Sequence analysis

FastMotif: spectral sequence motif discovery

Nicoló Colombo¹ and Nikos Vlassis²,*

¹Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg and ²Adobe Research, San Jose, CA, USA

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on September 18, 2014; revised on March 20, 2015; accepted on April 9, 2015

Abstract

Motivation: Sequence discovery tools play a central role in several fields of computational biology. In the framework of Transcription Factor binding studies, most of the existing motif finding algorithms are computationally demanding, and they may not be able to support the increasingly large datasets produced by modern high-throughput sequencing technologies.

Results: We present FastMotif, a new motif discovery algorithm that is built on a recent machine learning technique referred to as Method of Moments. Based on spectral decompositions, our method is robust to model misspecifications and is not prone to locally optimal solutions. We obtain an algorithm that is extremely fast and designed for the analysis of big sequencing data. On HT-Selex data, FastMotif extracts motif profiles that match those computed by various state-of-the-art algorithms, but one order of magnitude faster. We provide a theoretical and numerical analysis of the algorithm’s robustness and discuss its sensitivity with respect to the free parameters.


Contact: vlassis@adobe.com

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

In the last decades, due to the advent of new sequencing technologies, motif discovery algorithms have become an essential tool in many computational biology fields. In cell biology, sequence motif discovery plays a primary role in the understanding of gene expression through the analysis of sequencing data and the identification of DNA-transcription factor binding sites (Annala et al., 2011; Berger et al., 2006; Cheng et al., 2013; Wei et al., 2010; Zhang et al., 2013; Zhao et al., 2012).

Various experimental techniques are nowadays available to extract DNA-protein binding sites in vivo [ChIP-Seq (Johnson et al., 2007)] and in vitro [protein binding microarray (PBM) (Berger and Bulyk, 2009), HT-Selex (Jolma et al., 2010, 2013; Kinzler and Vogelstein, 1990; Tuerk and Gold, 1990)]. Thanks to the quantity and quality of the generated data, HT-Selex is considered one of the most promising high-throughput techniques for studying transcription factor binding affinity in vitro [see the recent work (Orenstein and Shamir, 2014) for a quantitative comparison between HT-Selex and other high-throughput techniques as ChIP-Seq and PBM]. In the HT-Selex protocol, tens of thousands enriched DNA fragments are obtained through a series of incubation/selection cycles. In each cycle, an initial pool containing randomized ligands of length 14–40 bp is incubated with an immobilized DNA-binding protein. Bound ligands are amplified by polymerase chain reaction steps (PCR), sequenced and then used as initial pool for a next cycle, until the pool is saturated (Jolma et al., 2010, 2013; Kinzler and Vogelstein, 1990; Tuerk and Gold, 1990; Zhao et al., 2009).

Due to the high but not exact specificity of transcription factor binding affinities, enriched DNA fragments in a dataset typically contain similar but not exactly conserved instances of the same binding motif. This calls for algorithms that can extract simple and intuitive binding motifs that are robust to such stochasticity (Berger and Bulyk, 2009; Jolma et al., 2010; Zhao et al., 2009). In the simplest case, the binding preferences are modelled by means of consensus sequences, obtained by selecting a few deterministic character strings that are over-represented in the dataset. A more flexible
representation is provided by Position Weight Matrices (PWM) that describe binding sites as probability distributions over the DNA alphabet. Based on the simplifying assumption that the total binding energy is a site-by-site sum of single protein-nucleobase interactions, PWM's are only approximate models of the true transcription factor preferences. A debate is still open on whether such an approximation gives a satisfactory picture of the DNA-protein interaction or if it is a too simplified reduction of the real biological process (Badis et al., 2009). More sophisticated models, that go beyond the PWM representation by taking into account multi-base probability distributions or long-distance interactions, have been proposed and tested in the literature (Badis et al., 2009; Bulyk, 2002; Chen et al., 2007; Mathelier and Wasserman, 2013; Santolini et al., 2013). However, in most cases, these improvements did not bring about cogent evidence against the simpler and more intuitive approach based on position independent distributions (Zhao and Stormo, 2011).

In statistics and machine learning, factorized (aka product) distributions like PWMs and their linear combinations (aka mixtures) are commonly used in modelling empirical distributions from various kinds of data, and an important problem is how to estimate such models from data (Lindsay, 1995; Titterington, 1985). In pioneering work, Chang (1996) showed that it is possible to infer a mixture of product distributions via the spectral decomposition of ‘observable’ matrices, i.e. matrices that can be estimated directly from the data using suitable combinations of the empirical joint probability distributions (Chang, 1996). Extensions and improvements of this idea have been developed more recently in a series of remarkable works, where the spectral approach is applied to a larger class of probability distributions, and robust versions of the original method have been analysed theoretically (Anandkumar et al., 2012a, c; Hsu et al., 2012; Mossel and Roch, 2006).

In this article, we further develop the original spectral technique of Chang (1996) and study its application to the problem of learning probabilistic motif profiles from noisy sequencing data. The key observation is that, under the PWM approximation, motif discovery reduces to the more general problem of learning a mixture of product distributions, and hence, it is possible to extract motif profiles from sequencing data using usual spectral decompositions.

