Bioinformatics, 31(16), 2015, 2623–2631
doi: 10.1093/bioinformatics/btv208
Advance Access Publication Date: 16 April 2015
Original Paper

OXFORD

Sequence analysis

FastMotif: spectral sequence motif discovery

Nicoló Colombo[1] and Nikos Vlassis[2,*]

[1]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg and [2]Adobe 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.

Availability and implementation: The Matlab code of FastMotif is available from http://lcsb-portal.uni.lu/bioinformatics.

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 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://lcsb-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 deterministic 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, c; Boots *et al.*, 2011; Chang, 1996; Hsu *et al.*, 2012; Mossel and Roch, 2006) [see (Balle *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 (Quattoni *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 $\{x_1, x_2, \dots\}$ 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 $(O_x, O_y, \dots)$ 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 ℓ, 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 matrix entries corresponding to the sequence letters.

De novo motif discovery algorithms compute one or more binding motifs from a set of enriched sequences, i.e. a set of sequences that contain, with high probability, several instances of the protein binding site. FastMotif computes the binding motifs in three steps: First the enriched sequences dataset is modelled using a special class of probability distributions (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 4×3 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 $P_{\text{motif}}(I,J,K) = PWM(I,1)\,PWM(J,2)\,PWM(K,3)$, 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 $P_{\text{sample}}(I,J,K) = WP_{\text{motif}}(I,J,K) + (1-W)P_{\text{background}}(I,J,K)$, 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_{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_{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_{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 sub-sequence $[I,J,K]$ in S_3, normalized by the total number of elements in S_3. Since I, J, K assume discrete values in an alphabet of four letters, the triple empirical distribution $\hat{P}(I,J,K)$ 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 4×4 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 PWM_{motif} and the background matrix $P_{\text{background}}$, 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 PWM_{motif} and $P_{\text{background}}$.

The last step consists of identifying PWM_{motif} with the binding model and $P_{\text{background}}$ 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 $k = 3, \dots, 15$ the expected length is defined as

$$\ell \;=\; \arg\max_{k} \frac{(O_k - E_k)^2}{E_k} \qquad (1)$$

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 ($\sim$5000 sequences).

Given an expected motif length ℓ of the target binding site, we define S_ℓ to be the collection of all overlapping sub-sequences of length ℓ in the dataset. Equivalently, S_ℓ can be thought as the set of all records of a sliding window of length ℓ 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_ℓ is modelled by a convex combination of PWM models, that is $S_\ell \sim \sum_{r=1}^{p} w_r PWM_r$ with $w_r \geq 0$ and $\sum_{r=1}^{p} w_r = 1$. Intuitively, the p PWMs 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 $\hat{P}(I, J, K)$ is the frequency of the subsequence $[I, J, K]$ in S_3. Pairwise joint probability matrices are defined analogously and can be obtained from $\hat{P}$ by marginalization, for example $\hat{P}_{12} = \sum_K \hat{P}(\cdot, \cdot, K)$. 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 $\hat{P}_i = \sum_K \hat{P}(\cdot, \cdot, K)[\mathbf{e}_i]_K$, where $i \in \{A, C, G, T\}$ and $\mathbf{e}_i$ is a 4D basis vector with 1 on the i-th entry and 0 elsewhere.

Assuming a p-component mixture model $\sum_{r=1}^{p} w_r PWM_r$, we define each component as $PWM_r = [X_r, Y_r, Z_r] \in [0, 1]^{4\times 3}$, where each column X_r (idem 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 (X_r) of the r-th matrix PWM_r, 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:

$$\hat{P} = \sum_{r=1}^{p} w_r X_r Y_r Z_r, \qquad \hat{P}_{12} = X \mathrm{diag}(w) Y^T, \tag{2}$$

$$\hat{P}_i = X \mathrm{diag}(w) \mathrm{diag}(\mathbf{e}_i^T Z) Y^T, \qquad i \in \{A, C, G, T\}, \tag{3}$$

where $w = (w_r)$ is the vector of mixing weights, and $\mathrm{diag}(v)$ 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

$$\hat{M}_i = \hat{P}_i \hat{P}_{12}^{-1} = X \mathrm{diag}(\mathbf{e}_i^T Z) X^{-1}, \qquad i \in \{A, C, G, T\}. \tag{4}$$

