JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 15, Number 6, 2008

Pp. 547–564
DOI: 10.1089/cmb.2007.0084

Compound Poisson Approximation of the Number of Occurrences of a Position Frequency Matrix (PFM) on Both Strands

UTZ J. PAPE,[1,2] SVEN RAHMANN,[3] FENGZHU SUN,[4] and MARTIN VINGRON[1]

ABSTRACT

Transcription factors play a key role in gene regulation by interacting with specific binding sites or motifs. Therefore, enrichment of binding motifs is important for genome annotation and efficient computation of the statistical significance, the *p*-value, of the enrichment of motifs is crucial. We propose an efficient approximation to compute the significance. Due to the incorporation of both strands of the DNA molecules and explicit modeling of dependencies between overlapping hits, we achieve accurate results for any DNA motif based on its Position Frequency Matrix (PFM) representation. The accuracy of the *p*-value approximation is shown by comparison with the simulated count distribution. Furthermore, we compare the approach with a binomial approximation, (compound) Poisson approximation, and a normal approximation. In general, our approach outperforms these approximations or is equally good but significantly faster. An implementation of our approach is available at *http://mosta.molgen.mpg.de*.

Key words: binding site clumps, compound Poisson distribution, count statistics, DNA motif, overlapping occurrences, PFM, Position Frequency Matrix, *p*-value.

1. INTRODUCTION

TRANSCRIPTION FACTORS (TFs) play a key role in gene regulation by binding to genomic sequences (Alberts et al., 2002). Binding sites of TFs are often represented as position frequency matrices (PFMs) introduced by Stormo et al. (1982). Genome annotation and the understanding of genetic regulatory networks require the assessment of the statistical significance of TF binding site occurrences. Reliable results for the significance are given by *p*-values derived from the count statistic (Denise et al., 2001).

Exact calculation of the count distribution can be done by generating functions (Gentleman and Mullin, 1989; Hertzberg et al., 2005; Kleffe and Langbecker, 1990; Régnier, 2001) or other methods (Beckstette

[1]Computational Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany.
[2]Mathematics and Computer Science, Free University of Berlin, Berlin, Germany.
[3]COMET group, Genome Informatics, Technische Fakultät, Universität Bielefeld, Bielefeld, Germany.
[4]Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, California.

et al., 2006; Bejerano et al., 2004; Staden, 1989; Zhang et al., 2007). Unfortunately, the calculation of exact p-values is NP-hard (Zhang et al., 2007). Hence, many efforts are concerned with the asymptotic word count distribution using normal approximations (Brendel et al., 1986; Leung et al., 1996; Waterman, 2000; Prum et al., 1995), (compound) Poisson approximations (Chryssaphinou and Papastavridis, 1988; Godbole, 1991; Schbath, 1995; Robin, 2002; Roquain and Schbath, 2007), and large deviation results (Denise et al., 2001; Reinert et al., 2000). Both the exact and the asymptotic approaches require the enumeration of all compatible words which are words encoded in the PFM yielding a hit on the sequence. The number of compatible words grows exponentially with the length of the PFM. Thus, calculation is inefficient for long PFMs.

In this article, we propose an approximation based on the compound Poisson distribution for the number of occurrences of a PFM without enumerating all compatible words. Furthermore, we incorporate both strands of the DNA molecules, which is important for palindromic PFMs. We explicitly consider dependencies of overlapping hits. The approach outperforms existing approximations (Schbath, 1995; Waterman, 2000; Roquain and Schbath, 2007). In contrast to the most recent exact calculation (Zhang et al., 2007), its complexity neither depends on the number of compatible words nor on the sequence length.

The next section introduces the statistical framework (Rahmann et al., 2003) and develops the approximation for the count statistic. Since we explicitly model the self-overlap of the PFM that results in dependencies, we can calculate two characteristic values for the self-overlap and the palindromicity of a PFM. Finally, the simulation for the comparison of the approaches is described, as well as the competing approaches (Schbath, 1995; Waterman, 2000; Roquain and Schbath, 2007). The Results section contains the comparison of the approaches. A discussion and an outlook are given in the final section.

2. METHODS

Our statistic for the number of binding sites uses the probability of a detected binding site, as well, as the self-overlap of the PFM. For example, a PFM with the consensus "CTAACT" has a higher probability to find two hits overlapping in two positions than to find two independent hits. Counting the number of binding sites while taking care of the self-overlap has already been discussed for a single word (Guibas and Odlyzko, 1981; Robin and Schbath, 2001) and a small set of given words (Reinert et al., 2000). We give two reasons for avoiding the Chen-Stein approach proposed by Reinert et al. (2000) and Roquain and Schbath (2007): First of all, the enumeration of all compatible words encoded in the PFM is computationally demanding and only possible for small PFMs. Second, the incorporation of the complementary strand can lead to two hits at one position. In terms of word counting, this means that the words in the set of compatible words are not necessarily different contradicting one important assumption for the (compound) Poisson approximation. Therefore, we use the discrete nature of the PFM score to compute the probabilities of overlapping hits (Pape et al., 2006). Based on these probabilities, we use a generalization of the Poisson distribution to model overlaps. We couple a probability vector for the number of hits with a Poisson distribution. This is a so called stopped-sum distribution (Johnson et al., 1995) or compound Poisson distribution. This distribution is widely used for word count statistics.

2.1. The PFM framework

Each PFM represents a binding site. It contains specific probabilities for each nucleotide at every position. We assume that the binding sites of each TF are described by only one PFM. An extension to more than one PFM is not trivial but, in general, possible. The position specific scoring matrix (PSSM) $\Psi_{\kappa,\sigma}$ is chosen to be the log-likelihood ratios of the nucleotide distribution of the PFM and the background probabilities π_σ for every position κ and for nucleotides $\sigma \in \Sigma$. We denote the length of the PSSM by ℓ. The background model is an i.i.d. model only defined by the GC content. Since we require the distribution of hits on both strands to be equal, we need this restriction. We call this a symmetric i.i.d. background model which can also be justified by Chargaff's second law. Furthermore, in contrast to coding sequence, there is no motivation to handle both strands in the upstream region differently.

Using the PSSM, we can assign a score to every position of the potential binding site depending on the observed nucleotide. Sliding a window of length ℓ over the sequence $\sigma_0 \dots \sigma_n$ and summing up the scores

FIG. 1. The figure clarifies notation: The lower 5′-3′ strand is the leading strand. Given a motif CTAT, there are two overlapping occurrences on the shown sequence region. $Y_1 = Y'_3 = 1$ indicate a hit starting at position 1 on the leading strand and another hit ending at position 3 on the complementary strand. This definition of Y_j and Y'_j simplifies notation.

in each window, yields a score S_j for every position j of the sequence

$$S_j = \sum_{\kappa=0}^{\ell-1} \Psi_{\kappa,\sigma_{j+\kappa}}.$$

From now on, we consider the letters σ_j of the sequence as random variables. Thus, S_js are random variables, too.

Considering the complementary strand requires additional notation: In general, we use the same variables with a prime for this purpose. We call the strand of the corresponding gene the 5′-3′ strand. Correspondingly, the complementary strand is called the 3′-5′ strand. In contrast to the 5′-3′ strand, we assign a hit to the position at the complementary strand where the actual hit ends (Fig. 1). This means, that S'_j refers to the score of the nucleotides $\sigma'_{j+\ell-1} \dots \sigma'_j$ where σ'_k denotes the complementary letter of σ_k.

We call a position a *hit* if the corresponding window yields a score s higher than a certain threshold t. Denoting a hit at position j by the indicator random variable $Y_j = 1$, we obtain the definition: $Y_j := \mathbf{1}[S_j \geq t]$. Similarily, a detected binding site on the complementary strand at position j is denoted by $Y'_j = 1$ (Fig. 1).

2.2. Statistics for binding site detection

The threshold t controls the probabilities α and β given by $\alpha := \mathbb{P}_{H_0}(S_j \geq t)$ and $\beta := \mathbb{P}_{H_1}(S_j < t)$ where H_0 is the null model corresponding to a random sequence and H_1 the model for the binding site. Using the convolution of the position-specific score distributions according to the background model (Rahmann, 2003), we can compute the threshold t depending on the choice of α. The parameter α has to be very small as the expected number of false positives on a sequence of length n is $2n \cdot \alpha$ for both strands. To control the power of the PFM to seperate compatible from non-compatible words, we set the threshold as described by Pape et al. (2006). There, the threshold is set such that α and β are balanced on a reasonable level. In general, the count statistic is robust against the actual threshold as long as neither all positions are hits nor none because α is incorporated into its computation.