We present FastMotif, a new motif finding algorithm that is faster than other sequence discovery tools and is designed for processing noisy high-throughput datasets. Based on a new and more stable version of the spectral techniques introduced in (Anandkumar et al., 2012a; Corless et al., 1997; Mossel and Roch, 2006), our algorithm is robust to model misspecification, is not prone to local optima, and it can be adapted to searching for motifs of arbitrary length. In addition, the method is completely general and, upon minor modifications, can be used for sequence discovery over any sequence alphabet and for analyzing datasets with binding affinity scores (Berger and Bulyk, 2009; Johnson et al., 2007). Throughout this work, we assume that transcription factor specificities are well described by product distributions, i.e. PWM's, and leave for future work the spectral inference of more advanced motif representations.

Finally, a comment on the claimed optimality of FastMotif. It has been shown [see for example Anandkumar et al. (2012c)] that in the limit of infinitely many data and under no model misspecification, a spectral method is statistically consistent, that is, it always recovers the true underlying model. FastMotif, being a spectral method, inherits this optimality property by the way of the uniqueness of matrix eigen decomposition. This is in marked contrast to algorithms like Expectation Maximization (EM) that can easily get trapped in poor local optima even under favourable sample conditions. In practice, we expect the output of FastMotif to approach optimality when the true binding model approaches a PWM and the size of the training dataset is big enough, as for example in the case of HT-SELEX data.

The article is organized as follows: In Section 2, we give a brief overview of related work; in Section 3, we describe FastMotif and its application to sequencing data; and in Section 4, we show our results. More mathematical details about spectral approaches in general and FastMotif in particular can be found in the Supplementary Material. The Matlab code of FastMotif is available from http://kbs-portal.uni.lu/bioinformatics.

2 Related work

2.1 Motif finding

The literature on sequence motif discovery is vast. We refer to (Das and Dai, 2007; Sandve et al., 2007; Simcha et al., 2012; Tompa et al., 2005) for reviews and additional references. There are two main classes of motif finding algorithms, probabilistic and word-based. Probabilistic algorithms search for the most represented ungapped alignments in the sample to obtain determinstic consensus sequences, PWM models, or more advanced models that take into account multi-base correlations (Bulyk, 2002; Badis et al., 2009; Chen et al., 2007; Mathelier and Wasserman, 2013; Santolini et al., 2013). Word-based algorithms search the dataset for deterministic short words, measure the statistical significance of small variations from a given seed, or transform motif discovery into a kernel feature classification problem (Lee et al., 2011; Leslie et al., 2002; Vert et al., 2005). Our method and well known motif discovery algorithms as MEME (Bailey and Elkan, 1994) and STEME (Reid and Wernisch, 2011), belong to the probabilistic class, while the two algorithms we have used for evaluating our results, namely the method used in (Jolma et al., 2013) and DREME (Bailey, 2011) are word-based algorithms. The latter algorithms can also compute PWM models, so it is of interest to compare algorithms of different classes (Section 4).

2.2 Spectral methods

Spectral methods have been applied as an alternative to the EM algorithm (Dempster et al., 1977) for inferring various kinds of probability distributions, such as mixtures of product distributions, Gaussian mixtures, Hidden Markov models, and others (Anandkumar et al., 2012a, b, h; Boots et al., 2011; Chang, 1996; Hsu et al., 2012; Mossel and Roch, 2006) [see (Baile et al., 2014) for a recent review]. These methods are not as flexible as the EM algorithm, but they are not prone to local optima and have polynomial computational time and sample complexity. Various spectral decomposition techniques have been proposed: Chang's spectral technique (Chang, 1996; Mossel and Roch, 2006), a symmetric tensor decomposition method (Anandkumar et al., 2012a), and an indirect learning method for inferring the parameter of Hidden Markov Models (Hsu et al., 2012). The practical implementation of the spectral idea is a non-trivial task because the stability of spectral decomposition strongly depends on the spacing between the eigenvalues of the empirical matrices. In (Anandkumar et al., 2012a; Mossel and Roch, 2006) certain eigenvalue separation guarantees for Chang's spectral technique are obtained via the contraction of the higher (order three) moments to Gaussian random vectors. In the tensor approach presented in (Anandkumar et al., 2012a), the non-negativity of the eigenvectors is ensured by using a deflating power method that generalizes usual deflation techniques for matrix diagonalization to the case of symmetric tensors of order three. A third possibility involves replacing the random vector of Chang's
spectral technique with an ‘anchor observation’ that, for each hidden state, ‘tends to appear in the state much more often than in the other states’ (Song and Chen, 2014) and guarantees the presence of at least one well separated eigenvalue (Arora et al., 2012; Song and Chen, 2014). Finally, as briefly mentioned in (Anandkumar et al., 2012a; Hsu and Kakade, 2013), the stability of Chang’s technique can be significantly improved through the simultaneous diagonalization of several random matrices. Here, we present a new approach based on the simultaneous Schur triangularization of a set of nearly commuting matrices (Corless et al., 1997).

2.3 Spectral methods and sequence analysis

To the best of our knowledge, spectral methods have not been applied so far to the problem of DNA sequence motif discovery that we address here. Nevertheless, spectral techniques have been applied to other types of sequence analysis problems, such as poly(A) motif prediction (Xie et al., 2013), chromatin annotation (Song and Chen, 2014), and sequence prediction (Quattioni et al., 2014). The techniques used in these works are all based, with minor modifications, on the spectral algorithm of Hsu et al. (2012) for learning Hidden Markov Models, in which a dataset of time-series of observed values \( \{x_1, x_2, \ldots \} \) is used to recover a single observation matrix \( (O_x) \) whose columns are the conditional probabilities associated with the hidden states. Our approach marks a significant departure from these methods by allowing the recovery of distinct observation matrices \( (O_x, O_y, \ldots) \) and hence the extraction of motif PWM’s. Finally, we note that a general technique for learning mixture of product distributions in the presence of a background has been recently presented (Zou et al., 2013); it would be interesting to study how this technique could be applied to the problem of sequence motif discovery.