According to (4), the observable matrices $\hat{M}_i$ are simultaneously diagonalizable for all $i \in \{A, C, G, T\}$. From the definition of $Z = [Z_1, \ldots Z_p]$, one has $\mathbf{e}_i^T Z = [[Z_1]_i, \ldots [Z_p]_i]$, where Z_r is the probability distribution at position 3 according to the r-th model. Given the set of observable matrices $\hat{M}_A, \hat{M}_C, \hat{M}_G, \hat{M}_T$, we can recover all entries of the matrix Z from the matrix of their eigenvalues $\Lambda_{ir} = \lambda_r(\hat{M}_i)$, where we denote by $\lambda_r(A)$ 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 (I, 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 $\hat{M}_i$ 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 $\hat{M}_i$, for $i = A, C, G, T$. Letting $\hat{M}_i = M_i + E$, where $M_i = X \mathrm{diag}(\mathbf{e}_i^T Z) X^{-1}$ and E is the misspecification term, it is then possible to obtain upper bounds for the eigenvalue estimation error in terms of $||E|| = \sqrt{E^T E}$. The standard procedure consists of choosing an arbitrary linear combination of the observable matrices $\hat{M}(\theta) = \sum_i \hat{M}_i \theta_i$, compute its eigenvectors and use them to approximately diagonalize all single $\hat{M}_i$, 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 θ_i is crucial and the stability of the method depends strongly on the eigenvalues separation of the matrix $\hat{M}(\theta)$ (Hsu *et al.*, 2012). Some separation guarantee can be obtained from classic concentration bounds on the Gaussian distribution, if the vector θ 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 θ 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 $\hat{P}_i$ and $\hat{P}_{12}$, and we can show that the resulting θ has good separation guarantees (depending on the previous estimate's error). Second, instead of using the eigenvectors of the matrix $\hat{M}$ to diagonalize the matrices $\hat{M}_i$, we estimate the eigenvalues of $\hat{M}_i$ using the orthogonal matrices appearing in the Schur decomposition of $\hat{M}$ 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 I, 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

