BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 10 2012, pages 1328–1335
doi:10.1093/bioinformatics/bts147

Sequence analysis

Advance Access publication March 29, 2012

A subspace method for the detection of transcription factor binding sites

Erola Pairó[1,2,*], Joan Maynou[3,4], Santiago Marco[1,2] and Alexandre Perera[3,4,*]

[1]Institut de Bioenginyeria de Catalunya, Baldiri Reixach 4, 08028 Barcelona [2]Departament d'Electrònica, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, [3]CIBER de Bioingeniería, Biomateriales y Biomedicina and [4]Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial (ESAII), Universitat Politècnica de Catalunya, Pau Gargallo 5, 08028 Barcelona, Spain

Associate Editor: John Quackenbush

ABSTRACT

Motivation: The identification of the sites at which transcription factors (TFs) bind to Deoxyribonucleic acid (DNA) is an important problem in molecular biology. Many computational methods have been developed for motif finding, most of them based on position-specific scoring matrices (PSSMs) which assume the independence of positions within a binding site. However, some experimental and computational studies demonstrate that interdependences within the positions exist.

Results: In this article, we introduce a novel motif finding method which constructs a subspace based on the covariance of numerical DNA sequences. When a candidate sequence is projected into the modeled subspace, a threshold in the *Q*-residuals confidence allows us to predict whether this sequence is a binding site. Using the TRANSFAC and JASPAR databases, we compared our *Q*-residuals detector with existing PSSM methods. In most of the studied TF binding sites, the *Q*-residuals detector performs significantly better and faster than MATCH and MAST. As compared with Motifscan, a method which takes into account interdependences, the performance of the *Q*-residuals detector is better when the number of available sequences is small.

Availability: http://r-forge.r-project.org/projects/meet

Contact: epairo@ibecbarcelona.eu, alexandre.perera@upc.edu

Supplementary information: Supplementary data (1, 2, 3 and 4) are available at *Bioinformatics* online.

Received on September 9, 2011; revised on March 1, 2012; accepted on March 23, 2012

1 INTRODUCTION

Deoxyribonucleic acid (DNA) sequence motifs are short sequence patterns with biological function. In the gene promoter region, there are DNA sequence motifs which hint at the interaction between the gene regulation machinery and the nucleic acids. They are involved in several DNA and ribonucleic acid (RNA) processes, such as the binding of some proteins to DNA, the ribosome binding to mRNA, and mRNA processing (D'haeseleer, 2006). Protein biosynthesis starts with a transcription process. This process, for example in eukaryotes, is led by several types of RNA polymerase that require special DNA sequences in promoters and a set of transcription factor (TF) proteins.

Due to the importance of gene regulation, a major problem in molecular biology is to discover the location of the transcription factor binding sites (TFBSs) within the genome. But the fact that most TFs bind to short, degenerate sequences makes it difficult to find sequence patterns to model the binding sites (Wasserman and Sandelin, 2004). Many algorithms try to characterize these patterns, and such algorithms may be classified into consensus-based methods or alignment-based methods (Pavesi *et al.*, 2004).

Most of the algorithms developed target the location of TFBSs. These follow one of two strategies: (i) to discover common binding sites into a set of unaligned sequences of co-regulated genes and (ii) to make use of the previous knowledge of sequences to search for a motif within a genome (Das and Dai, 2007; Elnitski *et al.*, 2006; Hannenhalli, 2008; Sandve and Drablos, 2006).

The algorithms which use the previous knowledge of the binding site sequences are mostly based on position-specific scoring matrices (PSSMs; Stormo, 2000). PSSM are matrices of frequencies of each nucleotide in each position of the binding site. Some examples of these algorithms are MATCH (Kel *et al.*, 2003), which uses information at each position to construct a PSSM; MAST (Bailey and Gribskov, 1998), based on the QFAST algorithm and part of MEME suite (Bailey and Elkan, 2006); rVISTA (Loots and Ovcharenko, 2004) which uses evolutionary data; and ITEME (Maynou *et al.*, 2010) which calculates the information loss of the binding sites. These models assume that the positions in binding sites are statistically independent. However, experimental evidence shows that TFBS have interdependences between positions (Bulyk *et al.*, 2002) and some computational studies suggest the same (Tomovic and Oakeley, 2007). These findings have motivated the development of new strategies which take into account position interdependences. Models based on Markov chains, such as WAM (Zhang and Marr, 1993), are restricted to modeling interdependences between adjacent positions. Other algorithms estimate non-adjacent interdependences using permuted Markov models (Zhao *et al.*, 2006); Bayesian networks (Barash *et al.*, 2003); variable order Bayesian networks (Ben-Gal *et al.*, 2005; Castelo and Guigó, 2004) or graphs (Naughton *et al.*, 2006). Detectors constructed using these techniques have higher accuracy, but require the tuning of many parameters for optimal operation which typically requires a large number of binding site instances. Additionally, most of these algorithms are computationally intensive.

On the other hand, a large body of knowledge exists for specific event detection in numerical sequences (signals), and the conversion

*To whom correspondence should be addressed.

of symbolical DNA sequences into numerical DNA sequences has been widely used in genomic signal processing to extract relevant biological information from DNA sequences. For example, numerical conversions have been used to identify protein coding regions by studying their periodicity (Anastassiou, 2001; Cristea, 2005; Shmulevich and Dougherty, 2007).

In this article, we propose a detector based on the Q-residuals of a numerical sequences covariance model. This contribution aims to study to what extent the covariance can capture information on position interdependences between binding sites. Our hypothesis is that, when projected into the subspace defined by the covariance, sequences belonging to the modeled TFBS should have smaller Q-residuals than chromosomic or random sequences, consequently Q-residuals could be used to detect binding sites. The proposed detector was compared with the PSSM-based methods, MAST and MATCH, using real genomic data. It was also compared with the Motifscan method which calculates interdependences between positions.