2.3. Count statistic

After this review of the detection of binding sites, we move on to the statistics of the number of detected binding sites. As previously mentioned, the probability of detecting a binding site by chance in a symmetric i.i.d. sequence model is α. Hence, the indicator random variables Y_j and Y'_j have a Bernoulli distribution with $\mathbb{P}_{H_0}(Y_j = 1) = \mathbb{P}_{H_0}(Y'_j = 1) = \alpha$ and $\mathbb{P}_{H_0}(Y_j = 0) = \mathbb{P}_{H_0}(Y'_j = 0) = 1 - \alpha$. Let X denote the number of binding sites in a region of length n of a sequence:

$$X = \sum_{j=0}^{n-\ell} (Y_j + Y'_j).$$

In fact, Y_j and Y'_j are defined on an infinite sequence but in practice we are concerned with finite sequences. Hence, the dependencies of Y_j and Y'_j are different at the beginning and the end of the sequence. These boundary effects are negligibly small under the rare hit assumption (Barbour et al., 1992). The rare hit assumption holds because we set α and the threshold t such that only a very small fraction of all possible words have a score greater than or equal to the threshold.

Now, we can compute the enrichment of binding sites on a sequence using the probability $p = \mathbb{P}_{H_0}(X \geq x)$ as p-value where x is the observed number of binding sites. Although we know the probability of Y_j and we assume the symmetric i.i.d. sequence model, calculation of p is not straightforward due to dependencies between Y_js and Y'_js. The dependencies are caused by self-overlap of the PFM and by incorporation of the complementary strand both leading to overlapping binding sites.

2.4. Computing probabilities of clumps

Dealing with overlapping hits requires a refined vocabulary: A hit and its overlapping hits together can be defined as a clump. The size of the clump corresponds to the number of contained overlapping hits. Also a single hit without any overlaps is called a clump of size 1. Thus, a clump is a left- and right-maximal set of overlapping hits on both strands.

Now, we incorporate the notion of clumps into our definitions. We assume that the number of clumps N is distributed as a Poisson random variable $\mathcal{P}(r)$ with unknown rate parameter r. The size Z_i of the clumps are assumed to be identically and independently distributed by an unknown probability vector $\vec{\theta}$ (Fig. 2). Both assumptions can be justified by the fact that they hold for word counting (Robin, 2002). The number of counts per sequence is given by $X = \sum_{i=1}^{N} Z_i$. X follows a compound Poisson distribution $\mathcal{CP}(r, \vec{\theta})$.

In the remaining part of this section, we show how to compute approximations of the unknown parameters of rate r and the probability vector $\vec{\theta}$. First of all, we reduce the computation of r to the computation of $\vec{\theta}$. Then, we start with the probability θ_1 of having exactly one hit. Based on θ_1, we recursively compute the remaining parameters. Furthermore, the analysis of the resulting formulation discovers two characteristic values describing the self-overlap of the PFM.

2.4.1. Computing the rate r. The parameter r is the rate of clump occurrences. We cannot use the probability α of a false positive directly to compute it because α is the probability for hits including overlapping hits. In contrast, we need the rate for non-overlapping hits which is equal to the rate of clumps. Using the law of total probability, we obtain:

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}(X \mid N)] = \mathbb{E}[N\mathbb{E}(Z)] = \mathbb{E}[N]\mathbb{E}[Z] = r\mathbb{E}[Z].$$

Thus, we can express r in terms of $\vec{\theta}$: $\mathbb{E}[X]$ is the expected number of hits. The probability of a hit by chance is α as we defined the threshold in this way. Hence, the expected number of hits is given by the fact that a hit can occur at each sequence position on each strand, thus

$$r = \frac{\mathbb{E}[X]}{\mathbb{E}[Z]} = \frac{2\alpha(n - \ell + 1)}{\sum_{i>0} i\,\theta_i}. \tag{1}$$

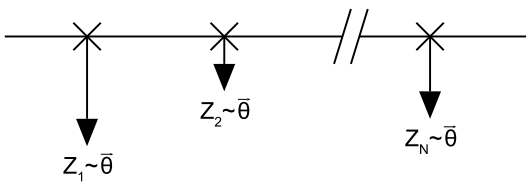

FIG. 2. The horizontal line symbolizes the sequence. At certain position (marked by a cross), a clump occurs. The size of each clump is modeled by an i.i.d. probability vector $\vec{\theta}$. The number of clumps N is distributed as a Poisson random variable $\mathcal{P}(r)$.

2.4.2. Parameters for the probability vector. The probability vector $\vec{\theta} = (\theta_i)_{i>0}$ contains probabilities for the different sizes of clumps $i > 0$. Here, we show how to compute approximations $(\tilde{\theta}_i)_{i>0}$ of these probabilities. At first, we focus on one strand of the sequence only. Subsequently, we extend the approach to deal with the complementary strand as well.

θ_1 corresponds to the event that there is exactly one hit at a certain position j while no overlapping hits occur given the hit at position j. Using the fact that an overlapping hit is a hit within the range of the length ℓ of the PFM, we obtain

$$\theta_1 = \mathbb{P}_{H_0}(Y_{j-\ell+1} = 0, \ldots, Y_{j-1} = 0, Y_{j+1} = 0, \ldots, Y_{j+\ell-1} = 0 \mid Y_j = 1). \tag{2}$$

The conditional probability on the right hand side of Equation (2) is hard to compute because the events in the collection $(\{Y_{j+k} = 0\})_{-\ell+1 \le k \le \ell-1, k \ne 0}$ are not independent, given $\{Y_j = 1\}$. However, in a first order approximation we pretend that conditional independence holds and compute

$$\tilde{\theta}_1 = \prod_{k=-\ell+1, k\ne 0}^{\ell-1} \mathbb{P}_{H_0}(Y_{j+k} = 0 \mid Y_j = 1) = \prod_{k=-\ell+1, k\ne 0}^{\ell-1} (1 - \mathbb{P}_{H_0}(Y_{j+k} = 1 \mid Y_j = 1)). \tag{3}$$

Due to the symmetric i.i.d. random sequence, we can prove the symmetry $\mathbb{P}_{H_0}(Y_{j-k} = 1 \mid Y_j = 1) = \mathbb{P}_{H_0}(Y_{j+k} = 1 \mid Y_j = 1)$ by applying the law of conditional probabilities and substituting $j = j' + k$:

$$\begin{aligned}
\mathbb{P}_{H_0}(Y_{j-k} = 1 \mid Y_j = 1) &= \frac{\mathbb{P}_{H_0}(Y_{j-k} = 1, Y_j = 1)}{\mathbb{P}_{H_0}(Y_j = 1)} \\
&= \frac{\mathbb{P}_{H_0}(Y_{j'+k-k} = 1, Y_{j'+k} = 1)}{\mathbb{P}_{H_0}(Y_{j'+k} = 1)} \\
&= \mathbb{P}_{H_0}(Y_{j'+k} = 1 \mid Y_{j'} = 1).
\end{aligned} \tag{4}$$

Symmetry follows due to the symmetric i.i.d. background model. Thus, we obtain for Equation (3)

$$\tilde{\theta}_1 = \left(\prod_{k=1}^{\ell-1} [1 - \mathbb{P}_{H_0}(Y_{j+k} = 1 \mid Y_j = 1)] \right)^2. \tag{5}$$

Next, we extend the approach to both strands by continuing to assume conditional independence for all hits (Pape et al., 2006) and simplify notation by

$$\gamma_k = \mathbb{P}_{H_0}(Y_{j+k} = 1 \mid Y_j = 1), \qquad \gamma'_k = \mathbb{P}_{H_0}(Y'_{j+k} = 1 \mid Y_j = 1),$$

where Y' refers to the hit random variable on the other strand. Hence, Equation (5) becomes

$$\tilde{\theta}_1 = (1 - \gamma'_0) \prod_{k=1}^{\ell-1} (1 - \gamma_k)^2 (1 - \gamma'_k)^2. \tag{6}$$

This term contains the probability of two overlapping hits $\mathbb{P}_{H_0}(Y_{j+k} = 1 \mid Y_j = 1)$ for k as given above in γ_k, and correspondingly in γ'_k for the reverse strand. Thus, we need to compute the probability of the score at position $j + k$ exceeding the threshold given there is a hit at position j: $\mathbb{P}_{H_0}(S_{j+k} \ge t \mid Y_j = 1)$. In previous work (Pape et al., 2006), we used the nucleotide distribution given by the PFM for the overlapping part. Unfortunately, the performance of the approximation varies signficiantly between different PFMs. Therefore, we model the dependencies explicitly.