3 Materials and methods

Binding site models represent the binding preferences of DNA-binding proteins via probability distributions over the set of all possible ‘words’ of some given length. If the DNA-protein interaction is assumed to be the sum of single protein-nucleobase interactions, these probability distributions can be represented by PWM’s (Stormo et al., 1986; Stormo, 2000). Given the length of the binding site \( \ell \), a position weight matrix is a \( 4 \times \ell \) matrix whose columns are interpreted as the probability distribution associated to the various position within the binding site. Then, according to the factorizability assumption, the total binding score of a particular sequence is obtained by summing (in the log domain), over all positions, the probability distribution (mixture of product distributions). Then the model parameters are estimated using a powerful machine learning technique (spectral method). Finally, the binding profile is identified as one of the components in the obtained model. Next we provide a high-level description of the above three steps, in a simple (ideal) case where a single binding site (no secondary motif) of length three is over-represented in the dataset, and the protein binding preferences are exactly described by a \( 4 \times 3 \) position weight matrix PWM.

In the first step we assume that the dataset is drawn from a suitable probability distribution. Given three consecutive letters in the sample, one should consider two cases: (i) the three letters belong to one instance of the binding site or (ii) the three letters belong to the background. When the three letters belong to a binding site their probability is given by the binding model

\[
P_{\text{motif}}(I, J, K) = \text{PWM}(I) \text{PWM}(J) \text{PWM}(K),
\]

where \( I, J, K \) take values from the set \( \{A, C, G, T\} \). When the three letters do not belong to a binding site, their probability is given by the (constant) background distribution

\[
P_{\text{background}}(I, J, K) = B(I) B(J) B(K),
\]

where \( B(I) \) is the frequency of the letter \( I \in \{A, C, G, T\} \) in the dataset.

Then the probability of observing three consecutive letters \( (I, J, K) \) at a random position in the sample is given by

\[
P_{\text{sample}}(I, J, K) = \text{PWM}(I) \text{PWM}(J) \text{PWM}(K) \cdot \text{PWM}(I) \text{PWM}(J) \text{PWM}(K),
\]

where \( W \) is the overall probability of finding a binding site in the sample. In other words, each overlapping subsequence of length three in the dataset can be assumed to be drawn either from \( P_{\text{motif}} \) or \( P_{\text{background}} \), with probability given by the coefficient \( W \) and \( 1 - W \), respectively. In particular, letting \( S_3 \) be the set of all overlapping sub-sequences of length 3, one can write \( S_3 \sim P_{\text{sample}} \).

The second step consists of recovering the entries of both \( P_{\text{motif}} \) and \( P_{\text{background}} \) from the dataset. In a general, that would require solving a hard non-convex constrained optimization problem, with a large number of parameters. Spectral methods provide a powerful technique to obtain the entries of \( P_{\text{motif}} \) and \( P_{\text{background}} \) directly. The parameters are obtained from the eigenvalues of suitable ‘observable’ matrices, computed by counting joint frequencies in \( S_3 \). The observable matrices are formed out of the empirical pairwise and triple probability distributions, defined as the probabilities of observing two or three consecutive letters in \( S_3 \). For any three letters \( I, J, K \), where each letter belongs to \( \{A, C, G, T\} \), the corresponding triple empirical probability, denoted by \( \hat{P}(I, J, K) \), is obtained by the number of occurrences of the triple \( I, J, K \) in the dataset. Since \( I, J, K \) assume discrete values in an alphabet of four letters, the triple empirical probability \( \hat{P}(I, J, K) \) is a \( 4 \times 4 \times 4 \) multi-dimensional array (tensor), obeying \( \sum_{I,J,K} \hat{P}(I, J, K) = 1 \). Equivalently, the triple empirical distribution can be represented as a set of four \( 4 \times 4 \) matrices, namely \( \hat{P}(I, J, A), \hat{P}(I, J, C), \hat{P}(I, J, G), \hat{P}(I, J, T) \), whose columns sum to one. The main idea behind FastMotif is that, assuming the factorizability of the position weight matrix \( \text{PWM}_{\text{motif}} \) and the background matrix \( \text{PWM}_{\text{background}} \), triple and pairwise empirical distributions can be written in a very simple form, as we will explain next. Moreover, by multiplying different empirical distributions together, it is possible to form ‘observable’ matrices, whose eigenvalues are directly related to the entries of \( \text{PWM}_{\text{motif}} \) and \( \text{PWM}_{\text{background}} \).

The last step consists of identifying \( \text{PWM}_{\text{motif}} \) with the binding model and \( \text{PWM}_{\text{background}} \) with the background distribution. This is done via an exact \( P \)-value test where the matrices are used to define a classifier to distinguish between sequences containing the binding site and randomly reshuffled sequences.