2 MATERIALS AND METHODS

2.1 Data

TFBS sequences were extracted from the TRANSFAC 7.0 2005 public database (Wingender *et al.*, 2000) and from JASPAR 2010 (Portales-Casamar *et al.*, 2010). For the JASPAR database, the motifs with 10 or more sequences were extracted. To carry out the study, we selected 43 motifs corresponding to *Homo sapiens*, 25 from *Mus musculus*, 11 from *Rattus norvegicus*; a further 10 were randomly chosen from all the TFBS available for *Drosophila melanogaster*. For the TRANSFAC database, the 108 motifs with >10 sequences were chosen. These motifs were multiple-aligned using the CLUSTALW2 algorithm (Larkin *et al.*, 2007) with default parameters. The alignment was performed N times, where N is the number of sequences for each motif, using a leave-one-out cross validation (L.O.O.) procedure. The 23 TFBS motifs having a core with >5 consecutive positions without gaps at each step of the L.O.O. procedure were used to compare our method to the existing PSSM algorithms. These binding sites correspond to eukaryotic organisms of different level of complexity, ranging from *Saccharomyces cerevisiae* to *H.sapiens* and including *D.melanogaster*, *R.norvegicus*, *M.musculus* and *Gallus gallus*. The number of selected sequences from JASPAR totalled 89 motifs. The relation of the 89 JASPAR motifs and the 23 TRANSFAC motifs is given in the Supplementary Material 2, and a summary of the TF used for each organism can be seen in Table 1.

All promoter sequences from the organisms used, with the exception of *S.cerevisiae*, were extracted from the Eukaryotic promoter database (EPD) sequences (Schmid *et al.*, 2006), using the EPD version based on EMBL release 105 (September, 2010). The sequences located at the positions from −1000 to 500 relative to the transcription start site (TSS) were used to construct the background model, consisting of the nucleotide frequencies for the promoters of each organism. In *S.cerevisiae*, the extracted sequences correspond to promoter sequences in Chromosomes 1 and 16 of the EMBL chromosome database (Kanz *et al.*, 2005), release 94 (March, 2008).

Table 1. Information about motifs used for each organism

Organism	JASPAR	TRANSFAC	Total
Saccharomyces cerevisiae	0	7	7
Drosophila melanogaster	10	3	13
Mus musculus	25	4	29
Rattus norvegicus	11	4	15
Homo sapiens	43	4	47
Gallus gallus	0	1	1
Total	89	23	112

In each organism, we randomly chose two promoter sequences of length 1501 nucleotides for use as background sequences. In *D.melanogaster*, we used the sequences from −1000 to 500 relative to TSS of *FAF* gene as Background 1 and the same range of nucleotides from gene *CG*12170 as Background 2. In *M.musculus*, the same range of nucleotides was set and the *Igk′T* gene was used as Background 1, whereas gene *Igk′MPC*11 was used as Background 2. In *R.norvegicus*, Background 1 was extracted from the myosin $LC3_fP2$ gene and Background 2 from *PSBPC*2. For *H.sapiens*, the promoter corresponding to Background 1 was in the region of the gene *RPS*9*P*2+ whereas the promoter corresponding to Background 2 was relative to *PSMA*2 TSS. In the study of *G.gallus*, Background 1 was relative to *apoVLDLII* TSS and Background 2 relative to *a′A—globin* TSS. Finally, in *S.cerevisiae*, the Background 1 sequence generally corresponded to positions 44 730–46 230 in Chromosome 1. However, an exception was made for ABF1 binding sites, since ABF1 binding sites are present in that promoter; for ABF1 Background 1, the sequence used corresponded to the positions 678 930–680 430 in this organism's Chromosome 16 whereas, for Background 2, the positions from 11 410 to 12 910 in Chromosome 1 were used in all the organism's studied binding sites.

2.2 Preprocessing

The aligned matrix of DNA sequences had to be converted to a rectangular matrix of numerical sequences. The first step was to translate symbolic DNA to numerical sequences using the conversion process proposed by Silverman and Linske (1986), where each nucleotide is placed at the vertex of a regular tetrahedron as in Equation (1):

$$\begin{aligned} A &\equiv (0,0,1) \\ C &\equiv \left(-\frac{\sqrt{2}}{3},\frac{\sqrt{6}}{3},-\frac{1}{3}\right) \\ G &\equiv \left(-\frac{\sqrt{2}}{3},-\frac{\sqrt{6}}{3},-\frac{1}{3}\right) \\ T &\equiv \left(2\frac{\sqrt{2}}{3},0,-\frac{1}{3}\right) \end{aligned} \tag{1}$$

where A, C, G and T are points in 3D Euclidian space corresponding to the a, c, g and t nucleotides, respectively. This conversion was chosen because it is symmetric for all nucleotides and is widely used in genomic signal processing (Liew *et al.*, 2005).

After conversion, each DNA sequence of length M became a sequence of length $3\times M$, concatenating numerical vectors corresponding to each nucleotide. Then, the N sequences belonging to the same TF were arranged in matrix format. The result was an $N\times(3M)$ matrix of numerical DNA.

Where gaps were produced during the alignment process, we imputed the numerical value of these gaps into the mean of the chromosome, taking into account the nucleotide probability distribution of the background organism and the conversion process. The location of the gaps within the tetrahedron is thus given by Equation (2):

$$\mathrm{GAP}=P(a)A+P(c)C+P(g)G+P(t)T \tag{2}$$

In this equation, GAP is a three-element vector corresponding to the position of the gap within the tetrahedron; A, C, G and T are the positions of a, c, g and t nucleotides in the vertexes of the tetrahedron; $P(a)$, $P(c)$, $P(g)$ and $P(t)$ are the nucleotides probabilities in the promoter of the organism. Only those positions where the information was available for at least half of the sequences were imputed, the others were neglected.