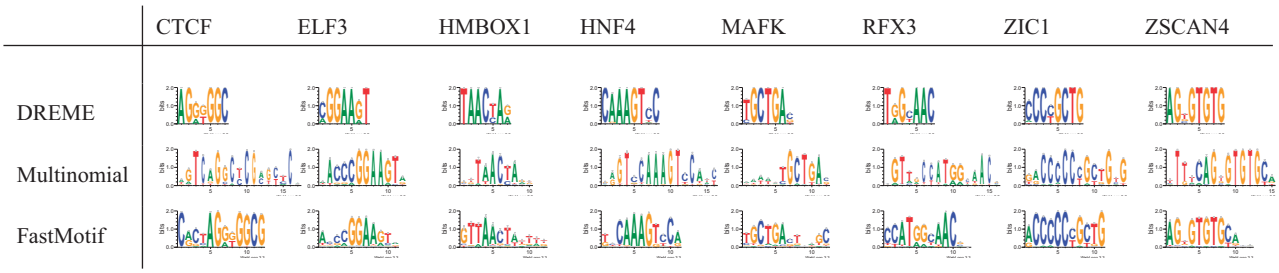

Fig. 1. PWM computed by different algorithms on the datasets of Jolma *et al.* (2013). 'Multinomial' is the algorithm described in Jolma *et al.* (2013) and DREME (Bailey, 2011) is the motif discovery algorithm of the MEME suite designed for processing big datasets. All logos were obtained using `weblogo 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 consecutive bases of length $n > 1$. For example, choosing $n = 2$, the set S_ℓ of all sub-sequences of length ℓ over the DNA alphabet $\{A, C, G, T\}$ is transformed into the equivalent set of sub-sequences of length $\frac{\ell}{2}$ over the grouped alphabet $\{AA, AC, \ldots, TT\}$ and one has $D = 4^n = 16$. Because higher dimensional variables are able to capture interdependences 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 $HPWM_r = [\tilde{X}_r, \tilde{Y}_r, \tilde{Z}_r] \in [0, 1]^{D \times 3}$ whose columns are probability distributions over the alphabet of grouped variables. The corresponding 4D models are then obtained from the set S_ℓ as follows. According to the model assumption, the set S_ℓ of all sequences of length ℓ is the direct sum of exactly p subsets each of them containing the sub-sequences of length ℓ generated by the r-th model. Letting, S_ℓ^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_ℓ^r. To select whether a sub-sequence has to be included in a subset S_ℓ^r, we define a scoring function $f_r(s) = \log(HPWM_r(s)/HB(s))$, where HB is a background model computed over S_ℓ. The subsets S_ℓ^r are defined by

$$S_\ell^r = \{s \in S_\ell : \Pr[f_r(b) > f_r(s)] < \epsilon_{\text{match}}\} \tag{5}$$

where b is a random sequence of length ℓ and ϵ_{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 ξ_r such that $\Pr[f_r(b) > \xi_r] < \epsilon_{\text{match}}$ is true over a finite set of $N_{\text{match}} \sim \frac{1}{\epsilon_{\text{match}}}$ random sequences. Finally, we admit a sequence $s \in S_\ell$ to the set S_ℓ^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 (McLeay 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 form 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), RXF3 (cycle 3, 195 356 sequences), ZIC1 (cycle 3, 267 963 sequences) ZSCAN4 (cycle 3, 68 378 sequences), CTCF (cycle 4, 134 566 sequences), ELF3 (cycle 3, 78 124 sequences), and MAFK (cycle 3, 144 041 sequences). All sequences have length ~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

	HMBOX1	HNF4A	RFX3	ZIC1	ZSCAN4	CTCF	MAFK	ELF3
DREME	2337	14217	84858	9315	9550	58695	23971	12057
Multinomial	1243	1405	18904	29410	1665	10153	10118	6125
FastMotif	1351	19243	92281	8866	10435	57656	18384	17914

