Statistical detection of co-operative transcription factors with similarity adjustment

Utz J. Pape\textsuperscript{1,2} and Holger Klein\textsuperscript{1} and Martin Vingron\textsuperscript{1}

\textsuperscript{1} Dept. of Computational Biology, Max Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany, (e-mail: utz.pape@molgen.mpg.de)
\textsuperscript{2} Dept. of Mathematics and Computer Science, Free University of Berlin, Takustr. 9, 14195 Berlin, Germany.

Abstract: Statistical assessment of cis-regulatory modules (CRMs) is a crucial task in computational biology. Usually, one concludes from exceptional co-occurrences of DNA motifs that the corresponding transcription factors are co-operative. However, similar DNA motifs tend to co-occur in random sequences due to high probability of overlapping occurrences. Therefore, it is important to consider similarity of DNA motifs in the statistical assessment. Based on previous work, we propose to adjust the window size for co-occurrence detection. Using the derived approximation, one obtains different window sizes for different sets of DNA motifs depending on their similarities. This ensures that the probability of co-occurrences in random sequences are equal. Applying the approach to selected similar and dissimilar DNA motifs from human transcription factors shows the necessity of adjustment and confirms the accuracy of the approximation.

Our previously published statistics can only deal with non-overlapping windows. Therefore, we extend the approach and derive Chen-Stein error bounds for the approximation. Comparing the error bounds for similar and dissimilar DNA motifs shows that the approximation for similar DNA motifs yields large bounds. Hence, one has to be careful using overlapping windows. Based on the error bounds, one can pre-compute the approximation errors and select an appropriate overlap-scheme before running the analysis. Software and source code are available at http://mosta.molgen.mpg.de.

1 Introduction

An important goal in computational biology is to decipher the transcriptional regulation of genes. Interaction of nearby transcription factors (TFs) initiate or inhibit transcription of a gene [Fic96, YBD98]. They bind mainly upstream of genes to DNA by recognizing TF-specific sequences which can be summarized to a DNA motif. The set of DNA motif occurrences upstream of a gene is called a cis regulatory module (CRM, [BNP\textsuperscript{+}02]). A CRM is a sequence region with dense clusters of DNA motif occurrences as demonstrated experimentally [CMW\textsuperscript{+}03, HGL\textsuperscript{+}04] and computationally [Wag99, LMNP03]. TFs, which combinatorially regulate genes, are called co-operative. Such TFs are assumed to have exceptionally many DNA motif occurrences approximate to each other. Thus, a significant number of co-occurrences of the corresponding DNA motifs can be used to assess the strength of co-operativity.
CRMs can be detected using *ab initio* discovery of new (e.g. [ZW04, GL05]) or based on known DNA motifs. We assume that the DNA motifs are known. Many approaches have been proposed integrating data of different kind for improving CRM prediction [PSC01, YLZQ06]. Since the main characteristic of CRMs is their high local density of DNA motif occurrences, one essential data source is always the DNA sequence annotated with DNA motif occurrences. Here, we focus on DNA motifs represented by position frequency matrices (PFMs) [Sto00]. Other approaches compute the co-operative binding energy of multiple sites of TFs [GS01, FFY04] using thermo-dynamical models.

![Figure 1: Two different approaches to detect CRMs](image)

Figure 1: Two different approaches to detect CRMs: Upper panel illustrates approaches which are based on short distances between DNA motif occurrences. Lower panel visualizes detection of CRM considering occurrences in windows.

Based on the PFM representation, [Guh06] classifies the approaches to find CRMs into hidden Markov models [CRB97, FHW01] and occurrence-based approaches. We further divide the occurrence-based approaches into two categories, (i) relying on small distances between DNA motif occurrences [WF98, Wag99] and (ii) based on co-occurrences of DNA motifs in a small window [BNP+02, HL02, FSHW02, KV07]. The method to compute statistical significance is a difficult problem [Kri04] and can be solved by (i) assuming position independence of occurrences [WF98, Wag99, FSHW02] or (ii) employing randomizations [HL02, BHVvr07] or (iii) exact calculation [BCR+07].