2.3 Definition of the subspace method

A covariance subspace model was computed for each binding motif using a principal component analysis (PCA) of the numerical DNA sequence representation (Pearson, 1901). To carry out the PCA, first the covariance of the numerical DNA matrix was calculated, then the data projected into the subspace where the covariance matrix is diagonal. In this subspace, relatively few components explain most of the covariance, thus reducing the dimensionality of the problem. This yields a bilinear decomposition of the set of aligned DNA sequences as defined in Equation (3):

$$X = AB^{\mathrm{T}} + E \tag{3}$$

where X is a $N \times (3M)$ TFBS numerical matrix, with N being the number of TFBS sequences and M the number of TFBS positions. A is the projected data, consisting of an $N \times n\mathrm{PCS}$ matrix called scores, where nPCS is the number of principal components chosen to construct the subspace. B is the $(3M) \times n\mathrm{PCS}$ loading matrix which defines the subspace into which data is projected, and E is the $N \times (3M)$ error matrix.

The covariance is a $3M \times 3M$ matrix which captures the covariances between the numerical positions. When it is diagonal, no interdependences exist between positions of a specific binding site. This information is, in our model, explained in the loadings which are almost zero when a position is conserved, and which differ from zero (either in a positive or negative sense) when a position varies. In the Supplementary Material 1, an example of the covariance matrix and the loadings for the DL binding sites, where covariances exists, is presented.

The detector was built using the Q-residuals, which are the square of the Euclidean distance from a sequence to the subspace generated by the Principal Components model. Given a candidate sequence, the Q-residuals can be calculated using Equation (4):

$$Q = EE^{\mathrm{T}} \tag{4}$$

where E is the $3M$ error vector obtained from projecting the sequence into the Principal Components subspace, and Q is the Q-residual of the candidate sequence.

The model should explain most of the variance and, as outlined above, sequences belonging to the studied TF should have smaller Q-residuals than the other sequences. Defining a threshold in Q-residuals should be sufficient to allow distinguishing between TFBS and sequences not belonging to the modeled TFBS. The threshold chosen is based on the Q-residuals statistics (Jackson, 2004), resulting in a confidence interval for a sequence belonging to our model. The Q-residuals distribution corresponding to the modeled TFBS sequences are first converted into a new $N(0,1)$ quantity C (i.e. C is normally distributed with mean $\mu = 0$ and variance $\sigma = 1$). The quantile with the desired confidence interval can be then calculated from this normal distribution. The constructed detector depends on the number of principal components chosen.

2.4 Comparison to PSSM algorithms

To compare our detector to existing PSSM methods, the MEET R package (available in the R-forge project http:// r-forge.r-project.org/projects/meet), was developed (Pairó *et al.*, 2011). This R package allows us to combine several alignment methods with different algorithms to search for TFBS within a large sequence. The package can be configured to call external alignment methods including CLUSTALW2, MUSCLE (Edgar, 2004), and MEME which has as an internal multiple alignment method. The proposed Q-residuals method is compared both with MAST and with an implementation of the MATCH algorithm which takes into account the probability distribution of the nucleotides in the promoter sequences of each organism.

To implement MATCH, the algorithm explained in Kel *et al.* (2003) was used, however, the background nucleotide probability distribution specific for each organism was also used. To detect a motif, first the PSSM matrix was calculated. Then, using this matrix, the information of each position was calculated as in Equation (5).

$$I(i) = \sum_{B=A,C,G,T} f_{i,B} ln(\frac{f_{i,B}}{P_B}) \tag{5}$$

where $I(i)$ is the information of position i, $f_{i,B}$ is the frequency of the B nucleotide in this position and P_B is the background probability of the B nucleotide. The Score of a sequence of length M was calculated as in Equation (6).

$$\mathrm{Score} = \sum_{i=1}^{M} I(i) f_{i,b_i} \tag{6}$$

where f_{i,b_i} is the frequency of the corresponding b_i nucleotide for the sequence in position i and $I(i)$ is the information in the same position. Finally, a SimilarityScore for the sequence and the core (first five consecutive more conserved positions), as explained in Equation (7) was used to discriminate between TFBS and other sequences as in the MATCH program (publicly available in TRANSFAC 7.0).

$$\text{Similarity Score} = \frac{\text{Score–Min}}{\text{Max–Min}} \tag{7}$$

Max and Min being the maximum and minimum possible scores for a candidate sequence.

Comparison with the MAST algorithm was done using the downloadable MEME 4.4.0 source available at the MEME suite—this allowed us to combine different alignment algorithms to construct the PSSM and then use the PSSM as an input to MAST. To calculate the PCA model and the Q-residuals in R, the pcaMethods R package was used (Stacklies *et al.*, 2007).

CLUSTALW2 with the default parameters, gapextend $= 0.2$, gapopen $=$ 10 was used to align the sequences in all the methods compared in TRANSFAC.

2.5 Validation

The MEET R package performs a double L.O.O. to calculate the ROC curves, the Area under ROC curve (AUC), and the errors associated with them. Given a motif of N sequences, first a sequence A is removed and inserted into the background sequence. Then, the remaining $N-1$ sequences of the same motif are used for a standard L.O.O. to construct models with $N-2$ sequences. These $N-2$ sequences are first aligned and the chosen algorithm is applied to build a model. Finally, each one of the $N-1$ models of the L.O.O. is used to detect the sequence A within the known position of the background. After that, sequence A is again inserted into the group and another sequence B is used, this whole process being repeated N times. As the location of the true positives is known, the threshold of the detectors can be varied in order to generate the N different ROC curves and AUCs. Thresholding is detector-specific; for Q-residuals it is the residuals statistics of the PCA model, for MATCH it is the sequence similarity and for MAST it is the p-value. Once the N ROC curves are generated, the SD is used to estimate the variability of the ROC curve points and the AUC.