2.4.3. Probability of two overlapping hits. Here, we compute the exact value for $\mathbb{P}_{H_0}(S_{j+k} \ge t \mid Y_j = 1)$. The event $\{Y_j = 1\}$ is equal to the event $\{S_j \ge t\}$ due to the definition of the indicator random

variable Y_j. We define the set of scores which are assumed to be integers by

$$\mathcal{S} := \left\{ s : \sum_{\kappa=0}^{\ell-1} \min_{\sigma\in\Sigma} \Psi_{\kappa,\sigma} \le s \le \sum_{\kappa=0}^{\ell-1} \max_{\sigma\in\Sigma} \Psi_{\kappa,\sigma} \right\}. \tag{7}$$

Then, the scores greater than or equal to the threshold can be defined by $\mathcal{S}_t := \{s \in \mathcal{S} : s \ge t\}$. Using these definitions, we can express $\mathbb{P}_{H_0}(Y_{j+k} = 1 \mid Y_j = 1)$ in terms of a two-dimensional score distribution

$$\begin{aligned}\gamma_k = \mathbb{P}_{H_0}(Y_{j+k} = 1 \mid Y_j = 1) &= \frac{\mathbb{P}_{H_0}(S_{j+k} \ge t, S_j \ge t)}{\mathbb{P}_{H_0}(S_j \ge t)} \\ &= \frac{1}{\alpha} \sum_{s\in\mathcal{S}_t} \sum_{s'\in\mathcal{S}_t} \mathbb{P}_{H_0}(S_{j+k} = s', S_j = s). \end{aligned} \tag{8}$$

The overlapping probabilities γ'_k can be computed correspondingly.

Considering the scores as state space, the S_js become a first-order Markov chain (Fu and Koutras, 1994) because the score only depends on the score of the previous position. Hence, Equation (8) can also be written in terms of its transition matrix to the kth power. For the sake of simplicity, we focus on the two-dimensional score distribution for each k. This distribution can be computed by the two-dimensional convolution of the position specific score distributions. An efficient dynamic programming algorithm is presented in Section 2.5.

2.4.4. Probability of an i-clump with $i > 1$. We recursively compute the probability $\tilde{\theta}_i$ to have a clump with exactly i hits for $i > 1$. Without loss of generality, we assume that we count hits starting with Y_0, Y'_0, Y_1, Y'_1, Y_2, and so on. Considering a clump of size two, the first overlapping hit at a clump position j can either occur in the interval $k \in [j+1, j+\ell-1]$ on the same strand or in the interval $k \in [j, j+\ell-1]$ on the opposite strand. The idea is to cancel the probability in Equation (6) for each position k and to replace it with the probability of a hit at this position. We denote these *extension* factors ξ_k for a hit on the 5′-3′ strand and ξ'_k for a hit on the 3′-5′ strand. We obtain for a pair of hits for $0 < k < \ell$

$$\mathbb{P}_{H_0}(Y_{j+k} = 1, Y'_{j+k} = 0, Y'_j = 0, \{Y_{j+\kappa} = 0, Y'_{j+\kappa} = 0\}_{-\ell<\kappa<k+\ell,\kappa\neq 0,k} | Y_j = 1) \approx \tilde{\alpha}_1 \cdot \xi_k,$$

$$\mathbb{P}_{H_0}(Y'_{j+k} = 1, Y_{j+k} = 0, Y'_j = 0, \{Y_{j+\kappa} = 0, Y'_{j+\kappa} = 0\}_{-\ell<\kappa<k+\ell,\kappa\neq k,0} | Y_j = 1) \approx \tilde{\alpha}_1 \cdot \xi'_k,$$

$$\mathbb{P}_{H_0}(Y'_j = 1, \{Y_{j+\kappa} = 0, Y'_{j+\kappa} = 0\}_{-\ell<\kappa<\ell,\kappa\neq 0} | Y_j = 1) \approx \tilde{\alpha}_1 \cdot \xi'_0.$$

For the definitions of ξ_k, ξ'_k, and ξ'_0, it is important to note that one also has to replace the probability at the other strand at position k with the probability γ'_0 for an exact palindromic hit at position k of the old hit except the new hit is on the 3′-5′ strand. In this case (ξ'_k), an exact palindromic hit is not possible (because we would have counted the hit before). We also replace the probabilities for hits at the subsequent positions given the hit at position j with the probabilities of a hit given the hit at $j+k$. Lastly, one has to extend the positions without a hit to the positions covered by the new hit but not by the former hit. Thus, we obtain the definitions for $0 < k < \ell$

$$\xi_k := \frac{\gamma_k}{1-\gamma_k} \cdot \frac{1-\gamma'_0}{1-\gamma'_k} \cdot \left(\prod_{\kappa=1}^{\ell-k-1} \frac{1-\gamma_\kappa}{1-\gamma_{k+\kappa}} \cdot \frac{1-\gamma'_\kappa}{1-\gamma'_{k+\kappa}} \right) \cdot \left(\prod_{\kappa=\ell-k}^{\ell-1} (1-\gamma_\kappa)(1-\gamma'_\kappa) \right), \tag{9a}$$

$$\xi'_k := \frac{\gamma'_k}{1-\gamma'_k} \cdot \left(\prod_{\kappa=1}^{\ell-k-1} \frac{1-\gamma_\kappa}{1-\gamma_{k+\kappa}} \cdot \frac{1-\gamma'_\kappa}{1-\gamma'_{k+\kappa}} \right) \cdot \left(\prod_{\kappa=\ell-k}^{\ell-1} (1-\gamma_\kappa)(1-\gamma'_\kappa) \right). \tag{9b}$$

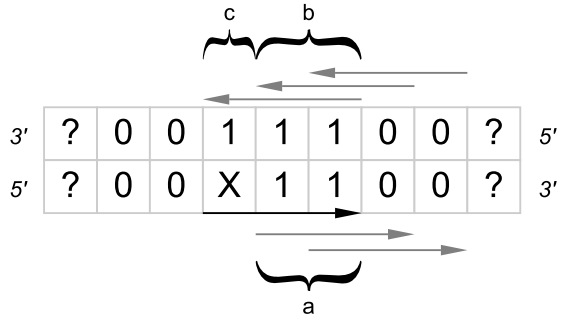

FIG. 3. X and the black arrow indicate the given hit, while 0s indicate where no hit is allowed, 1 denotes the possibility of an overlapping hit (marked by grey arrows), and ? is a hit or no hit. The letters a, b, and c label the three different type of hits: type (a) is a hit on the same strand, type (b) hits on the complementary strand but not palindromic, and type (c) is the palindromic hit.

For the other strand, we must not replace the exact palindromic hit because the hit is already on that strand. We have an overlapping hit if any of these encoded events occur. Thus, we can sum up the terms (Fig. 3)

$$\xi := \sum_{k=1}^{\ell-1} \xi_k, \qquad \xi' := \sum_{k=1}^{\ell-1} \xi'_k, \qquad \xi'_0 := \frac{\gamma'_0}{1-\gamma'_0}. \tag{10}$$

We split up the different types of hits into a, b, and c because they differ with respect to which hit can follow (see Fig. 4). Type (a) encodes a hit on the 5′-3′ strand. The hit can be followed by a hit on the same strand (a), a hit on the complementary strand (b), or an exact palindromic hit (c). Type (b) is a hit on the 3′-5′ strand excluding an exact palindromic hit. After it, types (a) and (b) can follow. An exact palindromic hit is not possible as the hit itself is on the 3′-5′ strand. Type (c) stands for the exact palindromic hit. Everything but another exact palindromic hit can follow.

Thus, an additional hit of type (a) can be preceded by a hit of type (a), (b), or (c). For (b) the same logic applies. In contrast, a palindromic hit (c) can only occur after a hit of type (a). Without these considerations, we would allow more than two hits at the same position. This gives us a linear system of recurrences to compute an approximation of $\vec{\theta}$. Again, assuming that conditional independence holds, we

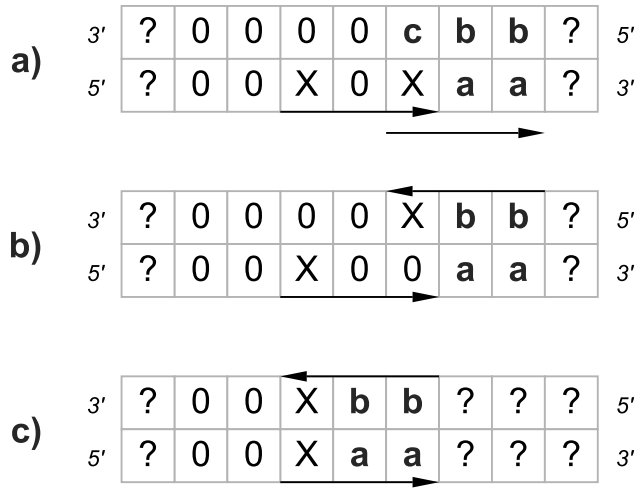

FIG. 4. The figure shows the three different types of hits (a), (b), and (c) in a situation of two previous hits. Furthermore, for each type of hit, the possible subsequent type of hit is indicated by the corresponding identifier (a, b, or c). While (a) and (b) can occur after each type of hit, (c) can only occur after type (a).

have $\tilde{\theta}_{i+1} := \tilde{\theta}_1(a_i + b_i + c_i)$, where

$$a_1 := \xi, \qquad a_{i+1} := (a_i + b_i + c_i)\xi, \tag{11a}$$

$$b_1 := \xi', \qquad b_{i+1} := (a_i + b_i + c_i)\xi', \tag{11b}$$

$$c_1 := \xi'_0, \qquad c_{i+1} := a_i\xi'_0. \tag{11c}$$

2.4.5. Closed formula. Although the above formulas suffice to compute an approximation for $\vec{\theta}$, the reformulation as a closed formula and its further analysis reveals interesting insights. At the end, we obtain two characteristic values which describe the self-overlap of the PFM.

