun F.Z., et al. 2009. Alignment free sequence comparison (I): statistics and power. *J. Comput. Biol.* 16, 1–20.

Ribeca, P., and Raineri, E. 2008. Faster exact Markovian probability functions for motif occurrences: a DFA-only approach. *Bioinformatics.* 24, 2839–2848.

Robin, S., and Daudin, J.J. 1999. Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Probabil.* 36, 179–193.

Robin, S., and Schbath, S. 2001. Numerical comparison of several approximations of the word count distribution in random sequences. *J. Comput. Biol.* 8, 349–359.

Royden, H.L. 1988. *Real Analysis.* Prentice Hall, Englewood Cliffs, NJ.

Schbath, S. 1995. Compound Poisson approximation of word counts in DNA sequences. *ESAIM Probabil. Statist.* 1, 1–16.

Schbath, S. 2000. An overview on the distribution of word counts in Markov chains. *J. Comput. Biol.* 7, 193–201.

Shan, G., and Zheng, W.M. 2009. Counting of oligomers in sequences generated by Markov chains for DNA motif discovery. *J. Bioinform. Comput. Biol.* 7, 39–54.

Takai, D., and Jones, P.A. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. USA* 99, 3740–3745.

Thiesen, H., and Bach, C. 1990. Target detection assay (TDA): a versatile procedure to determine DNA binding sites as demonstrated on SP1 protein. *Nucleic Acids Res.* 18, 3203–3209.

Tompa, M., Li, N., Bailey, T.L., et al. 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* 23, 137–144.

Tyagi, M., Sharma, P., Swamy, C.S., et al. 2006. Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet. *Nucleic Acids Res.* 34, W119–W123.

Vergne, N., and Abadi, M. 2008. Poisson approximation for search of rare words in DNA sequences. *Latin Am. J. Probabil. Math. Statist.* 4, 223–244.

Waterman, M.S. 1995. *Introduction to Computational Biology. Maps, Sequences and Genomes.* Chapman & Hall, New York.

Willmot, G.E., and Panjer, H.H. 1987. Difference equation approaches in evaluation of compound distributions. *Insurance Math. Econ.* 6, 43–56.

Yang, J.M., and Tung, C.H. 2006. Protein structure database search and evolutionary classification. *Nucleic Acids Res.* 34, 3646–3659.

Zhang, J., Jiang, B., Li, M., et al. 2007. Computing exact P-values for DNA motifs. *Bioinformatics* 23, 531–537.

Address correspondence to:
Dr. Fengzhu Sun
Molecular and Computational Biology Program
University of Southern California
1050 Childs Way
Los Angeles, CA 90089

E-mail: fsun@usc.edu