The position independence of binding site occurrences is strongly violated for (self-)similar PFMs [Wag99, PRSV08]. The significance calculation based on randomization also encounters problems for similar PFMs, hence, they are usually removed from the analysis [HL02]. In addition, incorporating the complementary strand, introduces further dependencies and worsen the results. The exact calculation [BCR+07] based on a Aho-Corasick automaton has high computational complexity such that solutions for longer PFMs are hard to obtain. Furthermore, the approach does not use the complementary strand.

In [PV08], we propose a fast and accurate approximation for the significance calculation of CRMs circumventing the position independence assumption, incorporating similarity between PFMs, and including the complementary strand. There, we define a CRM to be a sequence region, which we called window, of defined length where all DNA motifs of a given set have at least one occurrence. This is called the co-occurrence event. To get statistically significant CRMs, the length of the window has to be small such that the co-occurrence event is unlikely to happen by chance. We compute the probability of a CRM which is the probability of the co-occurrence event in a random sequence given a window length. Considering the overlap probabilities between the occurrences of the TF binding sites, we capture the (self-)similarities of the PFMs and most of the dependencies introduced by the complementary strand.
In this article, we extend the approach such that one can compute the length of the window for a specific set of DNA motifs by defining the probability of the co-occurrence event as parameter. We focus on pairs of DNA motifs. Intuitively, the results show that for similar PFMs the length of the window is smaller than for dissimilar PFMs given the same probability. Due to this computation, one can adjust the window size based on the similarity of the PFMs. Hence, by using different window sizes for sets of PFMs sharing different amounts of similarity between their PFMs, one can obtain equal co-occurrence probabilities for all sets. Therefore, follow-up analyses do not have to consider the similarity between PFMs anymore. Otherwise, similar PFMs would yield more co-occurrence events than dissimilar PFMs just due to their similarity. This would generally bias statistics based on the number of co-occurrence events. Hence, window size adjustment by considering the similarity of PFMs is necessary.

Figure 2: Proposed algorithm to compute co-operativity of a pair of TFs: First, divide sequence into windows. Second, count windows containing at least one hit of each TF. Compute corresponding count distribution under random sequence model to obtain $p$-value for co-operativity. Furthermore, one is interested in whether specific TFs are generally involved in the same CRMs. We call this co-operativity of TFs. In [PV08], we also show how to compute the significance of co-operativity. The sequence is divided into equal-sized non-overlapping windows covering the whole sequence. We compute a $p$-value for the number of observed CRMs (windows with the co-occurrence event) since we can derive the count distribution of CRMs. In case of non-overlapping windows the count distribution is exact besides the approximations in the calculation of the co-occurrence event. The accuracy of the approximation is shown by comparison with a simulation study [PV08]. In contrast, overlapping windows introduce further dependencies. Therefore, we show in this article how to compute error bounds using the Chen-Stein method. Applying these error bounds to selected sets of PFMs show that similar PFMs retrieve high approximation errors due to stronger dependencies between overlapping windows.

In the next section, we derive formulas for the window length and explicitly state the Chen-Stein error bounds. Furthermore, we describe the data set of human TFs and how the PFMs are selected. The Section Results applies the formulas for window length and the Chen-Stein error bounds to a selected pairs of TFs.
2 Methods

We assume that each TF is given by a PFM. For each position \( j \) of a sequence, we have an indicator random variable \( Y_j(A) \) which is 1 if the summed score at this position reaches the threshold. We denote the random variables for the complementary strand by a prime, e.g. \( Y_j'(A) \). The threshold can be controlled by the type I error \( \alpha := P(Y_j(A) = 1) = P(Y_j'(A) = 1) \) in a random sequence. The model for the random sequence is assumed to be an i.i.d. sequence defined by the GC content. We assume this simple background model, since we require the distribution of hits on both strands to be equal.

As stated before, a CRM is a window of given length \( w \) with at least one hit for TF \( A \) and one hit of TF \( B \). We split up the calculation of this co-occurrence event into three parts: Let \( N_w(A) = \sum_{j=1}^{w}(Y_j(A) + Y_j'(A)) \) denote the random variable for the number of hits of TF \( A \) in a random sequence of length \( w \) where we allow hits overlapping the boundary of the window. Now, we can state the probability \( p(w) \) of a CRM in a given window of length \( w \) by \( p(w) := P(N_w(A) > 0, N_w(B) > 0) \). Calculation using the inclusion-exclusion formula and transformations as described in [PV08] yields for the probability of the co-occurrence event \( p(w) \approx 1 - e^{-r_A \cdot w} - e^{-r_B \cdot w} + e^{-r_{AB} \cdot w} \) where \( r_A \) resp. \( r_B \) correspond to rates for the occurrence of TF \( A \) resp. \( B \) considering overlaps.