In the case of the Q-residuals detector, AUC was calculated for from 1 to 10 principal components; in the case of MATCH, the varying parameter was the Core Similarity, ranging from 0.5 to 0.95 in increments of 0.05. Only one set of ROC curves and AUCs was calculated in MAST because the length of the sequence (the parameter to optimize in MEME) is defined by the number of positions of the PSSM constructed using the aligned sequences.

The mean and the variance of AUC for the studied range of principal components were calculated for each motif. Models built using different numbers of principal components can have an equivalent performance when the AUC mean and the AUC variance are taken into account. Between these models, the one with smallest AUC variance averaged between Backgrounds 1 and 2 was chosen as the best model. The same criterion was used to choose the threshold of Core Similarity in the MATCH algorithm.

As the number of negative examples greatly exceeded the number of positive examples in this study, it was also convenient to compare the algorithms using precision–recall (PR) curves. There exists a unique correspondence between the PR curves and the ROC curves, and when an algorithm dominates in the ROC spaces it also dominates in the PR space, however, optimizing the AUC under the two different methods is not the same thing (Davis and Goadrich, 2006). To show that the PR curves confirm the results obtained with the ROC curves, we calculated the curves with the optimal parameters for each detector (Supplementary Material 3). The ROC curves, the AUC and the PR curves were calculated using the ROCR package (Sing *et al.*, 2005).

2.6 Interdependences between positions

The improvement in detection of Q-residuals should be linked to the interdependences between positions in each binding site. To study this relation, the mutual information $\mathrm{MI}_{i,j}$ between positions i and j of the binding sites was calculated using Equation (8):

$$\mathrm{MI}_{i,j} = \sum_{b_i,b_j} P_{b_i,b_j,i,j} \log_2 \frac{P_{b_i,b_j,i,j}}{P_{b_i,i},P_{b_j,j}} \qquad (8)$$

where b_i and b_j correspond to the nucleotides in the studied positions i,j and P_{b_i} is the probability of the b_i nucleotide in the position i. The joint probability of having nucleotide b_i in position i and b_j in position j is described by P_{b_i,b_j}. The Bayes factor (BF) described in Equation (9) was used to test the null hypothesis, H_0, of independence between positions i and j against H_1, the alternative hypothesis of dependence, in order to determine the significance of the dependencies found:

$$\mathrm{BF}(H_0;H_1) = \frac{\Gamma(\sum_{b_i,b_j}\alpha_{b_i,b_j}}{\Gamma(M+\sum_{b_i,b_j}\alpha_{b_i,b_j}} \prod_{b_i} \frac{\Gamma(N(b_i,i)+\alpha_{b_i})}{\Gamma(\alpha_{b_i})}$$

$$\prod_{b_j} \frac{\Gamma(N(b_j,j)+\alpha_{b_j})}{\Gamma(\alpha_{b_j})} \prod_{b_i,b_j} \frac{\Gamma(\alpha_{b_i,b_j})}{\Gamma(N(b_i,b_j,i,j)+\alpha_{b_i,b_j})} \qquad (9)$$

where M is the size of the bindings sites sequences, $N_{b_i,i}$ is the number of b_i nucleotides in position i and α refers to the parameter of the Dirichlet prior distribution. This measure was used in previous studies to show which positions of a TF have interdependences (Tomovic and Oakeley, 2007; Zhou and Liu, 2004). When $\alpha_{b_i,b_j} = 1$ and $\alpha_{b_i} = \sum_{b_j} \alpha_{b_i,b_j}$ the BF is related to the mutual information as shown in Equation (10) (Minka, 2003).

$$\log_2(\mathrm{BF}(H_0;H_1)) \approx -M\mathrm{MI}_{i,j} \qquad (10)$$

Formula (10), where $\mathrm{MI}_{i,j}$ is the mutual information and M the size of the binding sites, was used to calculate the BF, $\mathrm{BF}(H_0;H_1)$. As in Tomovic and Oakeley (2007), a threshold of $\mathrm{BF} < 0.1$ was set to indicate strong evidence of interdependences between positions. For each motif, the percentage of positions showing interdependences, I_{dep}, was calculated.

2.7 Comparison to Motifscan

Naughton *et al.* (2006), used 94 JASPAR (2006) motifs to compare Motifscan, a graph-based method which takes into account interdependences, to PSSM methods. To do the comparison, they calculated the ROC_N curves, where N is the number of sequences for the selected motif, and its AUC.

Using the same methodology and 93 of the 94 motifs of the old JASPAR version (the old version of the remaining one was not available), the AUC of the ROC_N curves was calculated for the Q-residuals detector, and the results were used to compare the detectors. The comparison between Motifscan and Q-residuals using the 93 JASPAR motifs is available as Supplementary Material 4.

3 RESULTS

In this section, we first present the results of the comparison between the Q-residuals detector, MATCH and MAST using the 112 motifs presented above and two different backgrounds for each organism. Then, we describe in more detail the comparison between MAST and Q-residuals, and show a study of the interdependences. We present an analysis of the computational time needed for each one of the studied detection algorithms, and finally we compare the Q-residuals detector to the results obtained with Motifscan.