We can write the recursive formulas (11a) to (11c) by matrix notation for $i > 0$

$$\begin{pmatrix} a_{i+1} \\ b_{i+1} \\ c_{i+1} \end{pmatrix} = \begin{pmatrix} \xi & \xi & \xi \\ \xi' & \xi' & \xi' \\ \xi'_0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} a_i \\ b_i \\ c_i \end{pmatrix} =: A \cdot \begin{pmatrix} a_i \\ b_i \\ c_i \end{pmatrix}.$$

Thus, we get the closed formula for $i \geq 0$ with $\vec{\xi} = (\xi, \xi', \xi'_0)^T$

$$\begin{pmatrix} a_{i+1} \\ b_{i+1} \\ c_{i+1} \end{pmatrix} = A^i \ \vec{\xi}.$$

Furthermore, we obtain $\tilde{\theta}_{i+1}$ using the recurrence formula for $i > 0$

$$\tilde{\theta}_{i+1} = (1, 1, 1) \cdot A^i \cdot \vec{\xi} \cdot \tilde{\theta}_1.$$

We decompose $A = B^{-1} \ \Lambda \ B$ where B contains the eigenvectors of A and the diagonal matrix Λ the corresponding eigenvalues λ_1, λ_2, λ_3 given by

$$\lambda_{1,2} = \frac{\xi + \xi'}{2} \pm \frac{1}{2}\sqrt{w}, \qquad \lambda_3 = 0,$$

with $w = (\xi + \xi')^2 + 4\xi \ \xi'_0$. Hence, we can denote $\tilde{\theta}_{i+1}$ in terms of the eigenvalues

$$\tilde{\theta}_{i+1} = (1 \ \ 1 \ \ 1) \ B^{-1} \ \Lambda^i \ B \ \vec{\xi} \ \tilde{\theta}_1 = (u\lambda_1^i + v\lambda_2^i)\tilde{\theta}_1 \tag{12}$$

where u and v are computed by solving the linear system

$$\begin{pmatrix} 1 & 1 \\ \lambda_1 & \lambda_2 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \tilde{\theta}_1 = \begin{pmatrix} \tilde{\theta}_1 \\ \tilde{\theta}_2 \end{pmatrix},$$

whereby $\tilde{\theta}_2 = (\xi + \xi' + \xi'_0)\tilde{\theta}_1$ using the recursive formula. Finally, we obtain the solution

$$u, v = \frac{w \pm (\xi + \xi' + 2\xi'_0)\sqrt{w}}{2w}.$$

In addition to the benefits of a closed formula, the expression in Equation (12) shows that the asymptotics of the clump size only depend on the first two eigenvalues λ_1, λ_2. Obviously, the following inequalities hold: $\lambda_2 \leq 0$ and $\lambda_1 > -\lambda_2$. Hence, the series $\tilde{\theta}_i$ converges to zero, $\lim_{i\to\infty} \tilde{\theta}_i = 0$, if $\lambda_1 < 1$. This condition holds if most of the words of length ℓ do not exceed the threshold which is in practice always true. In most cases, we can also assume $\lambda_1 < 1$ since for $\lambda_1 > 1$ we can transform the PSSM and the threshold to its complement which yields $\lambda_1 < 1$ and correspondingly use the complementary statistics.

2.4.6. Two characteristic values for PFMs. The eigenvalues can be used as descriptive values for the PFM: We call a PFM a palindrome if it allows two hits at the same position on both strands. Considering a PFM which is not a palindrome, we have $\gamma'_0 > 0$, thus, $\xi'_0 = 0$. Hence, matrix A has only rank 1 and we only obtain one non-zero eigenvalue λ_1 and $u = 1$. Therefore, $\tilde{\theta}_{i+1} = \lambda_1^i \tilde{\theta}_1$ decreases exponentially. Higher values of λ_1 decelerate convergence. Since $\tilde{\theta}_1$ is the probability of a clump of size one, which means that no overlap occurs, and λ_1 corresponds to the probability of an overlap, obviously $\tilde{\theta}_1 = 1 - \lambda_1$ holds. Hence, the clump size has a shifted geometric distribution. This case is similar to the compound Poisson model for one word considering only one strand described by Robin (2002). Discarding the complementary strand always leads to $\lambda_2 = 0$ and then both models are equivalent.

A palindromic PFM has $\lambda_2 < 0$ since $\xi'_0 > 0$. If ξ'_0 dominates the eigenvalues we obtain $\lambda_2 \approx -\lambda_1$. In addition, it follows that $v \approx -u$. Thus, $\tilde{\theta}_{i+1} \approx u[1 + (-1)^i]\lambda_1^i \tilde{\theta}_1$. This leads to $\tilde{\theta}_i \approx 0$ for odd i. It indicates that the probability of an odd clump size is approximately equal to zero. As we assumed a palindromic PFM with a very high probability of a palindromic hit, one almost always detects a hit on both or neither strands. In summary, λ_1 describes the speed of convergence of $\tilde{\theta}_i$ to zero and $-\lambda_2$ correlates with the tendency of palindromic hits.

2.4.7. The p-value for number of hits. Now, we can compute the approximations of the probability vector $\vec{\theta}$ and the rate parameter r using Equation (1). Using the approximations for the parameters, we can compute the distribution for the number of hits $x \geq 0$. Since the number of hits X is distributed as $\mathcal{CP}(r, \vec{\theta})$ we can apply formulas for the compound Poisson distribution (Kemp, 1967):

$$\mathbb{P}_{H_0}(X = 0) = \exp(-r),$$

$$\mathbb{P}_{H_0}(X = x + 1) = \frac{r}{x+1} \sum_{x'=0}^{x} (x + 1 - x')\theta_{x+1-x'} \mathbb{P}_{H_0}(X = x').$$

The p-value for the occurrence of $x \geq 0$ hits is computed by:

$$p = \mathbb{P}_{H_0}(X \geq x) = 1 - \sum_{x'=0}^{x-1} \mathbb{P}_{H_0}(X = x').$$

2.4.8. The p-value for number of clumps. The underlying Poisson process for the count statistic is given by $N \sim \mathcal{P}(r)$. Since N is the number of clumps, one can use $\mathcal{P}(r)$ to compute p-values p' for clumps as the count entity. In this case, we only need to compute the rate r. Equation (1) can be approximated by

$$\tilde{r} = \frac{2\alpha(n - \ell + 1)}{\sum_{i>0} i\,\tilde{\theta}_i}.$$

Next, we show how to compute $\tilde{r}$ efficiently by substituting the sum under the assumption $-1 < \lambda_2 \leq 0 \leq \lambda_1 < 1$. Using the expression in Equation (12), we obtain

$$\begin{aligned}
\sum_{i>0} i\,\tilde{\theta}_i &= \sum_{i>0} i\,(u\lambda_1^{i-1} + v\lambda_2^{i-1}) \cdot \tilde{\theta}_1 \\
&= \left(\frac{u}{\lambda_1} \sum_{i>0} i\,\lambda_1^i + \frac{v}{\lambda_2} \sum_{i>0} i\,\lambda_2^i \right) \cdot \tilde{\theta}_1 \\
&= \left(\frac{u}{(1-\lambda_1)^2} + \frac{v}{(1-\lambda_2)^2} \right) \cdot \tilde{\theta}_1.
\end{aligned}$$

Again, this equation has an interpretation in terms of the self-overlap of the PFM. In case of a non-palindromic PFM, the second term is equal to zero since $\lambda_2 = 0$ and, therefore, $v = 0$. From $\lambda_2 = 0$, it also follows that $w = (\xi + \xi')^2$, hence, $u = 1$. Since a non self-overlapping PFM has λ_1 near to zero, the above equation is equal to 1. Thus, the expected value of the clump size is equal to 1 and the rate for

clumps is $\tilde{r} = 2\alpha(n - \ell + 1)$. Furthermore, $\tilde{\theta}_1 = 1$ contains all the weights of the probability vector. Then, $\mathcal{CP}(\tilde{r}, (1, 0, \dots)) \sim \mathcal{P}(\tilde{r})$. Hence, applying the derived statistic to a non self-overlapping PFM leads to a Poisson process with rate $\tilde{r}$.