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

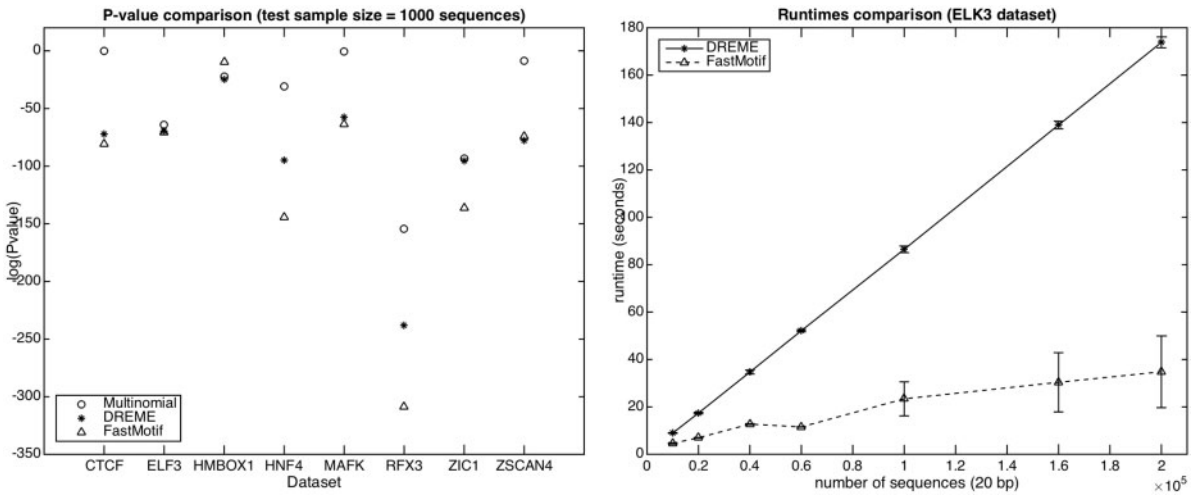

Fig. 3. Model quality (left): exact *P*-value of the models in Figure 1 computed via ground-truth tests. For each transcription factor, we show the logarithm of the average *P*-value obtained using the tool `ame` (McLeay and Bailey, 2010) over 20 distinct test datasets. The motifs computed by FastMotif obtained the best score for all the datasets except one. **Running times** (right): execution times of DREME and FastMotif plotted against the size of the dataset. All algorithms ran on the same machine and on the same datasets. For each dataset-size we have repeated the experiment 20 times and each point corresponds to the average runtime, with variances represented by the errorbars. For all the datasets considered, FastMotif has been about 10 times faster than DREME

search when one motif has been found. In Figures 1 and 2, we show respectively the models obtained by the different algorithms on the various datasets, and the corresponding number of sites used to compute the frequency matrices. All logos in Figure 1 were computed using the application `weblogo 3.3` (Schneider and Stephens, 1990). In some case the columns of the models computed by 'Multinomial' do not sum to the same values and we have reported the sum of the entries in the most deterministic column. In general, the numbers shown in Figure 2, compared with the total amount of sequences in the datasets, show that only a relatively small subset of sequences contains the expected binding site. The experimental interpretation of this fact goes beyond the scope of this paper, but we only note that, except for one case, the number of 'sites' used to compute the final model by 'Multinomial' is significantly smaller than for DREME and FastMotif. This can partially explain the difference in length between the models computed by FastMotif and 'Multinomial'. Because we do not know how the motif length is fixed by 'Multinomial', we do not discuss further the possible biological meaning of such discrepancies. Conversely, on approximately the same amount of sites, FastMotif is able to compute models that are two or more positions longer than the models computed by DREME. The number of binding sites used to compute the model by FastMotif depends on the choice of a user defined *P*-value matching threshold ϵ_{match} (Section 3), which is here set to its default value 0.001. Assuming the set S_ℓ of all sub-sequences of a given length ℓ to be drawn from a mixture of product distributions, the role of such matching threshold is to assign each sub-sequence in the dataset to the correct component in the mixture. We show in Figure 4 (centre) that, for HT-Selex data, small variations of ϵ_{match} do not significantly affect the quality of the output, but we expect that a fine tuning may be needed for more noisy data.

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 $\delta(\log y) = \frac{1}{y}\delta y$ 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

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 `discrete sample`. In total, we created 50 datasets, each containing 1000 sequences of length $L = 10 + 5 * n$, 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 `randseq`, 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 (ϵ_{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_ℓ, the set of sub-sequences of length ℓ. 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 (ϵ_{match}) is used to assign each sequence in S_ℓ 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