3.1 Detailed description of FastMotif

3.1.1 Estimating the motif length

In the earlier example, we assumed an a priori knowledge of the target motif length \( \ell = 3 \). In general, the expected length of the binding site is not known a priori, and it should be estimated via a statistical test over the sample. FastMotif obtains an estimation of the binding site length by measuring the Pearson test statistic of all sub-sequences of length \( k \) appearing in it. For \( k = 3, \ldots, 15 \) the expected length is defined as

\[
\ell = \arg \max_i \frac{(O_k - E_k)^2}{E_k}
\]

where \( O_k \) is the frequency of the most represented subsequence of length \( k \) in the random subset and \( E_k = 4^{-k} \) is an expected
theoretical frequency, based on the null hypothesis $P(A) = P(C) = P(G) = P(T) = \frac{1}{4}$. For speed reasons the statistical test is performed on a small random subset of the whole dataset (∼5000 sequences).

Given an expected motif length $\ell$ of the target binding site, we define $S_\ell$ to be the collection of all overlapping sub-sequences of length $\ell$ in the dataset. Equivalently, $S_\ell$ can be thought as the set of all records of a sliding window of length $\ell$ running over the sample.

### 3.1.2 Modelling secondary motifs

To relax the assumption of a single binding site, the two-component mixture in the earlier example is replaced in FastMotif by a $p$-component mixture with factors $PWM_r$, with $r = 1, 2, \ldots, p$. In this case, the set $S_\ell$ is modelled by a convex combination of $PWM_r$ models, that is $S_\ell \sim \sum_r \nu_r PWM_r$, with $\nu_r \geq 0$ and $\sum_r \nu_r = 1$. Intuitively, the $p$ $PWM_r$s are associated to the primary motif, the background, and the remaining $p-2$ secondary motifs, respectively. We note that, even if the dataset is expected to contain a single binding motif, working with a $p$-component mixture increases the robustness of the model, as it allows taking into account possible constant patterns that often appear in the background.

### 3.1.3 The spectral method

We now describe how we can extract the entries of each matrix $PWM_r$ from a set of ‘observable’ matrices using the spectral method. For illustration purposes, let us assume again that $\ell = 3$, and let $S_3$ denote the set of all 3-mers in the dataset. Let $\hat{P}$ be the empirical joint distribution whose entry $\hat{P}(l, j, k)$ is the frequency of the sub-sequence $[l, j, k]$ in $S_3$. Pairwise joint probability matrices are defined analogously and can be obtained from $\hat{P}$ by marginalization, for example $P_{12} = \sum_k \hat{P}(l, \cdot, k)$. More generally, the 3D array $\hat{P}$ can be transformed into a matrix by considering a linear combination of its 2D slices. For example, slicing on the 3-direction, one can select one of the four slices by setting $P_i = \sum_k \hat{P}(i, \cdot, k)$, where $i \in \{A, C, G, T\}$ and $\hat{e}_i$ is a 4D basis vector with 1 on the $i$-th entry and 0 elsewhere.

Assuming a $p$-component mixture model $\sum_r \nu_r PWM_r$, we define each component as $PWM_r = \{X_r, Y_r, Z_r\} \in [0, 1]^{k \times 3}$, where each column $X_r$ (i.e., $Y_r, Z_r$) is a probability distribution over the alphabet $\{A, C, G, T\}$. We also define a $4 \times p$ conditional probability matrix $X = \{X_1, \ldots, X_p\}$, whose $r$-th column is the first column of $X_r$ of the $r$-th PWM model, and analogously for $Y$ and $Z$. Then, assuming that the data are drawn exactly from a mixture of $p=4$ product distributions, and under the factorizability assumption of the PWM model, it is easy to verify the following identities:

$$\hat{P} = \sum_{i=1}^{p} \nu_i X_i Y_i Z_i,$$

$$\hat{P}_{12} = X_{diam} \hat{e}_i Y_i Z_i,$$

$$\hat{P}_i = X_{diam} \hat{e}_i Y_i Z_i, \quad i \in \{A, C, G, T\},$$

where $\hat{w} = (\nu_i)$ is the vector of mixing weights, and $\text{diam}(v)$ denotes a diagonal matrix with the entries of the vector $v$ on the diagonal.

The key idea in spectral methods is that, provided that (2) and (3) hold and the matrices $X$ and $Y$ are invertible, we can recover the entries of the matrix $Z$ from the eigenvalues of four ‘observable’ matrices defined as

$$M_i = \hat{P} \hat{P}_i^{-1} = X_{diam} \hat{e}_i Y_i Z_i^{-1}, \quad i \in \{A, C, G, T\}.$$  

(4)

According to (4), the observable matrices $M_i$ are simultaneously diagonalizable for all $i \in \{A, C, G, T\}$. From the definition of $Z = \{Z_1, \ldots, Z_p\}$, one has $\hat{c}_i^T Z = \{[Z_1], \ldots, [Z_p]\}$, where $Z_r$ is the probability distribution at position $r$ according to the $r$-th model. Given the set of observable matrices $M_A, M_C, M_G, M_T$, we can recover all entries of the matrix $Z$ from the matrix of their eigenvalues $\lambda_r = \hat{\lambda}_i(M_r)$, where we denote by $\lambda_i(A)$ the $r$-th eigenvalue of a matrix $A$. The entries of $X$ and $Y$ are obtained from analogous observable matrices, where the role of the three variables $(l, j, k)$ is interchanged.