2.5. *An efficient algorithm for computing overlap probabilities*

In Equation (8), we have to compute the joint event of two scores greater than or equal to the threshold. Given a position j and a shift k, the two scores induce a two dimensional distribution. The first component is the score s of the PSSM starting at position j. The second component is the score s' of the PSSM beginning at position $j + k$. As an example, consider a PSSM which only accepts 'CC'. In the case of a shift $k = 1$, the score s' can only exceed the threshold if s is above the threshold. Thus, both scores are not independent for $0 \leq k < \ell$. Since scores are the sum of the position specific scores $\Psi_{i,\cdot}$ we can decompose the score into each pair of positions $j + i$ and $j + i + k$ which point to the same sequence position, and, thus, to the same nucleotide. Then, pairs of scores are independent. Hence, we can use a dynamic programming algorithm.

The dynamic programming approach is often used for the computation of the one dimensional score distribution (Staden, 1989; Claverie and Audic, 1996; Wu et al., 2000; Rahmann, 2003; Rahmann et al., 2003; Beckstette et al., 2006). We extend this approach to two dimensions. Let $Q_i^{(k)}(s, s')$ denote the probability for a score s at the first $i + 1$ positions of the PSSM and a score s' at the first $i + 1 - k$ positions of the PSSM shifted by k positions. We compute this value by summing over the probabilities of the last step $i - 1$ which yield a score s and s' after observing any nucleotide with its respective score at step i. With $\Psi_{\kappa,\cdot} := 0$ for $\kappa < 0$ or $\kappa \geq \ell$, we obtain for $0 \leq k < \ell$ and $0 \leq i < \ell + k$

$$Q_{-1}^{(k)}(s, s') := \begin{cases} \text{undefined} & \text{if} \quad s \neq 0 \quad \text{or} \quad s' \neq 0, \\ 1 & \text{else}, \end{cases}$$

$$Q_i^{(k)}(s, s') := \sum_{\sigma \in \Sigma} Q_{i-1}^{(k)}(s - \Psi_{i,\sigma}, s' - \Psi_{i-k,\sigma}) \cdot \pi_\sigma.$$

After the last step, $Q_{\ell+k-1}^{(k)}(s, s')$ contains the probability to observe score s starting at position j and score s' starting at position $j + k$. Therefore, $\mathbb{P}_{H_0}(S_{j+k} = s', S_j = s) = Q_{\ell+k-1}^{(k)}(s, s')$ and, hence, we can solve Equation (8).

2.5.1. Speed improvement. The practical running time of the algorithm can be improved significantly by some modifications. The last k steps do not modifiy the scores starting at position j since $\Psi_{\kappa,\cdot} = 0$ for $\kappa \geq \ell$. Hence, instead of the k two-dimensional convolutions, we can obtain $Q_{\ell+k-1}^{(k)}$ in one step by using the one dimensional convolution of the last k positions (Fig. 5). Since Equation (8) sums over all scores

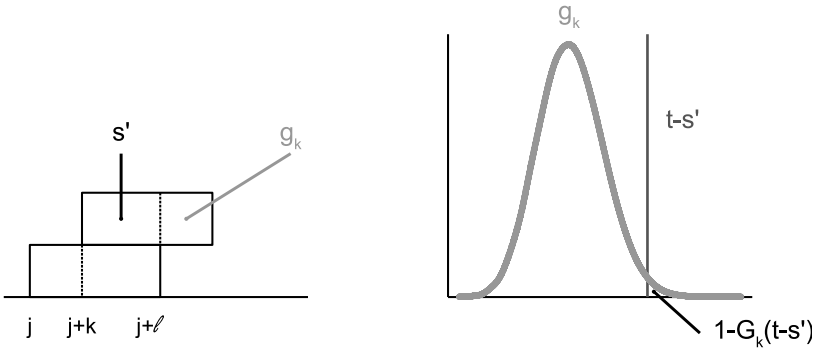

FIG. 5. The left part of the figure shows the sequence with two overlapping hits at position j and $j + k$. The score of the overlapping part of the second hit is given by s'. The score of the non-overlapping part is a random variable whose distribution is the convolution of the position-specific scores of the remaining positions of the PFM g_k. The right part of the figure visualizes this distribution and the probability of the second hit $1 - G_k(t - s')$.

s starting at position j, we can do the summation before adding the remaining scores:

$$R^{(k)}(s') = \sum_{s \in \mathcal{S}_t} Q^{(k)}_{\ell-1}(s, s').$$

Let f_κ denote the position specific score distribution at position κ of the PSSM and g_k denotes the score distribution for the non-overlapping part for a shift k:

$$g_k := f_{\ell-k} * \cdots * f_{\ell-1}.$$

Using the recursion $g_{k+1} = g_k * f_{\ell-k-1}$, we can use a dynamic programming approach for computing the convolution. Denoting the cumulative distribution for g_k with G_k, we can rewrite Equation (8) by:

$$\gamma_k = \mathbb{P}_{H_0}(Y_{j+k} = 1 \mid Y_j = 1) = \frac{1}{\alpha} \sum_{s' \in \mathcal{S}_t} [1 - G_k(t - s')] R^{(k)}(s').$$

We can further improve the speed by removing scores in each step of the calculation of Q which are too small to reach the threshold (see Beckstette et al. [2006] for the one-dimensional case). Hence, we define intermediate thresholds t_i by

$$t_i := t - \sum_{\kappa=i+1}^{\ell-1} \max_{\sigma \in \Sigma} \Psi_{\kappa,\sigma}. \tag{13}$$

In each step i, we can remove scores $s < t_i$ and $s' < t_{i-k}$. In addition, one can merge scores which will exceed the threshold t for sure. The corresponding intermediate thresholds t'_i are defined analogously to Equation (13) by substituting min with max. Then, in each step i, scores $s \geq t'_i$ and $s' \geq t'_{i-k}$ can be merged.

We can further speed up the algorithm by enhancing the effect of these improvements (see Beckstette et al. [2006] for a similar idea). Processing positions of the PFM with high information content first, discards many scores which can't exceed the threshold at all in the first steps. In addition, indefinite positions (processed at the end) often do not change the score significantly such that either the score has already been discarded or the score surely exceeds the threshold. This reduces the size of Q significantly.

In summary, the algorithm takes advantage from both a high and a low threshold t: On the one hand, the higher the threshold, the more scores can be removed in the beginning steps because scores will not be able to exceed the threshold at all. On the other hand, a low threshold yields many scores which surely exceed the threshold. As those scores can be merged, the number of different scores (size of Q) stays low.

2.5.2. Time complexity. The complexity of the algorithm for the computation of Q depends on the length of the PFM ℓ, the size of the set $\mathcal{S}$ of all scores, and the alphabet size $|\Sigma|$: $O(\ell^2 |\mathcal{S}|^2 |\Sigma|)$. The length of the PFM ℓ and the alphabet size $|\Sigma|$ are primitives and, therefore, cannot be reduced any further. In contrast, $|\mathcal{S}|$ is a constructed set, hence, we have to analyze its complexity. It is important to note that the size of $\mathcal{S}$ is independent of the threshold and, therefore, of the number of compatible words. Furthermore, $|\mathcal{S}|$ does not grow exponentially with increasing length of the PFM because the scores of a new column are only added to the overall scores. This only increases the size of $\mathcal{S}$ linearly with increasing PFM length.

3. COMPUTATIONAL RESULTS

3.1. Data

3.1.1. Sequences. To compare the new statistic with previous approaches, we use a simulation study. We simulate 100,000 sequences of length 10,000 with an arbitrarily selected GC content of 40% using an i.i.d. model. These sequences are annotated by binding sites of artificially constructed and real PFMs (see next paragraph). Counting the number of hits/clumps per sequence and computing the frequency for

each count, one retrieves a simulated count distribution for each PFM. Comparing the simulated count distribution with the theoretically approximated distributions, we can easily assess the accuracy of the approximations, as well, as comparing the approximations between themselves.

3.1.2. PFMs. We artificially construct four PFMs, each of them carrying a certain characteristic regarding self-overlap (Crooks et al., 2004) (Fig. 6):

- *Nothing*: a PFM without any self-overlaps
- *Palindrome*: a PFM with a likely hit on the complementary strand
- *Repeat*: a PFM where the suffix matches the prefix such that one expects overlapping hits in a chain
- *Repeatpalindrome*: a combination of the palindrome and the repeat.

Furthermore, we arbitrarily pick one real PFM from TransFac (Matys et al., 2003) with a self-overlapping structure to show that the gain in accuracy is relevant in practice. We select the palindromic PFM M00950 corresponding to the binding site of the MADS domain protein AGAMOUS-like 15 (AGL15) (Tang and Perry, 2003).