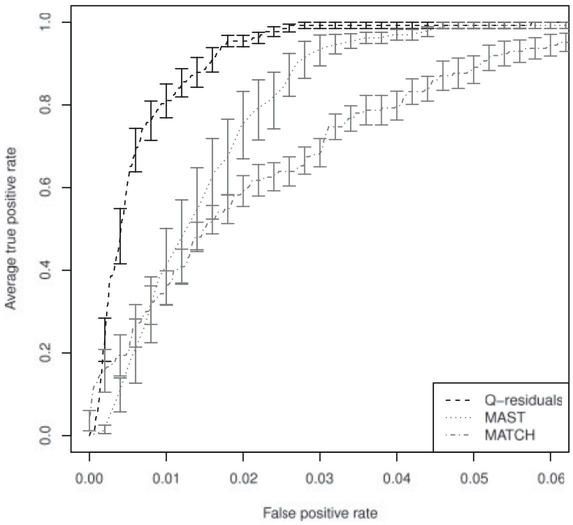

Fig. 1. ROC curve for Q-residuals, MAST and MATCH using the cMyB TF and the Homo Sapiens Background 1. The ideal number of components and the ideal MATCH Core Similarity were used to compute the ROC curve. The error bars correspond to the variation in detection using the L.O.O. cross validation. The figure shows the improvement of detection using Q-residuals.

One example of detection can be seen in the cMyB motif in Figure 1, a set of TFBSs for *H.sapiens*. The ROC curves show the performance of the three algorithms using the first background for *H.sapiens*. A significant improvement is observed when the Q-residuals detector is used in place of MAST or MATCH.

To visualize the performance of the three different detectors in all the studied TFs, Table 2 summarizes the results for Q-residuals, MATCH and MAST for the two different backgrounds in each organism for TRANSFAC. The best number of components (usually between 1 and 4) is shown, together with the mean AUC for each background and method. The results for all the studied TFs are available as Supplementary Material 2.

To quantify the differences in performance between the Q-residuals detector and the other algorithms, a Wilcoxon rank-test (Wilcoxon, 1945) was performed on the AUC distributions, using the null hypothesis that the two distributions are the same versus the alternative hypothesis that AUC using Q-residuals is closer to 1 than when MAST or MATCH are used. In Table 2 and the Supplementary Material 2, the increment in AUC and the significance of the test are displayed, and it can be seen that Q-residuals performs significantly better than Match in 57 of the 112 motifs studied and significantly better than MAST in 63 of them, with *p-value* <0.05.

To better visualize the detectors, we present the AUC box plots in Figure 2. These box plots represent the AUC and its variation when the L.O.O. is applied. Figure 2 shows the box-plots for the first background and the JASPAR motifs corresponding to *M.musculus*. In most cases, not only is the mean AUC closer to one in Q-residuals,

Table 2. Results for Q-residuals detector compared with MATCH and MAST algorithms, corresponding to the two backgrounds of each organism in TRANSFAC. The AUC shown for each method is the mean of the areas using the cross-validation method and the number of principal components for Q-residuals is chosen as the number of components with less variance in the AUC. The ΔAUC is the mean AUC improvement of Q-residuals versus MATCH and MAST, respectively. The level of significance corresponds to the p-value calculated when a Wilcoxon-rank test is performed, with the null hypothesis being that the AUC distributions using Q-residuals detector and the other algorithm are the same and the alternative hypothesis being that the AUC distributions calculated with the Q-residuals detector is closer to one. A description of the 89 JASPAR motifs and 23 TRANSFAC motifs can be found in the Supplementary Material 2

TF	nPCs	Q-residuals 1	Q-residuals 2	Match 1	Match 2	ΔAUC Match	MAST 1	MAST 2	ΔAUC MAST
ABF1	4	0.9991	0.9975	0.9902	0.9964	$5\cdot 10^{-3}$ ***	0.9957	0.9986	$1.14\cdot 10^{-3}$
BCD	3	0.9961	0.9952	0.9912	0.9884	$5.85\cdot 10^{-3}$***	0.9913	0.9947	$2.68\cdot 10^{-3}$*
CAT8	3	0.9998	0.9995	0.9971	0.9978	$2.21\cdot 10^{-3}$***	0.9999	0.9992	$9.02\cdot 10^{-5}$
CEBP β 35	3	0.9931	0.9965	0.9863	0.9878	$7.75\cdot 10^{-3}$ **	0.9936	0.9946	$6.66\cdot 10^{-4}$
cJun	1	0.9868	0.9915	0.9700	0.9813	$1.35\cdot 10^{-2}$ **	0.9575	0.9880	$1.64\cdot 10^{-2}$*
cMyB	1	0.9905	0.9907	0.9714	0.9714	$1.92\cdot 10^{-2}$***	0.9818	0.9869	$6.21\cdot 10^{-3}$*
DL	1	0.9982	0.9962	0.9835	0.9864	$1.23\cdot 10^{-2}$ ***	0.9682	0.9917	$1.73\cdot 10^{-2}$*
E2F	4	0.9997	0.9998	0.9991	0.9998	$3.00\cdot 10^{-4}$ *	0.9988	0.9995	$5.26\cdot 10^{-4}$
GAL4	1	0.9998	0.9999	0.9742	0.9759	$2.48\cdot 10^{-2}$ ***	0.9875	0.9653	$2.34\cdot 10^{-2}$*
GCN4	1	0.9988	0.9997	0.9936	0.9937	$5.68\cdot 10^{-3}$ ***	0.9951	0.9935	$5.06\cdot 10^{-3}$***
HNF1 α	9	0.9945	0.9940	0.9807	0.9850	$1.14\cdot 10^{-2}$ *	0.9943	0.9921	$2.1\cdot 10^{-3}$
HNF4 α	4	0.9957	0.9972	0.9870	0.9938	$6.05\cdot 10^{-3}$ *	0.9937	0.9957	$1.79\cdot 10^{-3}$
HNF6 α	1	0.9977	0.9996	0.9961	0.99358	$3.81\cdot 10^{-3}$***	0.9838	0.9949	$9.37\cdot 10^{-3}$*
IRF1	2	0.9992	0.9994	0.9727	0.9912	$1.74\cdot 10^{-2}$**	0.9970	0.9992	$1.22\cdot 10^{-3}$
IRF8	3	0.9991	0.9981	0.9926	0.9791	$1.28\cdot 10^{-2}$ ***	0.9928	0.9967	$3.86\cdot 10^{-3}$**
KR	3	0.9923	0.9965	0.9933	0.9838	$5.85\cdot 10^{-3}$ *	0.9926	0.9929	$1.69\cdot 10^{-3}$
LyF1	3	0.9952	0.9958	0.9689	0.9823	$1.99\cdot 10^{-2}$***	0.9903	0.9853	$7.68\cdot 10^{-3}$**
MIG1	1	0.9986	0.9954	0.9766	0.9475	$3.49\cdot 10^{-2}$ ***	0.9895	0.9896	$7.49\cdot 10^{-3}$*
NF κ B	2	0.9998	0.9999	0.9995	0.9999	$3.08\cdot 10^{-4}$*	0.9991	0.9998	$4.38\cdot 10^{-4}$ ***
p50	2	0.9996	0.9999	0.9995	0.9999	$4.86\cdot 10^{-5}$	0.9994	0.9998	$1.72\cdot 10^{-4}$ *
RFX1	7	0.9921	0.9969	0.9721	0.9867	$1.51\cdot 10^{-2}$ ***	0.9871	0.9837	$9.09\cdot 10^{-3}$*
ROX1	8	0.9998	0.9985	0.9997	0.9993	$-3.5\cdot 10^{-4}$	0.9996	0.9980	$3.40\cdot 10^{-3}$*
T3R α	6	0.9923	0.9919	0.9754	0.9852	$1.18\cdot 10^{-2}$***	0.9854	0.9757	$1.15\cdot 10^{-2}$**