### 3.1.4 Model misspecification

So far we have assumed that the model is exact, i.e., the data are drawn exactly from a mixture of $p$ product distributions and an infinite size sample is available. When the model is misspecified, (4) holds only approximately and the observable matrices $M_i$ are not exactly simultaneously diagonalizable. In this case, the entries of the conditional probability matrices can be recovered via an approximate diagonalization of the matrices $M_i$, for $i = A, C, G, T$. Letting $M_i = M_i + E$, where $M_i = X_{diam} \hat{e}_i Y_i Z_i^{-1}$ and $E$ is the misspecification term, it is then possible to obtain upper bounds for the eigenvalue estimation error in terms of $\|E\| = \sqrt{\sum E_{ij}^2}$. The standard procedure consists of choosing an arbitrary linear combination of the observable matrices $M(\theta) = \sum_i M_i \beta_i$ compute its eigenvectors and use them to approximately diagonalize all single $M_i$, for $i \in \{A, C, G, T\}$ (Anandkumar et al., 2012c; Hsu et al., 2012; Mossel and Roch, 2006). The choice of the linear combination coefficients $\theta$ is crucial and the stability of the method depends strongly on the eigenvalues separation of the matrix $M(\theta)$ (Hsu et al., 2012). Some separation guarantee can be obtained from classic concentration bounds on the Gaussian distribution, if the vector $\theta$ is chosen to be a random Gaussian vector of zero mean and unit variance (Mossel and Roch, 2006).

In FastMotif, we depart from the standard spectral procedure in two ways. First, in order to improve the stability of the spectral decomposition, we compute an ad hoc vector $\theta$ from a previous estimate of the conditional matrix $Z$. This estimate is obtained from the eigenvectors of a different observable matrix, obtained by exchanging the role of the sub-sequence positions in the definition of $P_i$ and $\hat{P}_{12}$, and we can show that the resulting $\hat{\theta}$ has good separation guarantees (depending on the previous estimate’s error). Second, instead of using the eigenvectors of the matrix $M$ to diagonalize the matrices $M_i$, we estimate the eigenvalues of $M_i$ using the orthogonal matrices appearing in the Schur decomposition of $M$ instead of its eigenvectors. Intuitively, we exploit the fact that the Schur form of a matrix $A$ is not unique on its upper-diagonal terms but it always contains the eigenvalues of $A$ on the diagonal (Corless et al., 1997). In the Results section, we provide a numerical comparison between the standard spectral approach and the simultaneous diagonalization approach of FastMotif based on the Schur decomposition, demonstrating the advantages of the proposed method (Fig. 4).

### 3.1.5 General setup

For simplicity, we have presented so far the simple special case of motifs of length three and have restricted the number of mixture components to $p = d$. Binding site models of general length $\ell > 3$ can be obtained by iterating the procedure described earlier with $l, j, k$ assigned to different positions within the motif. Regarding the number of components in the mixtures, one should distinguish between two situations. The case $p < d$ can be handled by reducing the size of the empirical matrices by means of a set of $p \times d$ orthogonal matrices obtained from the truncated singular value decomposition of $\hat{P}(x, y), \hat{P}(y, z)$ and $\hat{P}(x, z)$ (Anandkumar et al., 2012c). The case
Fig. 1. PWM computed by different algorithms on the datasets of Jolma et al. (2013). ‘Multinomial’ is the algorithm described in Jolma et al. (2013) and DREME (Bailey, 2011) is the motif discovery algorithm of the MEME suite designed for processing big datasets. All logos were obtained using weblogo 3.3 (Schneider and Stephens, 1990). Except for some discrepancies on the motif lengths, we observe very good agreement between the output of the three algorithms.

$p > d$ requires the definition of a new working space of dimensionality $D > p$, that can be obtained by considering grouped successive bases of length $n > 1$. For example, choosing $n = 2$, the set $S_r$ of all sub-sequences of length $\ell$ over the DNA alphabet $\{A, C, G, T\}$ is transformed into the equivalent set of sub-sequences of length $\ell_n$ over the grouped alphabet $\{AA, AC, \ldots, TT\}$ and one has $D = 4^n = 16$. Because higher dimensional variables are able to capture interdependencies between neighbouring positions of the binding sites, FastMotif maximizes the length of grouped variables.

Note that binding sites of general length $\ell > 3$ can be represented in terms of length-3 ‘high-dimensional’ binding sites. Given $\ell > 3$ it is always possible to define grouped variables of different lengths, say $n_1, n_2, n_3$, such that $n_1 + n_2 + n_3 = \ell$ and recover the binding model via the method described in the previous sections. However, in the case of grouped variables, the output of the spectral algorithm is a set of $p$ frequency matrices $HPWM = [X, Y, Z] \in [0, 1]^{D \times 3}$ whose columns are probability distributions over the alphabet of grouped variables. The corresponding 4D models are then obtained from the set $S_r$ as follows. According to the model assumption, the set $S_{r,1}$ of all sequences of length $\ell$ is the direct sum of exactly $p$ subsets each of them containing the sub-sequences of length $\ell$ generated by the $r$-th model. Letting, $S_r$ be the subset of sub-sequences generated by the model $r$, the corresponding 4D position weight matrix $PWM_r$, is defined by the frequencies of each letter $A, C, G, T$ at each position $i = 1, \ldots, \ell$ in $S_r$. To select whether a sub-sequence has to be included in a subset $S_r$, we define a scoring function $f_r(s) = \log \left( HPWM_r(s) / HB(s) \right)$, where $HB$ is a background model computed over $S_r$. The subsets $S_r$ are defined by

$$S_r = \{ s \in S_r : Pr[f_r(b) > f_r(s)] < \epsilon_{match} \} \quad (5)$$

where $b$ is a random sequence of length $\ell$ and $\epsilon_{match}$ is a user defined $P$-value matching threshold. In practice, since the computation of $Pr[f_r(b) > f_r(s)]$ can be expensive, we approximate the exact $P$-value computation by introducing a threshold $\xi_r$ such that $Pr[f_r(b) > \xi_r] < \epsilon_{match}$ is true over a finite set of $N_{match} \sim \frac{1}{\epsilon_{match}}$ random sequences. Finally, we admit a sequence $s \in S_r$ to the set $S_r$ if $f_r(s) > \xi_r$.