2.1 Calculate Window Size

In practice, the probability for the co-occurrence event is given as parameter and the window size has to be computed. In this case, we have to find the roots of

\[
1 - \exp(-r_A \cdot w) - \exp(-r_B \cdot w) + \exp(-r_{AB} \cdot w) - p. \tag{1}
\]

Using the Newton approach, we obtain following recursion starting from a chosen initial value \( w_0 \):

\[
w_{i+1} = w_i - \frac{1 - \exp(-r_A \cdot w_i) - \exp(-r_B \cdot w_i) + \exp(-r_{AB} \cdot w_i) - p}{r_A \exp(-r_A \cdot w_i) + r_B \exp(-r_B \cdot w_i) - r_{AB} \exp(-r_{AB} \cdot w_i)}. \tag{2}
\]

In case one requires a closed formula, one can also apply a Taylor expansion to the formula for the co-occurrence probability. E.g., the formula for a 2nd order expansion which already gives accurate results for small \( p \) is given with \( a = r_{AB} - r_A - r_B \) and \( b = r_A^2 - r_A^2 - r_B^2 \) by

\[
w(p) = \frac{a}{b} + \sqrt{\left(\frac{a}{b}\right)^2 + \frac{2p}{b}}. \tag{3}
\]
2.2 \textit{p}-value for Co-operativity

Previously, we showed how to compute the co-occurrence probability \( p(w) \) in a given window. To compute co-operativity, we suggest to decompose the sequence into non-overlapping windows of equal size and count the number \( x \) of CRMs (windows with the co-occurrence event). We define for each window \( i \) a Bernoulli random variable \( W_i \) which is 1 if the corresponding window contains a co-occurrence event and otherwise 0. Denoting the number of windows by \( m = n/w \) with sequence length equal to \( n \), we define \( W := \sum_{i=1}^{m} W_i \). The number \( W \) of windows with co-occurrence events is distributed as Poisson \( \mathcal{P}(\vartheta) \) with \( \vartheta = p(w) \cdot m \) if \( p(w) \to 0 \) and \( m \to \infty \).

2.3 Bounds for Overlapping Windows

Considering overlapping windows necessitates the step size \( s \) as parameter. The number \( m \) of windows becomes \( m = n/s - w + 1 \). We assume that \( n, s, w \) are chosen such that \( m, n, s, w \) are positive integers and \( s < w < \frac{1}{2}n \). Obviously, overlapping windows are dependent on each other. In this case, we can still use a Binomial or Poisson distribution but the dependencies lead to an error in the approximation. Using the Chen-Stein method [Che75], the error can be quantified. The quantification is done in terms of the total variation distance. Let \( U \) and \( V \) be any two random processes with values in the same space \( E \), then the total variation distance between their distributions (denoted by \( \mathcal{L}(\cdot) \)) is

\[
d_{TV}(\mathcal{L}(U), \mathcal{L}(V)) = \sup_{D \subseteq E} |P(U \in D) - P(V \in D)|
\]

(4)

where \( D \) is assumed to be measurable. Here, we focus on the Poisson Approximation since it obtains slightly better error bounds. Thus, we calculate the bound for \( d_{TV}(\mathcal{L}(W), \mathcal{P}(\vartheta)) \). Let denote \( I := \{i : 0 < i \leq m\} \) the index set of the Bernoulli variables. The main idea is to define for each Bernoulli variable \( W_i \) a neighborhood set \( B_i \subseteq I \) of random variables which have strong dependencies with \( W_i \). We also require \( i \in B_i \). In our case, there are only local dependencies since only overlapping windows are dependent on each other. Therefore, we capture all dependencies in the sets \( B_i \) which means that for each window \( i \) the set \( B_i \) contains the index \( i \) and the indices of overlapping windows to the left and to the right. Hence, we obtain the bound derived from Theorem 1 in [AGG90] using an improved bound [BHJ92] \( d_{TV}(\mathcal{L}(W), \mathcal{P}(\vartheta)) \leq \vartheta^{-1}(1 - e^{-\vartheta})(b_1 + b_2) \) with

\[
b_1 := \sum_{i \in I} \sum_{j \in B_i} E[W_i] \cdot E[W_j], \quad b_2 := \sum_{i \in I} \sum_{j \in B_i, j \neq i} E[W_i \cdot W_j].
\]