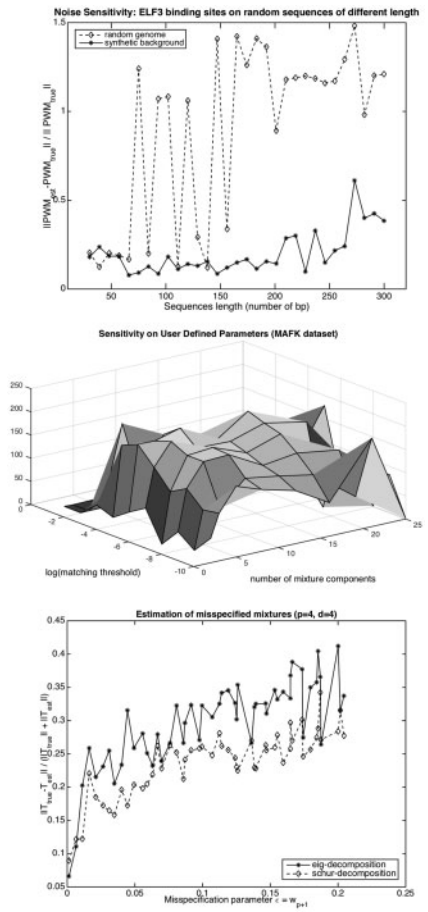

Fig. 4. Synthetic data experiment (top): distance between a ground-truth model and the FastMotif prediction on semi-synthetic data (dashed line) and synthetic data (solid). A set of motif instances, generated from a given binding site model (ELF3), have been inserted at random positions on random genome sequences (semi-synthetic data) or on randomly generated sequences (synthetic data) of increasing length. Better performances on the synthetic data can be explained by the presence of secondary motifs in the genome background. **Parameter sensitivity** (centre): *P*-value of the output model as a function of varying user-defined parameters, on HT-Selex data (MAFK dataset). For reasonable values of the parameters, the quality of the output is substantially constant (plateau). **Noise sensitivity** (bottom): a random four-components mixture $T_{\text{true}} \in [0,1]^{4\times4\times4}$ is perturbed by adding an extra component associated to the mixing weight $w_{p+1} = \epsilon \in [0, 0.2]$ and compared with its estimations T_{est} obtained via a standard spectral algorithm (Chang, 1996) (solid line) and FastMotif (dashed line). For each value of the misspecification parameter, we plot the average normalized distance $\frac{||T_{\text{true}} - T_{\text{est}}||}{||T_{\text{true}}|| + ||T_{\text{est}}||}$ over 50 equivalent experiments. In all cases, the Schur-based decomposition turned out to be more stable than the standard one

$p = 15$ 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

particular, FastMotif is built on a new and more stable spectral technique 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-components approximations. The weight of the fifth component $\epsilon \in [0, 0.2]$ has been used as a misspecification parameter. In Figure 4 (bottom), 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 ϵ. 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 position-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 distributions 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 arbitrary length. We have tested the algorithm on HT-Selex experimental data and produced PWM's that match the profiles obtained by other state-of-the-art motif finding algorithms, but one order of magnitude 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 algorithm 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-Selex 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 nucleobases), 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 sample 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 Laczny for useful discussions, comments and questions.

Funding

This work was supported by FNR-Luxembourg CORE grant 12/BM/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 variable models. *J. Mach. Learning Res.*, **15**, 2773–2832.

Anandkumar,A. *et al.* (2012b) Learning high-dimensional mixtures of graphical models. arXiv:1203.0697.

Anandkumar,A. *et al.* (2012c) A method of moments for mixture models and hidden Markov models. arXiv:1203.0683.

Annala,M. *et al.* (2011) A linear model for transcription factor binding affinity prediction in protein binding microarrays. *PLoS One*, **6**, e20059.

Arora,S. *et al.* (2012) Learning topic models-going beyond SVD. In: *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pp. 1–10. IEEE.

Badis,G. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.

Bailey,T. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: *International Conference on Intelligent Systems for Molecular Biology; ISMB*, Vol. 2, pp. 28–36.

Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-Seq data. *Bioinformatics*, **27**, 1653–1659.

Balle,B. *et al.* (2014) Methods of moments for learning stochastic languages: unified presentation and empirical comparison. In: Jebara,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.

Berger,M.F. and Bulyk,M.L. (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat. Protoc.*, **4**, 393–411.

Berger,M.F. *et al.* (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.

Boots,B. *et al.* (2011) Closing the learning-planning loop with predictive state representations. *Int. J. Robot. Res.*, **30**, 954–966.

Bulyk,M.L. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, **30**, 1255–1261.

Chang,J.T. (1996) Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Biosci.*, **137**, 51–73.

Chen,X. *et al.* (2007) RankMotif++: a motif-search algorithm that accounts for relative ranks of k-mers in binding transcription factors. *Bioinformatics*, **23**, i72–i79.

Cheng,Q. *et al.* (2013) Computational identification of diverse mechanisms underlying transcription factor-DNA occupancy. *PLoS Genet.*, **9**, e1003571.