*Significant at $p < 0.05$; **significant at $p < 0.005$; ***significant at $p < 0.001$

but the variance is also smaller, which suggests that the Q-residuals algorithm behaves more robustly.

An average of the PR curves obtained in each leave-one-out iteration is also presented as Supplementary Material 3, showing that, when these curves are used, the Q-residuals detector also performs better than the PSSM algorithms in most of the cases.

The percentage of positions showing interdependences, I_{dep}, varies among the studied binding sites as can be observed in Figure 2. A correlation test was performed between the I_{dep} and the improvement in binding site detection when Q-residuals detector was compared with MAST. The improvement in binding site detection was derived by subtracting the mean AUC for each binding site calculated using each method. Results show a significant correlation between the number of strong interdependent sites within a binding locus and the amount of improvement of the Q-residuals detector over MAST (as measured in terms of AUC). Performing the test on the results for JASPAR database gave a $p\text{-}value = 0.004$; the corresponding result for the TRANSFAC database was a $p\text{-}value = 0.04$.

The computational times of the Q-residuals detector, of MAST and of our R implementation of MATCH were compared for the detection of TFBS within promoter sequences. To compare the three algorithms, the MAST algorithm (MEME version 4.4.0), the C code for Q-residuals using the ideal number of components, and our implementation of MATCH algorithm in R with the ideal Core Similarity were used. The background corresponded to Background 1 for each organism—this consisted of 1500 nucleotides. The threshold for each method was set in such a way that the number of positives was similar. In the case of MAST a p-value of $p = 0.001$ was chosen, for Q-residuals a confidence interval of $C = 0.95$ was set, and for MATCH the Similarity was set to $S = 0.85$. The time was calculated for 100 iterations of the program. The average computational time in detection for the TRANSFAC database motifs are 0.003 ± 0.001 s using the Q-residuals detector, 0.0191 ± 0.001 s using MAST and 0.33 ± 0.03 s for the R implementation of MATCH. The results show that the Q-residuals detector is faster than MAST and the R implementation of MATCH in all the studied binding sites.

The Q-residuals detector was also compared with Motifscan, an algorithm which takes into account interdependences. Using the same criteria as Naughton *et al.* (2006), a 5% increase in the ROC_N AUC was required for an improvement to be considered significant. The results showed that in 34 of the 93 studied motifs Motifscan performs better than either the Q-residuals detector or the PSSMs methods, that Q-residuals is the best detector in 25 of the 93 motifs whereas PSSM is best in just 1 of them. The three detectors perform equally well in 16 motifs; Q-residuals and Motifscan are equally good and better than PSSM in 16 motifs; Q-residuals and PSSM