In a pre-processing step, we regularize the PFMs to ensure strictly positive frequencies. Thus, we add pseudocounts to the position specific distributions according to the information content of the position (Rahmann, 2003). In fact, positions with low information content are shifted towards the background distribution. For positions with high information content, the difference to the background distributions is enforced. Then, we compute PSSMs from the regularized PFMs by taking the log-likelihood ratio of the nucleotide frequencies of the binding site and the background model.

We set the threshold for each PFM according to Pape et al. (2006) ensuring that the probability α_{500} for at least one false positive in a sequence of length 500 for any higher threshold is 10% at maximum. Thus, in the case that one cannot balance α_{500} and β, we obtain $\alpha_{500} \approx 0.1$. Furthermore, in case of a balanced threshold, α_{500} and β will not be exactly equal due to the discrete nature of the score and, thus, of α_{500} and β. Applying this procedure, the number of compatible words for PFM "nothing" using a threshold $t = 136$ with $\alpha_{500} = 0.013$ and $\beta = 0.01$ is 1,142 only containing unique words. PFM "palindrome" ($t = 99$, $\alpha_{500} = 0.129$ and $\beta = 0.325$) encodes 48 words based on 24 unique words while PFM "repeat" ($t = 128$, $\alpha_{500} = 0.118$ and $\beta = 0.553$) has 702 words without any non-unique words and, finally, "repeatpalindrome" ($t = 125$, $\alpha_{500} = 0.157$ and $\beta = 0.649$) yields 50 words from which 25 are unique. Thus, the compatible set of both PFMs with a palindromic structure contain each word twice (for each strand once). The Transfac PFM 'M00950' ($t = 118$, $\alpha_{500} = 0.0785$ and $\beta = 0.0747$) has 846,976 words with 429,812 unique words.

3.2. Standard count statistics

In this section, we present the previous count statistics that we compare our approach with. They are applied on the set of compatible words after removing redundant words such that the assumptions are met.

3.2.1. Binomial and Poisson approximation. A simple approach to compute the p-value for the number of hits is the assumption of independence between hits. Then, the number of hits have a binomial distribution. We obtain the p-value $p_B = 1 - B(x; 2n, \alpha)$ where $B(\cdot)$ is the cumulative binomial distribution with parameters x for the number of successes, $2n$ the number of trials and α the probability of success. The number of successes corresponds to the number of detected hits. As we consider hits on both strands and assume independence, the number of trials is twice the length of the sequence.

A statistic for the number of clumps can also be derived. A clump is defined such that it does not overlap with a previous clump. Hence, the probability α' of a clump can be calcuated by

$$\alpha' = 2(1-\alpha)^{2\ell-1}\alpha + (1-\alpha)^{2\ell-2} \cdot \alpha^2.$$

The term $(1-\alpha)^{2\ell-1}$ is the probability of no hit on the $\ell - 1$ positions before the hit on both strands plus no hit on the complementary strand at the current position. As a hit can occur on both strands, we multiply by the factor 2. The last term corresponds to a clump starting with a palindrome. The p-value is computed by $p'_B = 1 - B(x; n, \alpha')$.

Both binomial distributions can also be approximated by a Poisson distribution with parameter $r_P \approx 2(n - \ell + 1)\alpha$ for hits and $r'_P \approx (n - \ell + 1)\alpha'$ for clumps.

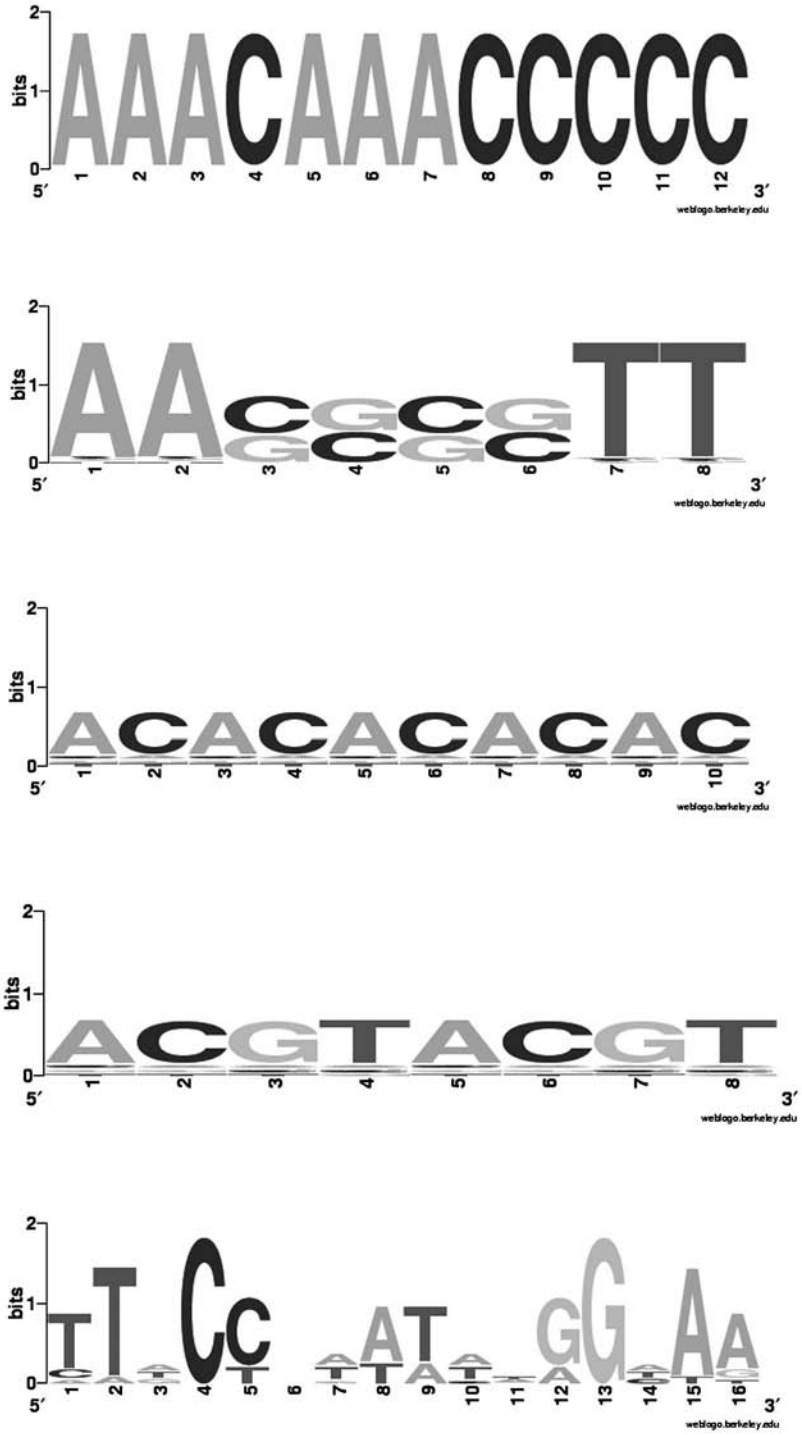

FIG. 6. Sequence Logos from left to right and top to bottom: "nothing," "palindrome," "repeat," "repeatpalindrome," and "M00950."

3.2.2. Chen-Stein Poisson approximation. Reinert and Schbath (1999) present a Chen-Stein Poisson approximation for the occurrences of multiple words. We can transform the PFM hit statistic to a word counting problem using the set of compatible words $\mathcal{W}$. Then, we can use the Chen-Stein approach to compute the Poisson approximation. There are several problems using this approach: First, enumeration of all words in $\mathcal{W}$ is not possible for longer PFMs. Second, the number of words in $\mathcal{W}$ might be fairly high which leads to large bounds on the total variation distance. Third, incorporation of the complementary strand can only be achieved by extending $\mathcal{W}$ by all reverse complementary words. This almost always breaks the necessary assumption for the approximation that no word is a substring of any overlap of any two other words. Therefore, we only use the approximation for the clump statistics because there only the weaker assumption that no word is a substring of any other word has to be fulfilled. This is indeed the case except for a palindrome.

In Roquain and Schbath (2007), an improved compound Poisson approximation is proposed. Since this approach only requires that no word is a substring of any other word for both the clump and the hit distribution, we also include this approach in the comparison. Still, the main drawback to enumerate all the compatible words remains. Furthermore, the approach involves multiple matrix multiplications where the two dimensions of the matrices are equal to the number of compatible words. This leads to numerical instabilities for large sets of compatible words.

3.2.3. Normal approximation. We can use another multiple word occurrence approach for the number of hits. Waterman (2000) shows how to compute the limiting covariance matrix as well as the limiting mean value. These values can be used as parameters for a normal distribution. Again, we need the set of all compatible words $\mathcal{W}$ for the computation.