The last step consists of isolating the model corresponding to the protein binding site from the other $p - 1$ models in the mixture. For that we use the ame tool of the MEME suite (McLey and Bailey, 2010), which allows computing the exact $P$-value of each model over a test dataset consisting of positive sequences from the dataset and negatives obtained by random reshuffling of the positives.

4 Results

The present version of FastMotif has been optimized to process datasets from HT-Selex experiments on transcription factor binding. Due to the enormous number of sequences, HT-Selex data are hard to analyse using other available motif discovery algorithms. Designed to fill this gap, FastMotif is able to process HT-Selex datasets in few seconds and extract high quality probabilistic binding models that match those produced by other state-of-the-art algorithms.

In this section, we present and discuss the performance of the algorithm on various HT-Selex datasets and compare the motifs produced by FastMotif with the ones produced by other algorithms. To test the robustness of FastMotif to model misspecification, we have also tested FastMotif on semi-synthetic data with increasing amounts of noise. Finally, we report some important theoretical features of the algorithm and discuss its sensitivity with respect to some user-defined parameters that can be tuned to optimize the search.

4.1 HT-Selex data

What distinguishes FastMotif from other sequence discovery algorithms is the inference technique based on spectral methods. HT-Selex is a domain where high quality and biologically meaningful outputs can be obtained directly from spectral learning.

To test FastMotif on real data, we have focused on the HT-Selex experiments described by Jolma et al. (2013). All data are available at the European Nucleotide Archive (ENA) database under accession number ERP001824. For a given transcription factor, we have downloaded the dataset corresponding to the Selex cycle used to compute the binding models published in the Supplementary Material of Jolma et al. (2013). We have selected the following datasets: HMBOX1 (cycle 4, 29 156 sequences), HNF4A (cycle 4, 80 491 sequences), RFX3 (cycle 3, 195 356 sequences), ZIC1 (cycle 3, 267 963 sequences), ZSCAN4 (cycle 3, 144 041 sequences). All sequences have length $\sim 20$ bp.

The models computed by FastMotif have been compared with the ones computed on the same datasets by two other algorithms: the algorithm described in Jolma et al. (2013) (which we will refer to as ‘Multinomial’), and the large-dataset motif discovery tool DREME of the MEME suite (Bailey, 2011).

Other available motif discovery tools, as for example the popular Expectation-Maximization algorithm MEME (Bailey and Elkan, 1994) or STEME (Reid and Wernisch, 2011), were unable to process files of the size of the original datasets and could not be included in the comparison. Moreover, since the code of ‘Multinomial’ is not available, we have only considered the PWM published in Jolma et al. (2013). Both FastMotif and DREME ran on the same machine with default settings: for FastMotif we set the number of mixture components $p = 15$ and the matching $P$-value threshold to 0.001, and DREME was launched with the option ‘--m1’ that stops the
search when one motif has been found. In Figures 1 and 2, we show matching threshold is to assign each sub-sequence in the dataset to be drawn from a mixture of product distributions, the role of such

A quantitative comparison between the logos shown in Figure 1 has been obtained by computing the exact $P$-value of each model on a series of ground-truth test samples. For each transcription factor we have created 20 test samples containing 1000 positive sequences from the original dataset and 1000 negatives. The negative sequences were obtained by reshuffling of the positives. Given such test samples and the models in Figure 1, the corresponding $P$-values were computed using the tool `ame` (McLeay and Bailey, 2010), launched as `>ame -fix-partition 1000 -bgfile <background file><model>`, where the option `-fix-partition 1000` fixes the size of the positive set and the background file contains the single letter frequencies in the test dataset. In Figure 3 (left), we show the logarithm of the average $P$-value over the 20 tests and the corresponding variances $\delta(y) = \frac{1}{y} dy$ as error bars. We observe that FastMotif obtains the best score for all the models shown in Figure 1, except for the HMBOX1 model where the low performance of FastMotif is probably due to the small size of the dataset. A possible weakness of this evaluation is the uncertainty about the ground-truth test sample, built on the probably false assumption that all sequences in a dataset contain the relevant binding site. However, since all algorithms were tested on the same datasets, we expect the true number of false positives to affect only mildly the validity of our comparison test.