Corless,R.M. *et al.* (1997) A reordered Schur factorization method for zero-dimensional polynomial systems with multiple roots, pp. 133–140. ACM Press.

Das,M.K. and Dai,H.-K. (2007) A survey of DNA motif finding algorithms. *BMC Bioinformatics*, **8**(Suppl. 7), S21.

Dempster,A.P. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)*, 1–38.

Hsu,D. and Kakade,S.M. (2013) Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In: *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, ACM, pp. 11–20.

Hsu,D. *et al.* (2012) A spectral algorithm for learning hidden Markov models. *J. Comp. Syst. Sci.*, **78**, 1460–1480.

Johnson,D.S. *et al.* (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.

Jolma,A. *et al.* (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.*, **20**, 861–873.

Jolma,A. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.

Kinzler,K.W. and Vogelstein,B. (1990) The GLI gene encodes a nuclear protein which binds specific sequences in the human genome. *Mol. Cell. Biol.*, **10**, 634–642.

Lee,D. *et al.* (2011) Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res.*, **21**, 2167–2180.

Leslie,C. *et al.* (2002) The spectrum kernel: a string kernel for SVM protein classification. In *Pacific symposium on biocomputing*, **7**, 566–575.

Lindsay,B.G. (1995) Mixture models: theory, geometry and applications. In: *NSF-CBMS Regional Conference Series in Probability and Statistics*, pp. 1–163. JSTOR.

Mathelier,A. and Wasserman,W.W. (2013) The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.*, **9**, e1003214.

McLeay,R.C. and Bailey,T.L. (2010) Motif enrichment analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics*, **11**, 165.

Mossel,E. and Roch,S. (2006) Learning nonsingular phylogenies and hidden Markov models. *Ann. Appl. Probab.*, **16**, 583–614.

Orenstein,Y. and Shamir,R. (2014) A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Res.*, **42**, e63.

Quattoni,A. *et al.* (2014) Spectral regularization for max-margin sequence tagging. In: *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 1710–1718.

Reid,J.E. and Wernisch,L. (2011) STEME: efficient EM to find motifs in large data sets. *Nucleic Acids Res.*, **39**, e126.

Sandve,G. *et al.* (2007) Improved benchmarks for computational motif discovery. *BMC Bioinformatics*, **8**, 193.

Santolini,M. *et al.* (2013) Beyond position weight matrices: nucleotide correlations in transcription factor binding sites and their description. arXiv:1302.4424.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Simcha,D. *et al.* (2012) The limits of de novo DNA motif discovery. *PLoS One*, **7**, e47836.

Song,J. and Chen,K.C. (2014) Spectacle: faster and more accurate chromatin state annotation using spectral learning. bioRXiv: 002725

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Stormo,G.D. *et al.* (1986) Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucleic Acids Res.*, **14**, 6661–6679.

Titterington,D.M. (1985) *Statistical Analysis of Finite Mixture Distributions. Wiley Series in Probability and Mathematical Statistics*. Wiley, Chichester, NY.

Tompa,M. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.

Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage t4 DNA polymerase. *Science*, **249**, 505–510.

Vert,J.-P. *et al.* (2005) Kernels for gene regulatory regions. In: *Advances in Neural Information Processing Systems*, pp. 1401–1408.

Wei,G.-H. *et al.* (2010) Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J.*, **29**, 2147–2160.

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

Zhang,Z. *et al.* (2013) Simultaneously learning DNA motif along with its position and sequence rank preferences through expectation maximization algorithm. *J. Comput. Biol.*, **20**, 237–248.

Zhao,Y. *et al.* (2009) Inferring binding energies from selected binding sites. *PLoS Comput. Biol.*, **5**, e1000590.

Zhao,Y. *et al.* (2012) Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics*, **191**, 781–790.

Zhao,Y. and Stormo,G.D. (2011) Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat. Biotechnol.*, **29**, 480–483.

Zou,J.Y. *et al.* (2013) Contrastive learning using spectral methods. In: *Advances in Neural Information Processing Systems*, pp. 2238–2246.