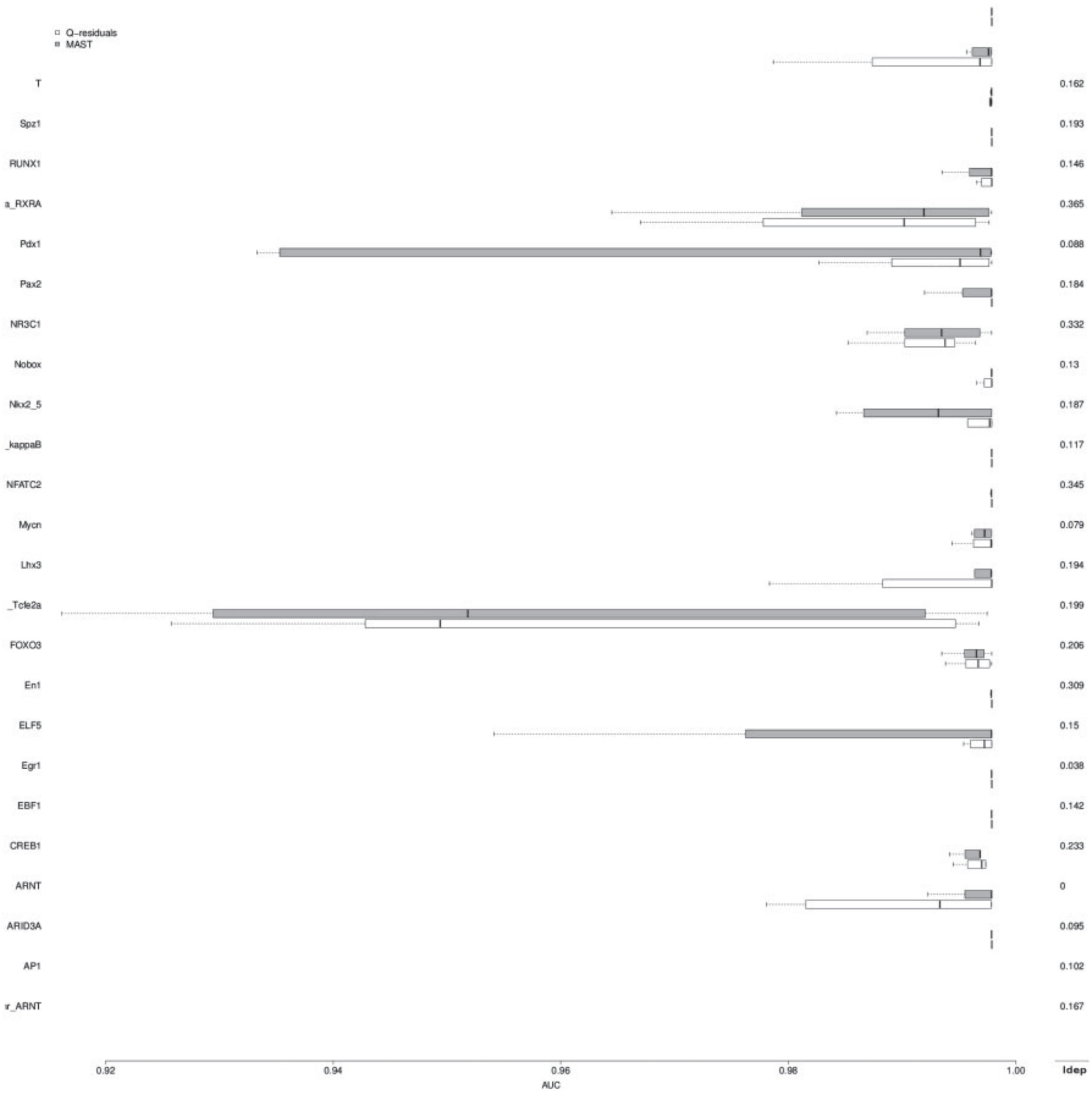

Fig. 2. Box plot of the AUC and its variation for the studied TFs, comparing the Q-residuals detector with the chosen number of components (in white) to MAST (in gray). The results correspond to the Background 1 of each organism. I_{dep} corresponds to the rate of positions within a binding site which have significant interdependences.

are better than Motifscan in 3 motifs; and Motifscan and PSSM are better than Q-residuals in 9 of the 93 motifs. The AUC performance is shown in the Supplementary Material 4. A visualization of the results in Figure 3 shows that the performance of Q-residuals is more sensitive to the number of positions. When the sequences are short, the number of false positives using the Q-residuals detector increases, leading to a smaller AUC. Motifscan performs better in this situation but, on the other hand, it needs more training sequences, so when the number of sequences is small, Q-residuals performs better than Motifscan. Focussing on the 37 motifs which have <20 sequences available, in 43.24% of the cases the AUC of Q-residuals shows it to be significantly the best algorithm, whereas Motifscan is best in just 27.02% of the instances. In most cases, even if Motifscan is significantly better than Q-residuals, the Q-residuals algorithm performs better than PSSM methods for this comparison also.

4 CONCLUSIONS

Calculating the residuals of the covariance model of the numerical TFBS has been demonstrated to be an effective method of detecting TFBS within real data, with better performance than existing MEME and MATCH methods.

The results show that, when there are no interdependences, our method is at least as good as the PSSM methods we used for comparison, but we also found a correlation between the improvement in AUC and the percentage of positions showing interdependences in a TF. This result proves that covariance can

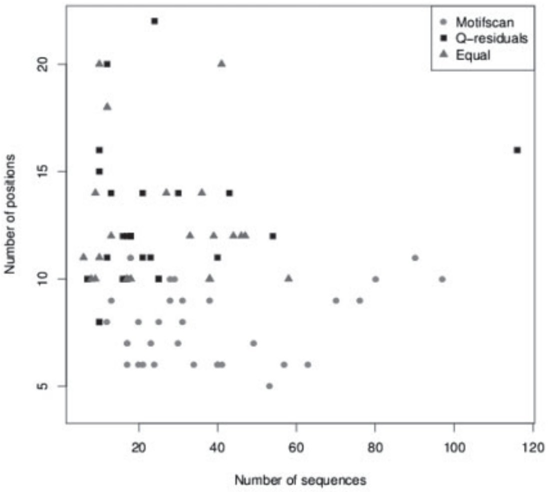

Fig. 3. Number of positions and number of sequences of the motifs where Motifscan was the best algorithm, (•); where *Q*-residuals was the best algorithm, (■); or where both perform equally well (<5% difference in AUC) in (▲). *Q*-residuals performs better for small number of sequences, but performs worse when the number of position per sequence is small.

capture position interdependences in TFBS, and that a covariance-based model can be useful in detecting TFBS within large databases.

When we compared the computational time of the *Q*-residuals detector and PSSM-based methods, we found that *Q*-residuals is faster; in contrast, other methods which take into account interdependences usually carry a high-computational cost. Another advantage of the *Q*-residuals detector, as compared with methods which take into account position interdependences, is that *Q*-residuals does not need a large amount of data in order to build a reliable detector.