We have already briefly commented on the limited length of the models computed by DREME, that restricts the search to sub-sequences of length $\ell = 8$ for speed reasons. As it is shown in Figure 1, the motifs computed by DREME coincide with the most deterministic part of the models computed by FastMotif. The presence of several flanking positions is perhaps one reason for the better classification performance of the motifs computed by FastMotif. Another reason of the slightly better performance of the models computed by FastMotif

<table>
<thead>
<tr>
<th>Model</th>
<th>HNF4A</th>
<th>RFX3</th>
<th>ZIC1</th>
<th>ZSCAN4</th>
<th>CTCF</th>
<th>MAFK</th>
<th>ELF3</th>
</tr>
</thead>
<tbody>
<tr>
<td>DREME</td>
<td>2337</td>
<td>14217</td>
<td>84838</td>
<td>9315</td>
<td>9550</td>
<td>58695</td>
<td>23971</td>
</tr>
<tr>
<td>Multinomial</td>
<td>1243</td>
<td>1405</td>
<td>18904</td>
<td>29410</td>
<td>1665</td>
<td>10153</td>
<td>10118</td>
</tr>
<tr>
<td>FastMotif</td>
<td>1351</td>
<td>19243</td>
<td>92281</td>
<td>8866</td>
<td>10435</td>
<td>57656</td>
<td>18384</td>
</tr>
</tbody>
</table>

Fig. 2. Number of sequences used to compute the binding models shown in Figure 1. In same cases, the columns of the PWM computed by ‘Multinomial’ (Jolma et al., 2013) do not sum to the same value. In those cases, we report the number of sites of the most deterministic columns.
is probably their fully stochastic profile, compared with the almost deterministic models computed by DREME.

Finally, in Figure 3 (right), we show the execution times of FastMotif and DREME on datasets of increasing size. ‘Multinomial’ has been excluded from this comparison because we could not have access to any estimation of its running time. FastMotif was about 10 times faster than DREME, in each of the datasets considered.

4.2 Sensitivity with respect to noise level

To test the robustness of FastMotif with respect to the level of noise in the data and model misspecification, we created semi-synthetic data and studied the performance of the algorithm in increasing levels of noise. The semi-synthetic data were generated by implanting instances of a stochastic motif on random genome sequences of increasing length. The motif instances were generated by sampling from a position weight matrix [the ELF3 binding site model computed by FastMotif on the corresponding dataset of [Jolma et al., 2013]] via the Matlab function `discreteSample`. In total, we created 50 datasets, each containing 1000 sequences of length \( L = 10 + 5 \times n \), with \( n = 1, \ldots, 50 \). FastMotif ran on each dataset and the obtained models were compared with the reference ELF3 model by computing the distance in norm between the two PWM. With default settings, FastMotif outputs three models ranked according to their \( P \)-value and we have chosen for the comparison the model of lowest \( P \)-value. Moreover, when the matrices computed by FastMotif and the reference model had different lengths we restricted the comparison to the eight most informative positions. In Figure 4 (top), we report the difference in norm between the reference and the estimated model as a function of the amount of noise in the dataset, i.e. the length of background random genome sequences. Due to the presence of secondary motifs in the random genome sequences, the lowest \( P \)-value motif that we chose for the comparison did not always correspond to the expected target motif. We expect this issue to be the main cause of the big jumps in the error function (dashed line) shown in Figure 4 (top). When the length of the sequences becomes too large, trivial secondary sub-sequences start being over-represented in the dataset and the algorithm cannot distinguish between them and the instances of the reference motif. We also repeated the experiment by replacing the random genome sequences with random sequences generated using the Matlab function `randn` and the jumps disappeared (solid line). In summary, the above analysis seems to indicate that any use of our algorithm to noisy in vivo data, as for example ChIP-Seq data, may require some pre-processing steps that we do not address here.

4.3 Sensitivity with respect to user-defined parameters

FastMotif depends on two user-defined parameters: the number of mixture components \( (p) \) and a \( P \)-value matching threshold \( (\epsilon_{\text{match}}) \). The number of mixture components \( (p) \) defines the number of PWM to be included in the probability distribution that is used to model \( S_t \), the set of sub-sequences of length \( \ell \). Extra components are ideally associated to secondary motifs or over-represented redundant parts of the background, and make the algorithm more stable in the case of noisy data. The matching threshold \( (\epsilon_{\text{match}}) \) is used to assign each sequence in \( S_t \) to the correct mixture component and extract the corresponding position weight matrix. Small matching thresholds make this selection restrictive and the final position weight matrix more deterministic, while bigger matching thresholds increase the number of sites used to compute the model, and hence the noise. A careful trade-off between the information content of the final logo and the number of enriched sites to be used (Fig. 2) is often required. In this article, we have always set the user-defined parameters to their default values \( p = 15 \) and \( \epsilon_{\text{match}} = 0.001 \). In Figure 4 (centre), we show the quality of the output as a function of such user-defined parameters. The plot shows that for reasonable (not extreme) values of the two free parameters, the quality of the output is rather constant (plateau).

Finally, we comment on the new fully probabilistic motif discovery approach implemented in FastMotif. One of the main features is that no deterministic consensus sequence is required to initialize the search and models are computed directly from the empirical joint frequency matrices. The search strategy is analogous to the one used by MEME (Bailey and Elkan, 1994), where the sequence discovery problem is translated to the problem of learning a mixture of product distributions. The key novelty of FastMotif is the spectral learning algorithm that is used to infer the parameters of the mixture. In
particularly, FastMotif is built on a new and more stable spectral tech-
nique based on the Schur decomposition of matrices (Section 3). The
new method comes with a theoretical analysis and we provide bounds of
the error of parameter estimation as a function of the amount of
noise (Supplementary Material). In particular, we generated a set of
triple joint probability distributions using a random five-components
mixture, and used FastMotif and the standard spectral algorithm
(Chang, 1996) to infer the parameters of the corresponding four-com-
ponents approximations. The weight of the fifth component \( \epsilon \in [0, 0.2] \) has been used as a misspecification parameter. In Figure 4 (bot-
tom), we compare the distance in norm between the original four-
components model, obtained by subtracting the fifth component, and
its estimation for increasing values of \( \epsilon \). In average, and for almost all
values of the misspecification parameter, the FastMotif decomposition
scheme performs better than the standard approach.