3.3. Comparison of the different approaches using simulated data

We compare the approaches based on the p-values since the statistic will mainly be used to retrieve p-values for observed number of hits/clumps. We present them after taking the logarithm to base ten. Therefore, the p-p plots show log-p-values. The x-axis always refers to the simulated distribution while the y-axis corresponds to the approximated distribution. The more points are located on the diagonal, the better the approximation. Furthermore, points below the diagonal correspond to underestimation of the p-values while points above the diagonal are conservative approximations.

3.3.1. Artificial PFMs. Figure 7A shows the p-p plots of the "nothing" PFM. Most of the points lie on the diagonal. Furthermore, there is no big difference between the binomial and Poisson approximations, as well, as the the approach from Roquain and Schbath (2007) and the new approach. Only for very small p-values, there is a subtle difference between the approximations: The binomial and Poisson approaches seem to slightly outperform the others. However, the very small p-values are based on very few sequences because such high numbers of hits/clumps do not occur very often. Therefore, we ignore these points for interpretation. Only the normal approximation underestimates the p-values systematically. As the Poisson approximation works better, obviously, the rare word assumption is fullfilled instead of the normal approximation assuming often-occurring words. The results for clumps do not differ. As there are no overlaps, both the hit and the clump statistics are similar. Obviously, the new approach captures this non-self-overlap.

Figure 7B contains the results for the PFM "palindrome." The single distribution lying on the diagonal corresponds to our new approach. The binomial, Poisson, the normal and the Roquain and Schbath (2007) approach substantially underestimate the p-values. For the binomial and Poisson approximation, this is due to the fact that the number of hits is higher since the PFM tends to hit on both strands the same time. Furthermore, there are always pairs of points very close to each other. This is due to the hit on the complementary sequence which always occurs with a hit on the $5'$-$3'$ strand: Having one hit on one strand implies a second hit on the other strand. Obviously, only the new approach can deal with this. In contrast, the Roquain and Schbath (2007) approach does not lead to a reasonable approximation. Since the set of compatible words contains each word twice but the approach can only deal with a set of unique words, the weak approximation is not surprising. For statistics of clumps, the Roquain and Schbath (2007) and the new

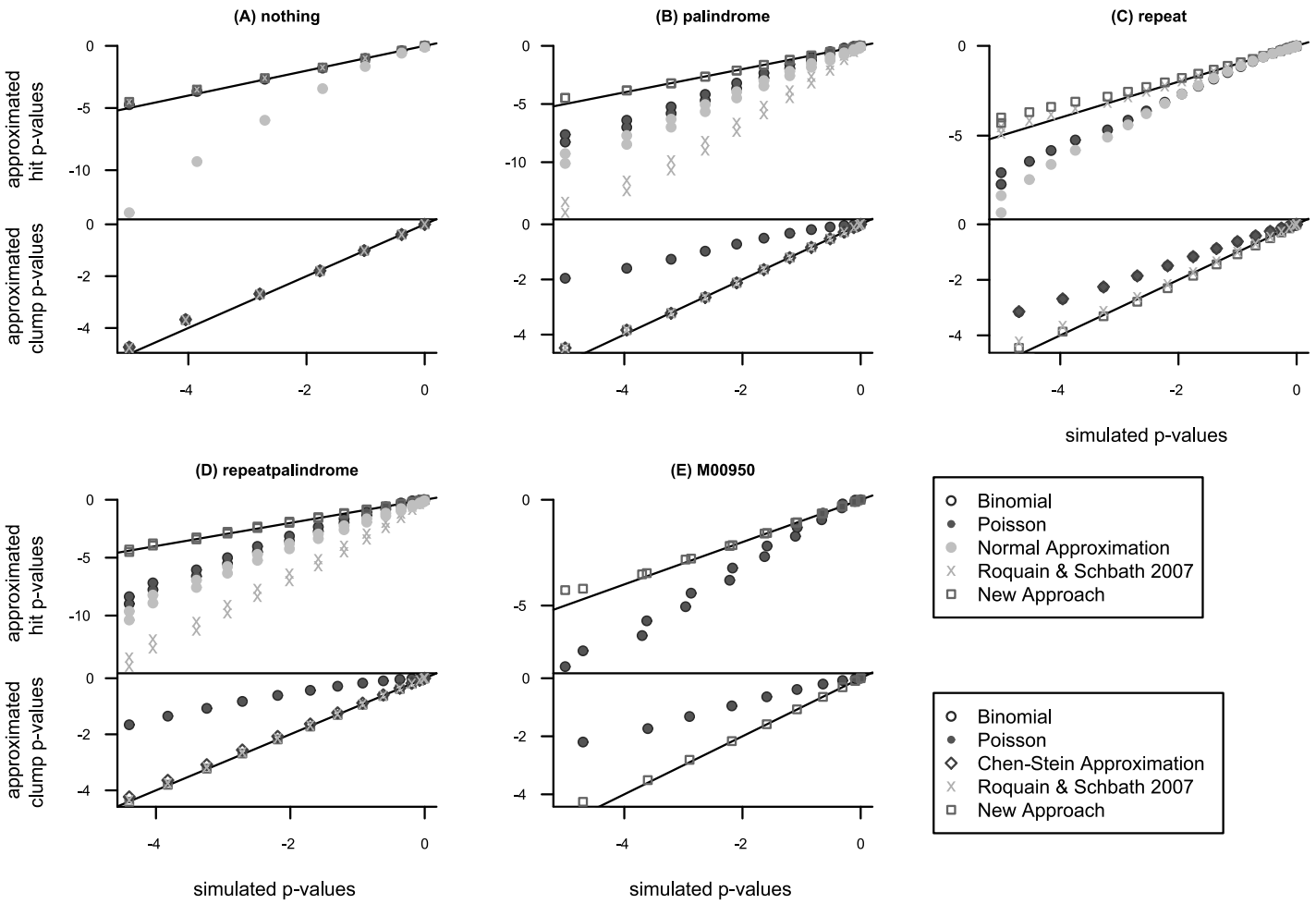

FIG. 7. Comparison of the simulated p-values (x-axis) with the approximated p-values (y-axis) in a log-scale p-p plot.

approach lie fairly on the diagonal, as well, as the Chen-Stein approximation. Here, the approximations for the Roquain and Schbath (2007) and the Chen-Stein approach work because for clumps redundant words in the set of compatible words have no influence.

Figure 7C compares the approximations for the "repeat" PFM. Binomial, Poisson, and normal approximations look very similar. Neither these nor the other two approaches lie on the diagonal, although the Roquain and Schbath and the new approach are more similar to the diagonal. In general, the Roquain and Schbath estimates are lower than the ones from the new approach and fit better to the diagonal. In addition, both approaches are conservative in contrast to the others which significantly underestimate the p-values. The approximations for the clump statistic are similar except that the new approach slightly underestimates the p-values for small number of clumps and overestimates them for higher number of clumps. In general, while the Roquain and Schbath approach obtains higher estimates leading to a more accurate approximation for small number of clumps but a slightly weaker approximation for higher number of clumps. The other approaches overestimate the p-values significantly. Here, the Chen-Stein approximation performs as bad as the binomial/Poisson approximation since the assumption that no word is a substring of the concatenation of any other two words is extensively violated.

In Figure 7D, the comparisons for the "repeatpalindrome" PFM is shown. In general, the approximations are similar to the ones of the "palindrome" PFM with some influence of the "repeat" PFM. This shows that the new approach can deal with both types of similarity at the same time in contrast to all other approaches. Only the new approach leads to a reasonable approximation of p-values for the number of hits.

3.3.2. PFM M00950. The results for the Transfac PFM M00950 are shown in Figure 7E. Using the balanced threshold to obtain a probability of a false positive in a region of 500 bp equal to 7.6%, the number of unique compatible words is equal to 429,812. Therefore, comparison with the Chen-Stein approximation, the normal approximation, and the Roquain and Schbath approach is not possible because these statistics could not be computed in a feasible amount of time. The comparison with the simulated p-values shows that the new approach fits very well. In contrast, the binomial/Poisson approximations show the typical significant deviation we have already seen for the artificial PFMs. Hence, in such a realistic framework, the new approach is the only possiblity to compute the count statistic without simulations.