(5)

The bound \( b_1 \) is straightforward to compute as it only contains the first moments. We have to consider the fact that the \( B_i \)s for the first and last few windows contain less dependent variables than windows in the middle of the sequence. Let \( r = w/s \), then for example, the first window has \( r - 1 \) overlapping windows, thus, \( |B_1| = r \) since we also include
Figure 3: The four disjoint events for two windows where the dark grey area indicates the overlap. Regions containing an $A$ or $B$ must necessarily contain at least one hit of the corresponding TF while $\bar{A}$ and $\bar{B}$ label regions where the respective TF must not occur. In blank regions, any TF and combinations of TFs might be present.

index 1 in the set. The second window additionally overlaps with the first window, thus, $|B_2| = |B_1| + 1$. The set size is incremented by 1 until the $r+1$th window as this window has equal number of overlaps to the left and to the right. At the end of the sequence, the set size is decremented in the same way. Hence, we obtain $b_1 = p(w)^2 (r(1 - r + 2m) - m)$.

The second bound $b_2$ is more complicated to calculate because it contains the second moments. Since we consider Bernoulli variables, the second moment is the probability that both variables are equal to one: $E[W_iW_{i+k}] = P(W_i = 1, W_{i+k} = 1)$. Considering only two TFs $A$ and $B$, we can write this probability in terms of the count random variables by decomposing it into four disjoint events as illustrated in Fig. 3.

Denoting the size of each non-overlapping part by $d = k \cdot s$ while the overlapping part has a length of $v = w - d$, we obtain for the second moment:

$$E[W_iW_{i+k}] = p(v) + (1 - e^{-drA})^2 \left[ 1 - e^{-vrB} - p(v) \right]$$

$$+ (1 - e^{-drB})^2 \left[ 1 - e^{-vra} - p(v) \right] + p(d)^2 e^{-vAB}. \quad (6)$$

To compute the bound, we observe that $E[W_iW_{i+k}]$ is independent of $i$ since all $W_i$ are identically distributed and have the same pairwise dependencies. Therefore, we clarify notation by defining $\zeta_k := E[W_iW_{i+k}]$. For the same reason, we also obtain $\zeta_k = E[W_iW_{i-k}]$. Using the further definition of $\zeta = \sum_{k=1}^{r-1} \zeta_k$, we yield for bound $b_2$ applying the same logic as above:

$$b_2 = 2 \cdot \sum_{i=1}^{r} \left[ \zeta + \sum_{k=1}^{i-1} \zeta_k \right] + 2(m - 2r)\zeta = 2 \left( m\zeta - r\zeta + \sum_{i=1}^{r} \sum_{k=1}^{i-1} \zeta_k \right). \quad (8)$$

Here, we assume that the empty sum ($\sum_{k=1}^{i-1} \zeta_k$ for $i = 1$) is equal to 0.
2.4 Data

The PFM set used here is the vertebrate_non_redundant_minFP set from the TRANSFAC database (v. 11.3) [MFG03]. Since despite the name the set contains more than one PFM per transcription factor (214 in total), we only select the first PFM per TF and obtain a set of 142 PFMs. Hence, we are left with a set of one PFM per TF. However, the remaining similarities between PFMs in this set are not negligible. To show this, we measure the similarity between all pairs of PFMs by the limiting covariance [PRV08]. Then, we select the pair of PFMs with highest similarity (0.0002): S8 (V$S8_01$) and CHX10 (V$CHX10_01$). We use this pair for our analysis. To assess the influence of similarity, we also select a very dissimilar pair of PFMs. Given S8, the most dissimilar PFM is HIC (V$HIC1_02$) with a similarity of $-0.000004$. Hence, we define a pair of similar PFMs S8 and CHX10 and a pair of dissimilar PFMs S8 and HIC.

All analyses regarding PFMs are performed based on a balanced type-I error ($\alpha$) in a sequence of length 500 controlled at a level of 1% (see [PGH06] for details). In a step called regularization, we add pseudo-counts to the position specific distributions of the PFM according to the information content of the position [Rah03]. Simulated sequences are generated i.i.d. with 50% GC content.