5 Conclusions
Under the approximation that TF-DNA binding affinities are pos-
ition-independent, the problem of finding over-represented motifs in a
set of sequences is equivalent to the problem of learning a mixture of
product distributions. The inference of mixtures of product distribu-
tions is a well-known problem in statistics and machine learning, and
powerful techniques have been developed to solve the problem based
on spectral decompositions. We described FastMotif, a spectral motif
discovery algorithm that is fast, robust to model misspecification, not
prone to local optimal solutions, and that can search for motifs of ar-
bitrary length. We have tested the algorithm on HT-Seqexperimen-
tal data and produced PWM’s that match the profiles obtained by
other state-of-the-art motif finding algorithms, but one order of mag-
nitude faster. FastMotif is based on a new approximate simultaneous
matrix diagonalization scheme, for which we provide theoretical and
numerical error bounds. We have analysed the robustness of the algo-
rithm theoretically and via numerical experiments on semi-synthetic
data. Moreover, we have studied the sensitivity of the algorithm on its
input parameters and shown that, at least for HT-Seqex data, small
variations around their default values do not affect the quality of the
extracted motifs. Designed for the analysis of large-scale transcription
factor binding data, the current version of FastMotif restricts the
search to sub-sequences of relatively small length (from 3 to 15 nucle-
obases), but the arbitrary-length case can be handled with minor
modifications. For increasing search ranges, the complexity of the
spectral decomposition part of FastMotif, which is linear in the sam-
ple size, is substantially unchanged.

Because of its flexibility and speed, FastMotif can be a useful
tool in many important biological applications that go beyond the
identification of transcription factor binding sites. In future work,
we would like to extend our method to the more challenging domain
of RNA-binding proteins and consider applications in other kinds of
amino acid sequence analysis.

Acknowledgments
We would like to thank AnimA Anandkumar, Anke Wienecke-Baldacchino,
Merja Heinaniemi, Matthieu Sainlez, Luis Salamanca and Cédric Laczný for
useful discussions, comments and questions.

Funding
This work was supported by FNR-Luxembourg CORE grant 12/8M/
3971381 HIBIO. The bulk of this work was carried out when N.V. was with
the Luxembourg Centre for Systems Biomedicine.

Conflict of Interest: none declared.

References
Anandkumar, A. et al. (2012a) Tensor decompositions for learning latent vari-
Anandkumar, A. et al. (2012b) Learning high-dimensional mixtures of graph-
Anandkumar, A. et al. (2012c) A method of moments for mixture models and
Annala, M. et al. (2011) A linear model for transcription factor binding affinity
Aorà, S. et al. (2012) Learning topic models-going beyond SVD. In: Fon-
tations of Computer Science (FOCS), 2012 IEEE 53rd Annual Sym-
posium on, pp. 1–10. IEEE.
Badis, G. et al. (2009) Diversity and complexity in DNA recognition by tran-
Bailey, T. and Elkan, C. (1994) Fitting a mixture model by expectation
maximization to discover motifs in biopolymers. In: International Confer-
ence on Intelligent Systems for Molecular Biology; ISMB, Vol. 2, pp.
28–36.
Bailey, T.L. (2011) DREMIE: motif discovery in transcription factor ChiP-Seq
Balle, B. et al. (2014) Methods of moments for learning stochastic languages:
unified presentation and empirical comparison. In: Jebra,T. and Xing,E.P.
(eds), Proceedings of the 31st International Conference on Machine
Learning (ICML-14), pp. 1386–1394. JMLR Workshop and Conference
Proceedings.
the comprehensive characterization of the DNA-binding specificities of tran-
Berger, M.F. et al. (2006) Compact, universal DNA microarrays to compre-
hsensively determine transcription-factor binding site specificities. Nat.
Boots, B. et al. (2011) Closing the learning-planning loop with predictive state
interdependent effects on the binding affinities of transcription factors.
Chang, J.T. (1996) Full reconstruction of Markov models on evolutionary
for relative ranks of k-mers in binding transcription factors. Bioinforma-
tics, 23, 172–179.
Cheng, Q. et al. (2013) Computational identification of diverse mechanisms
Corless, R.M. et al. (1997) A reordered Schur factorization method for zero-
dimensional polynomial systems with multiple roots, pp. 133–140. ACM
Press.
BMC Bioinformatics, 8(Suppl. 7), S21.
Dempster, A.P. et al. (1977) Maximum likelihood from incomplete data via
Hsu, D. and Kakade, S.M. (2013) Learning mixtures of spherical gaussians:
moment methods and spectral decompositions. In: Proceedings of the 4th
11–20.
Johnson, D.S. et al. (2007) Genome-wide mapping of in vivo protein-DNA
Jolma, A. et al. (2010) Multiplexed massively parallel SELEX for characteri-
zation of human transcription factor binding specificities. Genome Res.,
20, 861–873.
Jolma, A. et al. (2013) DNA-binding specificities of human transcription fac-


Xie,B. et al. (2013) Poly (a) motif prediction using spectral latent features from human DNA sequences. Bioinformatics, 29, i316–i325.