3.4. Characteristic values

Table 1 shows the characteristic values for each PFM. The "nothing" PFM has a low first eigenvalue while the second eigenvalue is equal to zero. Since the PFM has no self-overlap, these two characteristic values confirm the analysis given in the method section. For the "palindrome" PFM the equation $\lambda_1 \approx -\lambda_2$ holds because the only self-overlap is given by the the palindromic property of the PFM. The "repeat" PFM has a much higher first eigenvalue than the "nothing" PFM since it has a strong repeat-structure. Since there is no palindromic feature within the PFM, we obtain $\lambda_2 = 0$. In constrast, the "repeatpalindrome" PFM contains both self-overlaps, thus, $\lambda_2 < 0$ but $\lambda_1 \neq -\lambda_2$. Finally, the PFM "M00950" has also a clear palindromic self-overlap. All these observations are confirmed by the sequence logos (Fig. 6) and the resulting count statistics (Fig. 7). Thus, the characteristic values describe the self-overlapping features well. In the given cases, the self-overlap is clear for illustration purposes but in more difficult cases they shed light on the self-overlapping structure.

TABLE 1. THE TWO CHARACTERISTIC VALUES GIVEN BY THE EIGENVALUES OF MATRIX A FOR EACH PFM

	λ_1	λ_2	*Comment*
Nothing	0.0123	0.0000	λ_1 small, $\lambda_2 = 0$
Palindrome	0.0016	−0.0016	λ_1 small, $\lambda_2 = -\lambda_1$
Repeat	0.2599	0.0000	λ_1 large, $\lambda_2 = 0$
Repeatpalindrome	0.1792	−0.1526	λ_1 large, $\lambda_2 \approx -\lambda_1$
M00950	0.0455	−0.0435	$\lambda_2 \approx -\lambda_1$

4. DISCUSSION

We have proposed a new approximation for the count statistic of PFMs. In contrast to most previous works, we incorporate the complementary strand, which introduces further dependencies of overlapping hits. Due to explicit modeling of these dependencies, as well as dependencies between overlapping hits on the same strand, we are able to compute precise p-values for any PFM. Furthermore, we have shown how to compute two characteristic values describing the tendency of overlaps and palindromic hits of a given PFM with the same algorithm. The time complexity neither depends on the sequence length nor on the number of compatible words. Therefore, the algorithm is very efficient. It might be further improved using the Fourier transform with the convolution theorem (Press et al., 1992; Keich, 2005).

Comparison with other approaches shows that our approach has highest accuracy. Furthermore, most of the competing approaches enumerate all compatible words $\mathcal{W}$. Since $|\mathcal{W}|$ grows exponentially with the length of the PFM the overall-running time is exponential. Hence, the normal approximation (Waterman, 2000), the Chen-Stein approach (Reinert and Schbath, 1999), as well as the Roquain and Schbath (2007) approach, and the exact approach (Zhang et al., 2007) cannot generally be applied in practice (see Supplementary Material for a comparison of the running time at *http://mosta.molgen.mpg.de*).

A major drawback of the new approach is the restricted background model. So far, we only use a symmetric i.i.d. model defined by the GC content. Therefore, the statistics are symmetric between both strands. Further studies will show whether more complicated background models which conserve the symmetry can be used as well.

The count statistic can be used for the upstream annotation of genes as well as mapping of PFMs to certain genome regions. An extension of the approach to two or more PFMs yield many more applications like p-values for co-occurrences and similarity between PFMs (Pape et al., 2008).

ACKNOWLEDGMENTS

The authors thank Hugues Richard for discussion and advice, Stéphane Robin and the anonymous reviewer for their remarks, and Till A. Pape for support in designing the figures.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Alberts, B., Johnson, A., Lewis, J., et al. 2002. *Molecular Biology of the Cell*. Garland Science, New York.

Barbour, A.D., Holst, L., and Janson, S. 1992. *Poisson Approximation*. Oxford University Press, New York.

Beckstette, M., Homann, R., Giegerich, R., et al. 2006. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinform.* 7, 389.

Bejerano, G., Friedman, N., and Tishby, N. 2004. Efficient exact p-value computation for small sample, sparse, and surprising categorical data. *J. Comput. Biol.* 11, 867–886.

Brendel, V., Beckmann, J., and Trifonov, E. 1986. Linguistics of nucleotide sequences: morphology and comparison of vocabularies. *J. Biomol. Struct. Dyn.* 4, 11–21.

Chryssaphinou, O., and Papastavridis, S. 1988. A limit-theorem for the number of nonoverlapping occurences of a pattern in a sequence of independent trials. *J. Appl. Probab.* 25, 428–431.

Claverie, J.-M., and Audic, S. 1996. The statistical significance of nucleotide position-weight matrix matches. *Comput. Appl. Biosci.* 12, 431–439.

Crooks, G.E., Hon, G., Chandonia, J.-M., et al. 2004. Weblogo: a sequence logo generator. *Genome Res.* 14, 1188–1190.

Denise, A., Régnier, M., and Vandenbogaert, M. 2001. Assessing the statistical significance of overrepresented oligonucleotides. *Proc. WABI '01*, 85–97.

Fu, J., and Koutras, M. 1994. Distribution theory of runs: a Markov chain approach. *J. Am. Stat. Assoc.* 89, 1050–1058.

Gentleman, J.F., and Mullin, R.C. 1989. The distribution of the frequency of occurrence of nucleotide subsequences, based on their overlap capability. *Biometrics* 45, 35–52.

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

Guibas, L.J., and Odlyzko, A.M. 1981. String overlaps, pattern matching, and nontransitive games. *J. Comb. Theory Ser. A* 30, 183–208.

Hertzberg, L., Zuk, O., Getz, G., et al. 2005. Finding motifs in promoter regions. *J. Comput. Biol.* 12, 314–330.

Johnson, N.J., Kotz, S., and Kemp, A.W. 1995. *Univariate Discrete Distributions*. Wiley, New York.

Keich, U. 2005. sFFT: a faster accurate computation of the *p*-value of the entropy score. *J. Comput. Biol.* 12, 416–430.

Kemp, C.D. 1967. "Stuttering-Poisson" distributions. *J. Statist. Social Enquiry Society Ireland* 21, 151–157.

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

Leung, M., Marsh, G., and Speed, T. 1996. Over- and underrepresentation of short DNA words in herpesvirus genomes. *J. Comput. Biol.* 3, 345–360.

Matys, V., Fricke, E., Geffers, R., et al. 2003. TRANSFAC(R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* 31, 374–378.

Pape, U.J., Grossmann, S., Hammer, S., et al. 2006. A new statistical model to select target sequences bound by transcription factors. *Genome Inform.* 17, 134–140.

Pape, U.J., Rahmann, S., and Vingron, M. 2008. Natural similarity measures between position frequency matrices with an application to clustering. *Bioinformatics* 24, 350–357.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., et al. 1992. *Numerical Recipes in C. The Art of Scientific Computing*. Cambridge University Press, Cambridge, UK.

Prum, B., Rodolphe, F., and de Turckheim, E. 1995. Finding words with unexpected frequencies in deoxyribonucleic acid sequences. *J. R. Statist. Soc. Ser. B Methodol.* 57, 205–220.

Rahmann, S. 2003. Dynamic programming algorithms for two statistical problems in computational biology. *Proc. 3rd WABI*, 151–164.

Rahmann, S., Müller, T., and Vingron, M. 2003. On the power of profiles for transcription factor binding site detection. *Statist. Appl. Genet. Mol. B* 2.

Régnier, M. 2001. A unified approach to word occurrence probabilities. *Discrete Appl. Math.* 104, 259–280.

Reinert, G., and Schbath, S. 1999. Compound Poisson approximations for occurrences of multiple words, 257–275. In: Seiller-Moiseiwitsch, F., ed. *Statistics in Molecular Biology and Genetics*. IMS Lecture Notes, New York.

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

Robin, S. 2002. A compound Poisson model for word occurrences in DNA sequences. *J. R. Statist. Soc. C Appl.* 51, 437–451.

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.

Roquain, E., and Schbath, S. 2007. Improved compound Poisson approximation for the number of occurrences of multiple words in a stationary Markov chain. *Adv. Appl. Probab.* 39, 128–140.

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

Staden, R. 1989. Methods for calculating the probabilities of finding patterns in sequences. *Comput. Appl. Biosci.* 5, 89–96.

Stormo, G.D., Schneider, T.D., Gold, L., et al. 1982. Use of the "Perceptron" algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res.* 10, 2997–3012.

Tang, W., and Perry, S.E. 2003. Binding site selection for the plant MADS domain protein AGL15: an *in vitro* and *in vivo* study. *J. Biol. Chem.* 278, 28154–28159.

Waterman, M.S. 2000. Probability and Statistics for Sequence Patterns. In: *Introduction to Computational Biology*. Chapman & Hall/CRC, New York.

Wu, T.D., Nevill-Manning, C.G., and Brutlag, D.L. 2000. Fast probabilistic analysis of sequence function using scoring matrices. *Bioinformatics* 16, 233–244.

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

Address reprint requests to:
Utz J. Pape
Molecular Genetics
Ihnestr. 73
14195 Berlin, Germany

E-mail: utz.pape@molgen.mpg.de