3 Results and Discussion

First, we apply the formulas for the window size given a co-occurrence probability of $p = 0.01$ to both pairs of PFMs. The pair of similar PFMs S8 and CHX10 yields a window size of 54bp for both Newton iteration and Taylor expansion. Computing the co-occurrence probability for the window size 54bp yields exactly 0.01. Hence, both approximations are very accurate. The dissimilar pair S8 and HIC yields for the same given co-occurrence probability a window size of 297bp using Newton iteration and 281bp using Taylor expansion. The corresponding co-occurrence probabilities are 0.01 and 0.009. Hence, the Newton iteration is slightly more accurate than the Taylor expansion. In comparison to the similar pair, one yields a 5-fold larger window size. Since similar PFMs tend to have overlapping hits, their probability of co-occurrence which includes overlapping hits is high. Therefore, an occurrence of one PFM increases the probability of an occurrence of the other PFM. In contrast, dissimilar PFMs cannot overlap. Thus, presence of one PFM decreases the probability of an (overlapping) occurrence of the other PFM. Due to the big difference in the window sizes, it is very important to consider the similarity between PFMs. The presented approach shows one can simply adjust the window size. Hence, one would use a window size of 54bp for the similar pair and of 297bp for the dissimilar pair. Then, both pairs have equal co-occurrence probabilities.

We verify this prediction by a simulation study. After annotating 100 random sequences each of length 1,000,000 by the corresponding PFMs, we count the number of co-occurrence events given above window sizes. The histograms for both pairs are shown in Fig. 4. The left panel contains the histogram for the similar pair. The distribution has a mean of 0.007
and a standard deviation of 0.0006. Hence, the approximated co-occurrence probability of 0.01 is slightly biased to lower probabilities. The reason is that the approximation of the co-occurrence probability only considers first-order dependencies between occurrences. This means overlaps between more than two occurrences are ignored. The right panel of Fig. 4 shows the histogram for the dissimilar pair. The mean is 0.012 with standard deviation 0.002. Thus, the approximated probability is still within one standard deviation of the mean. Since the corresponding PFMs do not overlap, the first-order approximation yields more accurate results. In contrast, applying the same window size (e.g. 297bp) to both pairs would yield a co-occurrence probability of around 0.04 (retrieved by simulation) for the similar pair. Hence, the difference between co-occurrence probabilities decreases from almost 3–4-fold to quite comparable co-occurrence probabilities by adjusting the window size.

Based on the selected window sizes, one can compute Chen-Stein error bounds for the co-operativity $p$-value approximation. Using windows which overlap by 10% yields an error bound of 0.04 for the similar pair S8 and CHX10 on a sequence of length 1000bp. Hence, it will be difficult to obtain significant results since one cannot obtain $p$-values less than 0.04. In general, similar PFMs have a high approximation error for overlapping windows since overlapping occurrences induce high dependencies between two windows. In contrast, the dissimilar pair S8 and HIC has an error bound of 0.002. The bound is much smaller for two reasons: First, the window is much larger, thus, less windows are used for the sequence. Second, overlapping windows are less dependent due to small probabilities for the overlap of two occurrences. Hence, in case of dissimilar PFMs one can use overlapping windows and still obtain significant co-operativity.

In conclusion, we can state that detection of significant co-occurrences and co-operativity based on PFM occurrences is a difficult problem due to strong dependencies induced by similarity between PFMs. We show a reasonable approximation to adjust the window size such that co-occurrence and co-operativity probabilities are comparable between similar and dissimilar PFMs. Therefore, statistical follow-up analyses can ignore the similarity issue. In addition, we propose a new approximation for co-operativity using overlapping windows. Using the Chen-Stein technique, we can bound the approximation error. Results show that similar PFMs imply strong dependencies between overlapping windows. This leads to high approximation errors. In contrast, dissimilar PFMs yield low approxi-
mation errors. Based on our error bounds, one can pre-compute the approximation errors and select an appropriate overlap-scheme before running the analysis. In general, the approach can be extended to deal with sets of TFs. Furthermore, a more general sequence background model would be eligible.

References


Pape and Vingron. Statistics for Co-Occurrence of DNA Motifs. Accepted by IWAP 2008.