The ideal number of components was chosen following a robustness criterion, biasing sequence background independence. Usually the number of components which satisfies the above condition is small, models having between 1 and 4 components explain most of the variance of the motif. Differences in detection using a range of components are not always significant.

As compared with a method which takes into account interdependences, *Q*-residuals shows a significant performance improvement when the number of sequences is small, but it also shows a larger sensitivity to the number of positions. *Q*-residuals needs more positions than Motifscan or PSSM to decrease the number of false positives.

Funding: This work has been partially supported by the Spanish Ministerio de Ciencia y Tecnología through the TEC2010-20886-C02-02 and TEC2010-20886-C02-01 grants, and the Ramon y Cajal program. E.P. wishes to thank IBEC for financially supporting her PhD. E.P. and S.M. are part of the 2009SGR-0753 consolidated research group of the Generalitat de Catalunya, Spain. A.P. and J.M. are part of the 2009SGR-1395 consolidated research group of the Generalitat de Catalunya, Spain.

Conflict of Interest: none declared.

REFERENCES

Anastassiou,D. (2001) Genomic signal processing. *IEEE Signal Proc. Mag.*, **18**, 8–20.

Bailey,T. and Elkan,C. (2006) Meme: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.

Bailey,T. and Gribskov,M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.

Barash,Y. *et al.* (2003) Modeling dependencies in protein-DNA binding sites. In *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology, RECOMB '03*. ACM, New York, NY, USA, pp. 28–37.

Ben-Gal,I. *et al.* (2005) Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics*, **21**, 2657–2666.

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

Castelo,R. and Guigó,R. (2004) Splice site identification by idlbns. *Bioinformatics*, **20** (Suppl. 1), i69–i76.

Cristea,P. (2005) Representation and analysis of DNA sequences chapter 1, pp 15–65. In edited by: E.G. Dougherty, I. Shmulevici, Jie Chen, Z. Jane Wang *Genomic Signal Processing and Statistics*. Hindawi Publishing Corporation.

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

Davis,J. and Goadrich,M. (2006) The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine learning*, Pittsburg, PA.

D'haeseleer,P. (2006) What are DNA sequence motifs? *Nat. Biotech.*, **24**, 423–425.

Edgar,R. (2004) Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Elnitski,L. *et al.* (2006) Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res.*, **16**, 1455–1464.

Hannenhalli,S. (2008) Eukaryotic transcription factor binding sites–modeling and integrative search methods. *Bioinformatics*, **24**, 1325–1331.

Jackson,J.E. (2004) *A User's Guide to Principal Components*. John Wiley & Sons, pp. 36–40.

Kanz,C. *et al.* (2005) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **33** (Suppl. 1), D29–D33.

Kel,A. *et al.* (2003) MATCHTM: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.

Larkin,M. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Liew,A.W.-C. *et al.* (2005) Pattern recognition techniques for the emerging field of bioinformatics: a review. *Pattern Recogn.*, **38**, 2055–2073.

Loots,G. and Ovcharenko,I. (2004) rVista 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res.*, **32**, W217–W221.

Maynou,J. *et al.* (2010) Computational detection of transcription factor binding sites through differential Renyi entropy. *IEEE Trans. Inf. Theory*, **56**, 734–741.

Minka,T.P. (2003) Bayesian inference, entropy and the multinomial distribution. *Technical Report*. Microsoft Research. [online] Available: http://research.microsoft.com/ minka/papers/multinomial.html.

Naughton,B.T. *et al.* (2006) A graph-based motif detection algorithm models complex nucleotide dependencies in transcription factor binding sites. *Nucleic Acids Res.*, **34**, 5730–5739.

Pairó,E. *et al.* (2011) Meet: motif elements estimation toolkit. In *Proceedings of the IEEE Conference on Engineering in Medicine and Biology (EMBC 2011).* Boston, USA.

Pavesi,G. *et al.* (2004) In silico representation and discovery of transcription factor binding sites. *Brief. Bioinform.*, **5**, 217–236.

Pearson,K. (1901) On lines and planes of closest fit to systems of points in space. *Philos. Mag.*, **2**, 559–572.

Portales-Casamar,E. *et al.* (2010) Jaspar 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38** (Suppl. 1), D105–D110.

Sandve,G. and Drablos,F. (2006) A survey of motif discovery methods in an integrated framework. *Biol. Direct*, **1**, 11.

Schmid,C.D. *et al.* (2006) EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res.*, **34**, D82–D85.

Shmulevich,I. and Dougherty,E.R. (2007) *Genomic Signal Processing (Princeton Series in Applied Mathematics).* Princeton University Press, Princeton, NJ, USA.

Silverman,B. and Linske,R. (1986) A measure of DNA periodicity. *J. Theor. Biol.*, **118**, 295–300.

Sing,T. *et al*. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics (Oxford, England)*, **21**, 3940–3941.

Stacklies,W. *et al*. (2007) pcaMethods–a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, **23**, 1164–1167.

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

Tomovic,A. and Oakeley,E.J. (2007) Position dependencies in transcription factor binding sites. *Bioinformatics*, **23**, 933–941.

Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.

Wilcoxon,F. (1945) Individual comparisons by ranking methods. *Biometrics Bull.*, **1**, pp. 80–83.

Wingender,E. *et al*. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.

Zhang,M. and Marr,T. (1993) A weight array method for splicing signal analysis. *Comput. Appl. Biosci.*, **9**, 499–509.

Zhao,X. *et al*. (2006) Finding short DNA motifs using permuted Markov models. *J. Comput. Biol.*, **12**, 894–906.

Zhou,Q. and Liu,J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **20**, 909–916.